UNIVERSITY OF BIRMINGHAM

THEORETICAL AND PROBABILISTIC METHODS FOR TOXICOKINETIC PREDICTIONS IN DAPHNIA MAGNA AND THEIR APPLICATION TO ENVIRONMENTAL RISK ASSESSMENT

by

Jacob-Joe Collins

A thesis submitted to the University of Birmingham for the degree of

Doctor of Philosophy

School of Biosciences

College of Life and Environmental Sciences

University of Birmingham

May 2024

ABSTRACT

The use of *in silico* and *in vitro* methods, commonly referred to as new approach methodologies (NAMs), has been proposed to support environmental (and human) chemical safety decisions, ensuring enhanced environmental protection. Toxicokinetic (TK) models developed for environmentally relevant species are fundamental to the deployment of a NAMs-based safety strategy, enabling the conversion between external and internal chemical concentrations, although they require historical TK data and robust physical models to be considered a viable solution. *Daphnia magna* is a key model organism in ecotoxicology albeit with limited quantitative TK data, as for most invertebrates, resulting in a lack of robust TK models. Moreover, current *D. magna* models are chemical specific, which restricts their applicability domain.

The overall aim of this thesis was to advance theoretical and probabilistic methods for TK predictions in *Daphnia magna* for use in environmental risk assessment (ERA). Firstly, to address current data limitations a *D. magna* TK dataset was collated from the literature and developed into an R package named *AquaTK*. Subsequently, a proof-of-concept Bayesian framework was developed to predict steady-state concentration ratios from the data. The application of the Bayesian framework to ERA was illustrated with an atrazine case study that showed prediction improvements (uncertainty reductions) with increasing amounts of data availability.

A substantial fraction of chemicals in the *AquaTK* dataset are ionisable at environmentally relevant pHs, therefore, the effect of ionisation on TK predictions was investigated within the Bayesian framework. Inferred steady-state concentration ratios were compared

with predictions from the Bayesian predictive model and the state-of-the-art non-lipid organic matter (NLOM) model. Predictions from the Bayesian model did not improve after accounting for ionisation but the prediction errors were lower relative to the NLOM model. Additionally, the largest prediction errors occurred primarily for neutral chemicals, indicating that factors beyond ionisation also should be considered.

A protein surface-binding (PSB) model was developed to integrate protein binding, which is a key pharmacokinetic parameter, as a function of *D. magna* protein fraction and external concentration. A theoretical upper bound for the PSB model was determined and evaluated against the *AquaTK* dataset, which highlighted that the bound holds for a range of external concentration scenarios. This will have positive implications for ERA where risk assessors can predict concentration ratios under any exposure scenario with minimum data requirements and without the need for *in vivo* data.

To integrate biotransformation into TK models requires quantitative data for biotransformation products (BTPs) that can only be obtained with standards. However, the lack of commercially available standards for BTPs presents a significant challenge. Semi-quantification methods that predict concentrations from ionisation efficiency values have been developed. Therefore, a random forest regression model to predict relative ionisation efficiency values was developed on experimental parent and BTP data and showed promising results compared to other studies, in addition to robust predictions on unseen data. This was an important first step in the semi-quantification of BTPs for TK modelling purposes with further work required to predict concentrations of BTPs without standards.

A consistent theme throughout this research is the use of *in silico* NAMs with an "open data" approach to sharing and generating data for ERA. Overall, this work provides the foundations of a modern ERA that does not require animal testing through the increased use of theoretical and probabilistic methods. Further work should endeavour to integrate all these methods into a predictive tool for ERA.

ACKNOWLEDGEMENTS

Firstly, I would like to extend my gratitude to my project supervisor Professor Mark R. Viant for the opportunity to undertake this PhD. Thank you for guiding me through covid and allowing me to follow my research interests with no reservations and your full support. Your encouragement and knowledge were invaluable to completing this research.

To Dr. George Fitton - it is hard to put into words how grateful I am for your time and dedication to helping me on this project. Thank you for taking on the role of primary industrial supervisor. Your patience with my countless number of questions, commitment to teaching me, and morale support were indispensable.

Thank you also to Dr. Joe Reynolds for your Bayesian expertise and mathematical support, your help and knowledge were essential to the success of this project. I also need to extend my gratitude to Dr. Bruno Campos and Dr. Claudia Rivetti for your constant support, guidance and knowledge and making visits to Unilever so enjoyable!

I want to extend my thanks to Dr. Leonardo Contreas for teaching me about machine learning techniques and answering all of my questions! To Patrik Enki and Tym Pietrenko thank you for the scientific discussions around my project and your contributions to the paper in review.

Special thanks to the Phenome Centre Birmingham and the fourth-floor metabolomics team that includes Dr. Andy Southam, Dr. Martin Jones, Dr. Ralf Weber, Lauren Cruchley-Fuge, and Ossama Edbali for conducting and processing the LC-MS experiments for the RIE project in Chapter 5 and your integral advice on the application of these

methods to my research. Also special thanks to David Epps for all your administration help and problem solving over the last four years.

Thank you to the Biotechnology and Biological Sciences Research Council (BBSRC) and Unilever for providing funding for this project

To my Mum and Dad thank you for always supporting and believing in me. I would not have been able to achieve any of this without you. Thank you to my Nan and Grandad who instilled a curious mindset in me from a young age. To my Sister thank you for your understanding and encouragement.

To my friends you have been the best distraction over the last four years even with the constant exclamation for me to get a"real job" or to "leave school". Every one of you has contributed to my success.

Finally, thank you to Ailish - your encouragement, patience, and support cannot be understated. Thank you for always listening to me discuss my work even when you had no clue what I was talking about and believing in me no matter what.

PUBLISHABLE MATERIAL

The content in Chapter 2 has been submitted for publication in Aquatic Toxicology and has been reviewed with recommended minor revisions and modifications at the time of thesis submission:

"A Proof-of-Concept Multi-Tiered Bayesian Approach for the Integration of Physiochemical Properties and Toxicokinetic Time-Course Data for *Daphnia magna*"

Authors

Jacob-Joe Collins[1,*], Joe Reynolds[2], Bruno Campos[2], Patrik Engi[2], Claudia Rivetti[2], Tymoteusz Pietrenko[2], Mark R. Viant[1], George Fitton[2]

[1]School of Biosciences, University of Birmingham, Birmingham, B15 2TT, United Kingdom

[2]Unilever, Safety and Environmental Assurance Centre, Colworth Science Park, Sharnbrook, Bedfordshire MK44 1LQ, United Kingdom

Authors' contributions

- J-JC: conceptualisation, data curation, formal analysis, investigation, methodology, writing – original draft, software, writing – review & editing.
- JR: conceptualisation, methodology, software, writing – original draft, writing –

review & editing.

- BC: project administration, supervision, writing – review & editing.
- PE: data curation, methodology, software.
- CR: supervision.
- TP: data curation, methodology, software.
- MRV: project administration, supervision, writing – review & editing.
- GF: conceptualisation, methodology, software, supervision, writing – original draft, writing – review & editing.

All authors read and approved the final manuscript.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

3Rs Reduction, replacement and refinement

ADME Absorption, distribution, metabolism (biotransformation) and elimination

BTPs Biotransformation products

CAS Chemical Abstracts Service

CMP Canadian Chemicals Management Plan

EC_{50} 50% maximal effect concentration

ECHA European Chemicals Agency

EFSA Environmental Food Safety Authority

ERA Environmental risk assessment

EU European Union

FI-MS Flow injection-mass spectrometry

fup Fraction unbound in plasma

IE Ionisation efficiency

LD_{50} Lethal dose for 50% of the population

MCMC Markov chain Monte Carlo

MoA Mode of action

MS Mass spectrometry

NAMs New approach methodologies

NLOM Non-lipid organic matter

NRC National Research Council

OECD Organisation for Economic Cooperation and Development

PBTK Physiologically based toxicokinetic

PCA Principal component analysis

PEC Predicted environmental concentration

PFAS Perfluoroalkyl substances

pKa Acid-dissociation constant

PNEC Predicted no effect concentration

PoDs Point of departure

PSB Protein surface-binding

qIVIVE Quantitative *in vitro-in vivo* extrapolations

QSARs Quantitative structure-activity relationships

REACH Registration, Evaluation, Authorisation and Restriction of Chemicals

RF Response factor

RIE Relative ionisation efficiency

RMSE Root mean square error

TD Toxicodynamic

TK Toxicokinetics

TSCA U.S. Toxic Substance Act

US United States

US EPA US Environmental Protection Agency

WHO World Health Organisation

CHAPTER 1 : INTRODUCTION

1.1 Current environmental risk assessment landscape

1.1.1 Global chemical estimates

There has been an exponential increase in the variety and volume of chemicals used globally in recent decades [256]. The chemical industry has grown at a significant rate since the 1960s, where between 1965 and 2006, the number of chemicals registered on the Chemical Abstracts Service (CAS) database increased from around 211,000 to 89,000,000 [35, 86]. Recent estimates reported that over 174,000 chemicals were pre-registered with the European Chemicals Agency (ECHA), 100,000 were inventoried in United States (US) commerce, and more than 350,000 chemicals were available commercially [267, 256]. Furthermore, there are an estimated 600 new chemicals registered every year with the US Environmental Protection Agency (US EPA) [262]. These chemicals have a wide range of applications within food additives, cosmetic ingredients, cleaning products, pharmaceuticals, biocides, and industrial chemicals [86]. However, throughout the 20th century, a substantial amount of available chemicals were released without prior safety information or risk assessment [262]. This meant that information concerning environmental or toxicological profiles was missing for many chemicals [86].

1.1.2 Environmental risk assessment and regulations

Environmental risk assessment (ERA) aims to quantify the effect of anthropogenic stressors, such as chemical use, with the aim of identifying potential effects on organisms or ecosystems [102]. This involves a culmination of methodologies to determine the risk of chemicals on the environment [22]. ERA can be divided into a four step framework developed by the National Research Council (NRC) of the US in 1983 [37]:

1. problem formulation (including hazard identification)
2. hazard characterisation
3. exposure assessment
4. risk characterisation

Problem formulation is a critical first step in ERA that involves the comprehensive analysis of all the key factors that need to be considered for risk assessment [239]. Firstly, the problem formulation phase identifies the objectives, scope, and purpose of the ERA [239]. For example, the risk assessment could be centred around a specific set of chemicals or hazardous exposure from a particular source, such as wastewater effluent [239]. A key part of problem formulation is the identification of the hazard or stressor and the information needed to assess the likelihood and potential impacts on the environment and the effects of exposure [77]. This includes the definition of appropriate endpoints for assessment and measurement [238, 22].

The second step of ERA is hazard characterisation. This step aims to quantify the probability of adverse effects through dose-response relationships. Dose-response relationships quantify the change in an observable effect as a result of a change in dose [179]. This results in a predicted no effect concentration (PNEC), which is the estimated chemical concentration expected not to have an adverse effect [5]. Traditionally, methods such as the acute or chronic toxicity test can be conducted to establish the effects on representative environmentally relevant species, such as aquatic vertebrates, invertebrates,

and algae [248]. An important assumption is that the most sensitive species is chosen as the protective benchmark for all species. This creates thresholds that protect all species within a given compartment, air, water, or soil [248].

Exposure assessment is the next step. The exposure assessment involves the quantification of the internal and external concentrations [221]. This yields a measured or estimated environmental concentration, referred to as the predicted environmental concentration (PEC). Exposure assessment can be divided into the environmental fate that describes factors that influence the chemicals availability and exposure and toxicokinetics (TK) that describes the fate of the chemical in a given organism [22]. Finally, the risk characterisation step involves analysing and evaluating the data from the previous steps and establishing the potential risk from exposure to the chemical [221]. A standardised way of characterising risk is by calculating a risk quotient by dividing the PEC by the PNEC [32].

A considerable number of regulatory and advisory bodies undertake ERA worldwide, which include ECHA, US EPA, the Organisation for Economic Cooperation and Development (OECD), Environmental Food Safety Authority (EFSA), and the World Health Organisation (WHO) [22]. ERA is required under chemical regulations, such as the Canadian Chemicals Management Plan (CMP), Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) in the European Union (EU), and the U.S. Toxic Substance Act (TSCA) in order to further push an international endeavour to ensure the safety of chemicals [129]. As an example, in Europe, the REACH regulatory framework specifies standardised toxicity tests for key trophic levels, such as primary producers, primary consumers, and secondary consumers [32]. Consequently, this makes plants, invertebrates, and fish key model organisms for ERA [108].

1.2 *Daphnia magna* as a model organism

Daphnia are one of the most extensively studied model organisms with a vast number of publications spanning multiple scientific disciplines, such as ecology, evolution, genomics

and toxicology [241, 9]. Specifically, *Daphnia magna* commonly referred to as the water flea, are of particular interest to ERA, due to their key ecological role. *D. magna* provide a link between primary production and the higher trophic levels of the aquatic food chain [112, 148, 9]. *D. magna* act as a primary consumer feeding on algae but are also predated on by other invertebrates and fish [250]. Populations of *D. magna* are found in the pelagic zone of freshwater environments, ranging from small temporary pools to lakes [241, 9].

1.2.1 Anatomy, morphology & physiology

D. magna are some of the largest *Daphniidae* with adults ranging from 5 to 6mm [135]. The anatomy and physiology of *D. magna* can be seen in Figure 1.1. The most distinctive anatomical features include the compound eye and two-branched antennae [246]. *D. magna* have a chitinuous shell called a carapace, which encloses their inner wall and provides protection. The carapace is made up of colourless chitin, however, due to haemoglobin production the *D. magna* have a reddish colour [206, 250]. Rather than a system of blood vessels like fish, *D. magna* have an open circulatory system, which contains haemolymph that surrounds all the organs [84]. *D. magna* have 10 pairs of appendages, which include antennae, antennules, mandibles, and maxillae, with five thoracic limbs located at the trunk and an abdominal claw [250].

D. magna are only capable of collecting food particles by filtration, which makes them unable to actively select food particles [47]. The movement of the thoracic limbs generates a current that sucks water between the filtering limbs and the carapace valves. The sucked in water is enclosed by the end of the limbs and then forced through the setules, which act as sieves on the filtering limbs [47]. This filtration is not only essential for feeding but also respiratory exchange. The importance of feeding current for oxygen uptake has been identified using a combination of microscopy and special optical techniques [210].

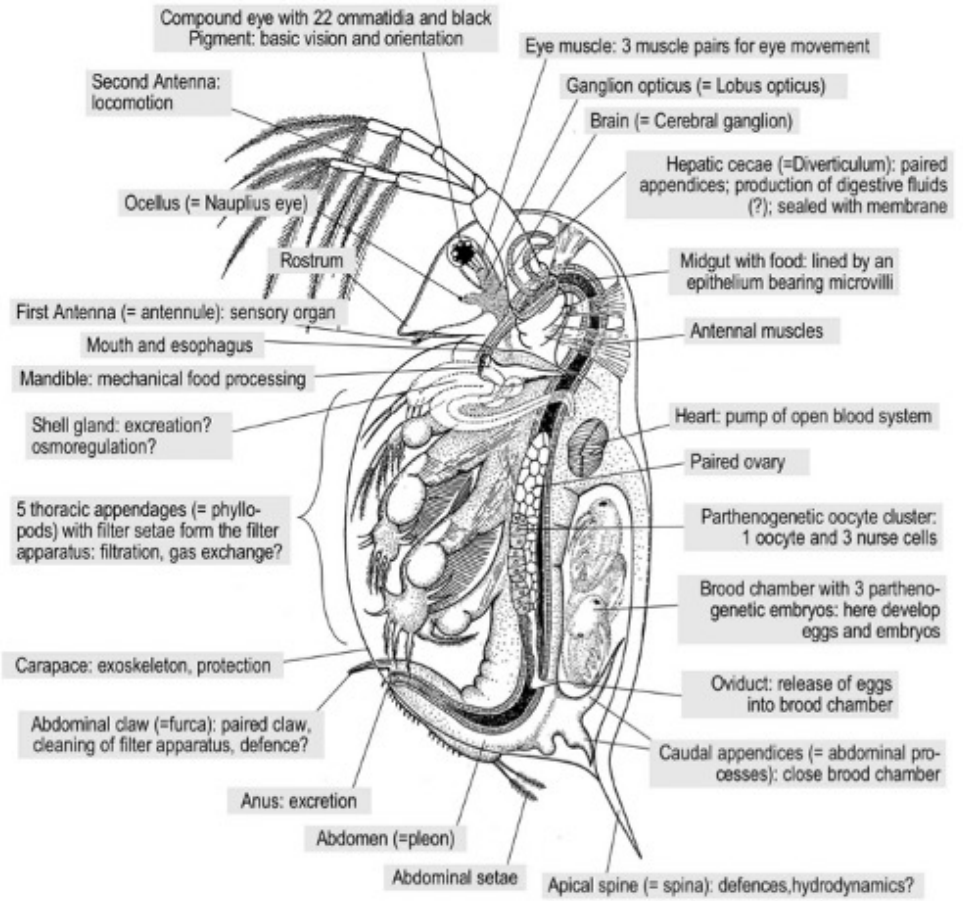

Figure 1.1: The functional anatomy of *Daphnia magna* from [84].

1.2.2 Life cycle & development

Under favourable conditions the *D. magna* reproductive cycle can transition between asexual and sexual phases, also known as cyclic parthenogenesis [84, 133, 250]. In the parthenogenic phase, the *D. magna* population is made up entirely of females, which produce diploid eggs in the ovary (Figure 1.2). If feeding and environmental factors allow a female will produce parthenogenetic eggs after each adult molt [84] . Females can produce eggs in the brood chamber every 3-4 days, which can last up to 2 months in a laboratory setting [84].

D. magna primarily reproduces through parthenogenesis, however, environmental conditions, such as droughts or extreme cold, in addition to external factors, such as scarcity of food or high population density can activate sexual reproduction and form males iden-

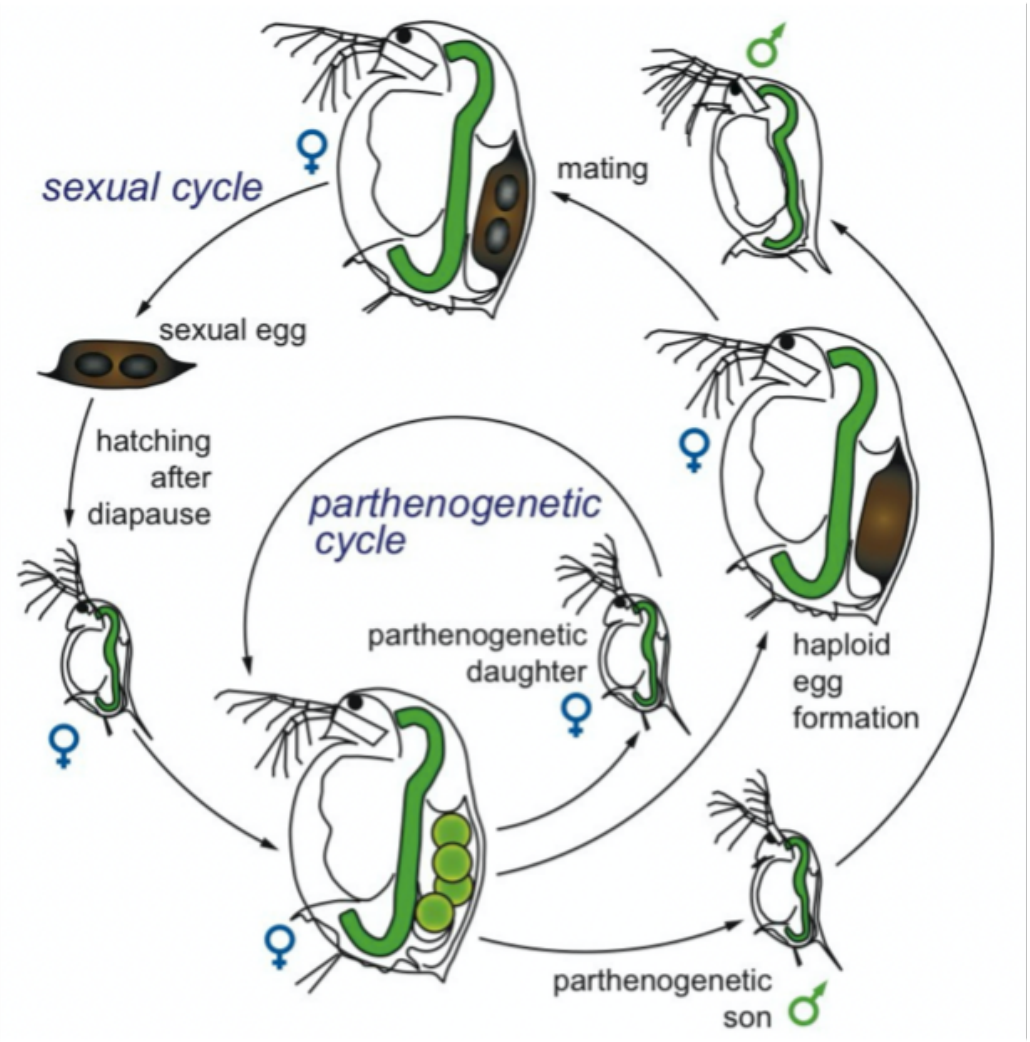

Figure 1.2: Life cycle of a cyclic parthenogenetic *Daphnia magna* from [84].

tical to the mother during a special parthenogenic event (Figure 1.2) [241]. During these environmental or external stimuli, it is possible for *D. magna* to produce haploid resting eggs by meiosis, which requires fertilisation by males and a period of dormancy. Resting eggs are encapsulated by a structure called an ephippium, which protects the eggs [84]. Development of these eggs is only resumed under favourable external stimuli, such as rising temperature, to establish a new population [241].

Sexual reproduction can be easily removed using specific culture conditions in the laboratory, resulting in genetically identical *D. magna*. Consequently, leading to the removal of any genetic variation that could impact ecotoxicology studies. Additionally, the short reproductive cycles, high fecundity, ubiquitous nature, ease of handling, and short life

span have made *D. magna* an important experimental organism for the US EPA and the OECD [246, 250].

1.2.3 *Daphnia magna* in ecotoxicology

Daphnia have an extensive range of biomarkers related to their behavioural and physiological responses, which are induced by chemical substances and environmental factors [250]. The accumulation of contaminants in organisms is influenced by their size [46]. It has been argued that the greater the surface area to volume ratio the greater the uptake rate in small organisms [137, 135]. A brief search of Google Scholar (https://scholar.google.com) with "*Daphnia magna*" as a keyword resulting in over 580,000 related publications. Furthermore, using "*Daphnia magna* risk assessment" and "*Daphnia magna* ecotoxicology" as individual keywords resulted in over 118,000 and 49,000 publications, respectively. This highlights the extensive use of *D. magna* as a key model organism in ecotoxicology and risk assessments.

1.3 Review of *in silico* new approach methodologies

1.3.1 What are new approach methodologies and why are they important?

New approach methodologies (NAMs) can be defined as *in silico*, *in vitro*, and *in chemico* based approaches that can be used to support chemical safety decisions [129, 263]. NAMs incorporate data from a wide variety of resources, such as computational chemistry, high-throughput toxicity testing, toxicogenomics, exposure science, and new animal models [129]. The introduction of the LD_{50} (lethal dose for 50% of the population) in the 1920s led to a significant increase in animal use in toxicity testing. Throughout the 20th century several methods have been developed involving whole animal testing, however, these methods have ethical implications and are not always time or cost effective [96, 171]. There is a societal and regulatory push of the 3Rs principle, which encompasses the reduction, replacement, and refinement of animal-based toxicity studies through the use of

NAMs [49]. In recent years there has been a paradigm shift towards replacing these animal intensive methods in ERA with NAMs [129]. In 2020 Europe unveiled the "GREEN DEAL" with a Chemicals Strategy for Sustainability that promoted the replacement of animal testing with alternative approaches, such as NAMs and committed to the reduction of animal testing [221]. While the cosmetic industry in Canada, EU, and the US have the prohibition of animal testing under consideration or already in place [3], it does not directly apply to environmental testing.

1.3.2 Overview of *in silico* new approach methodologies and their development

In silico approaches are referred to as any computational simulation that replicates *in vitro* or *in vivo* experiments, in addition to assisting in the interpretation of complex toxicological experiments [200]. There has been significant focus on the replacement of traditional toxicity testing with *in silico* methods because they do not require any animal tissues or samples [251]. Additionally, a major advantage of *in silico* methods is that they can provide predictions for novel chemicals that have not been synthesised yet [214]. However, regulatory acceptance of *in silico* approaches is low because of the uncertainty of model predictions, which may be attributed to concerns of data quality and appropriateness of the data used, the chemical applicability domain, and the interpretation of how the input model features relate to the model predictions [3].

There is a wide range of computational tools under the *in silico* approach umbrella, with databases containing key toxicity data or chemical properties, prediction software for chemical descriptors, system biology simulation tools, predictive models generated from statistical packages and software, pre-built models available on-line or in application form, and visualisation tools [214]. Development of these models can be divided into 5 major steps [214]:

1. biological data acquisition

2. generate chemical descriptors
3. generate a predictive model
4. evaluate the accuracy of the model
5. interpret the model

There are many methods that can be utilised to generate a predictive model, which can encompass quantitative structure-activity relationships (QSARs) and read-across approaches.

QSARs correlate the structure of a chemical with physiochemical properties (e.g. lipophilicity), biological (e.g toxicity), and environmental fate properties to create predictive models [22]. The QSAR theory assumes that similar chemicals will have similar activities and properties, which allows estimation of toxicological endpoints to be deduced for those similar chemicals without endpoint values available [228]. For example, a QSAR was successfully developed for the acute toxicity of anionic surfactants in *D. magna* by correlating the EC_{50} (50% maximal effect concentration) with the partition coefficient of the chemical with R^2 values of 0.99 and 0.89 for alkylbenzene sulphonates and ester sulphonates, respectively [117]. QSARs can be used to predict the mode of action (MoA) of a chemical [22]. A chemical is assigned to a MoA group dependent on its chemical structure and the effect of the chemical on the model organism [45]. However, there is a lack of high-quality data and the varied classification systems can lead to varied results [45].

The read across method involves evaluating a toxic endpoint of an untested target chemical using results of the same endpoint for a tested chemical considered similar in terms of structure, properties, or biological effects [236]. There is a level of expertise required to successfully group chemicals and there is a lack of application of read across to ERA [22]. Statistical and numerical scoring can be applied to help group similar chemicals [22]. However, the major source of uncertainty in this method stems from the assumption underlying the similarity scoring between the untested target chemical and the library

of tested chemicals.[29]. Read across extrapolations can be integrated with QSARs to improve model predictions, which when combined with other *in vitro* and *in vivo* data in a weight of evidence approach can provide reliable characterisation of toxicological hazard [29].

1.4 Role of toxicokinetics as an *in silico* new approach methodology and its application in environmental risk assessment

TK is an important *in silico* NAM for environmental and human safety assessments. TK relies on understanding absorption, distribution, metabolism (biotransformation) and elimination (ADME) processes of a chemical within a biological system over time [59, 21, 26]. TK models can help provide a quantitative mechanistic framework of these ADME processes to establish the fate and effects of a chemical in an organism and its physiological processes [105, 120]. TK data is essential for route-to-route extrapolations, interspecies extrapolations, and mechanistic understanding [60]. TK studies allow the concentration at a target site to be understood, which can then be used to support toxicodynamic (TD) investigations [171]. TK models are particularly important for quantitative *in vitro-in vivo* extrapolations (qIVIVE) to relate *in vitro* assay data to the entire target organism [60]. Regulatory bodies are beginning to encourage the integration of TK data into ERAs. In REACH, TK data is not currently required, but it is advised that the kinetic profile is considered for human risk assessment [65]. Similarly, TK data is starting to be used within several OECD Test Guidelines to improve testing robustness, namely OECD Test Guideline 417, which states that *in vitro* tests can be used instead of *in vivo* animal testing as TK information [197, 65, 60]. Additionally, Test Guideline 417 provides information on utilising TK data to inform study design through dose selection, accumulation potential, and exposure routes [244].

TK models vary in complexity, ranging from the simplest one-compartment models to multi-compartment models, and at the most complex level, physiologically based toxicoki-

netic (PBTK) models. Visual representations of one-compartment, multi-compartment, and PBTK models can be seen in Figure 1.3. The simplest compartmental model consists of one compartment, where the organism is represented as one homogenous unit [149, 165]. However, compartmental models are not always based on anatomy or physiology [105]. Even without physiological relevance, compartments are still useful in describing and predicting TK when calibrated on specific chemicals. A one-compartment model represents the organism as a single homogeneous volume with chemical mass accumulation governed by Fick's theory of diffusion [94, 105]. As per Anderson. (2013) it is possible to consider the diffusion of a material i.e chemical, between two compartments (environment and organism) separated by a membrane through which a material can diffuse, though not necessarily with the same diffusion rates in the two directions i.e. uptake and elimination rates [11]. Therefore, it is the uptake rate and an elimination rate that modulate the uptake of external concentration and the elimination of internal concentration, respectively.

There are also multi-compartment TK models that are usually comprised of two [235] or three [24, 245] compartments. Multi-compartmental models may have compartments that represent specific organs. Generally, multi-compartment models have a central compartment and two peripheral compartments depending on the model complexity [105]. The mechanistic degree of the model can be linked to the number of compartments with each compartment usually reliant on empirical data to inform parameterisation of the model [105]. A multi-compartment model was utilised to model the TK of diethylhexyl phthalate in rainbow trout by introducing a separate gill compartment that included absorption and metabolism, due to the non-homogenous distribution of the chemical [24, 105].

PBTK models provide quantitative descriptions of ADME processes defined by the relationship between an organisms anatomy, biochemistry, and physiology [191, 140]. Typically, PBTK models are constructed of the specific organs or tissues essential in describing the ADME processes of the chosen chemical [105]. PBTK models can be parameterised

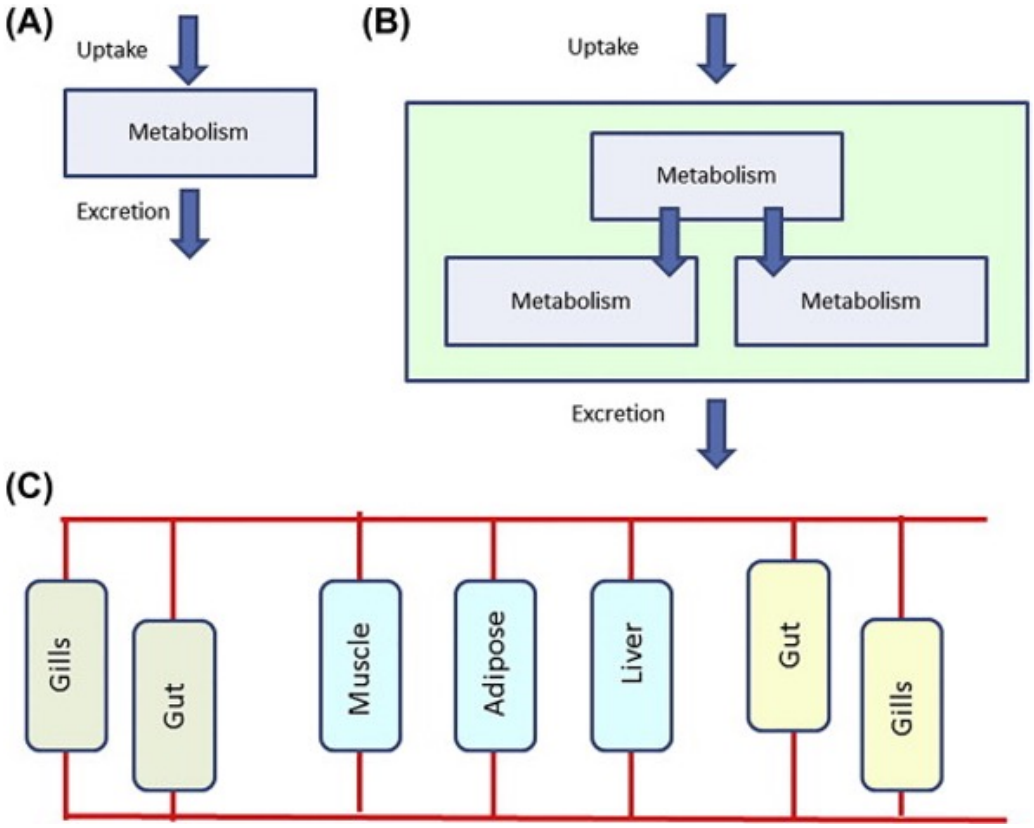

Figure 1.3: Examples of a one-compartment model (A), multi-compartment model (B) where the compartments do not necessarily represent physiological compartments, and a physiologically based toxicokinetic model (C) where compartments are physiologically relevant and movement between different compartments is taken into account. Figure from [192].

by *in vitro*, *in vivo*, and *in silico* data and allow extrapolation between different exposure scenarios [105]. PBTK models are built with the physiology of the organism in mind and have significant *a priori* knowledge, such as organ volumes and flow rates. These models are conventional in the pharmaceutical industry, however, the frameworks and exposure scenarios differ significantly with risk assessment of chemicals in the environment [185]. There are many examples of PBTK models for fish but models developed for invertebrates are sparse. A generalised fish PBTK has been developed for use in the ERA of pharmaceuticals [266]. A recent study has developed a PBTK model of perfluoroalkyl substances (PFAS) in zebrafish that can account for co-exposure scenarios of any number of chemicals [104]. There are no generic PBTK models for aquatic invertebrates and to develop a large number of chemical-specific models would be very resource inten-

sive [211]. Furthermore, development of only a single chemical-specific *D. magna* PBTK model would not be applicable to the significant number of chemicals that need to be assessed using ERA.

In the context of ERA, TK describes the internal concentration of a chemical within a given organism over time. This is important because adverse effects are not only dependent on the external environmental concentration but the concentration of chemical that penetrates into the interior of the organism. Utilising TK models enables the prediction of internal concentration of chemicals within organisms. These predictions are pivotal to accurately assessing the risk posed by exposure to chemicals. Overall, this enables informed decisions to guide environmental protection without the need for further animal toxicity testing.

1.5 Review of deterministic modelling and influential factors impacting chemical internal concentration predictions for environmental risk assessment

1.5.1 Long exposure environmental risk assessments

Aquatic environmental exposure to chemicals is typically over the lifetime of the organism [233, 183]. The accumulation of the chemical in the organism is typically assumed to be driven by the chemical concentration gradient, the difference between the external water concentration and the internal organism concentration and chemical partition profile. Over very long exposures, the chemical will reach steady state, e.g. an equilibrium is reached [94, 16, 105].

Steady-state internal concentrations relative to external concentrations (concentration ratio) provide a meaningful measure of chemical accumulation for long exposure environmental scenarios [149, 105]. Under controlled laboratory conditions the steady-state

concentration ratio is a widely used metric when assessing the accumulative potential of a chemical to an organism [164, 277].

1.5.2 Toxicokinetic diffusion models

The accumulation of chemical in an organism can be described by a mathematical model. Specifically, the exchange of chemical mass between the water and the organism can be modelled by a mathematical diffusion process [64]. A diffusion process driven by the concentration difference corresponds to a Fickian-like diffusion process [94]. Symbolically this is represented by the following differential equation,

$$\frac{dy}{dt} \propto \Delta y \tag{1.1}$$

where y is the internal concentration in kg/kg, t is time in days, and Δy is the scalar concentration difference between compartments. For the law to apply to ERAs, one must assume that the distribution of chemical in both the organism and the water are homogeneous, i.e., the equation has no dependence on spatial coordinates. The concentration difference is then defined as the difference between the internal and external concentrations,

$$\Delta y = y_w - y \tag{1.2}$$

where y_w is the external concentration. Equation 1.1 becomes,

$$\frac{dy}{dt} = D\,(y_w - y) \tag{1.3}$$

where D is the diffusion coefficient. TK models often assume diffusion is bidirectional, where the concentration can be allowed to diffuse at different rates in and out of compartments,

$$\frac{dy}{dt} = y_w \cdot k_{in} - y \cdot k_{out}. \tag{1.4}$$

This is explains why the diffusion process is often referred to as "Fickian-like". For the initial condition $y(t = 0) = 0$ Equation 1.4 has the particular solution,

$$y = y_w \cdot \frac{k_{in}}{k_{out}} \left(1 - e^{-t \cdot k_{out}}\right) \tag{1.5}$$

In relation to long exposures, which as suggested previously is typical for ERA, the rate of change in internal concentration is negligible, resulting in an approximate steady-state internal concentration,

$$y \approx y_w \cdot \frac{k_{in}}{k_{out}} \tag{1.6}$$

The ratio of internal to external concentrations is then directly expressed as the ratio of transfer rates ($y/y_w = \ k_{in} \ / \ k_{out}$). Given an external water concentration, the internal concentration is completely determined by the particular solution when the transfer rates are known.

The assumption that the accumulation of chemical in the organism follows Fickian-like behaviour means that the internal concentration can be predicted from actual measurements of the external concentration and transfer rate parameters. Measurements of the external concentration are relatively straightforward to obtain, which means predictions of the internal concentration depends on calculating uptake and elimination rates, which can be measured experimentally through TK studies. This approach is useful for ERA because the steps outlined can be performed with minimal difficulty, as seen in the literature and regulatory guidance.

A great deal of effort has been invested in determining semi-empirical expressions for the rates for different fish species, for many different chemicals, and for a range of physiological

and environmental parameters [114, 113, 16]. However, there is a lack of available models to predict the transfer rates of chemicals for *D. magna* with most studies being chemical specific, meaning the application to a wider range of chemicals is restricted. Based on the above there is a clear lack of available generic TK models for aquatic invertebrates, such as *D. magna* and further research is needed to investigate and develop generic simple one-compartment models specifically for *D. magna* that can be applied across an extensive range of chemicals for use in ERA. Furthermore, once a generalised model for *D. magna* has been developed the next logical step would be to create a species agnostic model where physiological differences that govern toxicokinetic processes are understood as they can vary across species space.

1.5.3 Factors influencing toxicokinetic processes

1.5.3.1 Lipophilicity - octanol-water partition coefficient (K_{ow})

The lipophilicity of a chemical is measured by the partition coefficient of a chemical between *n*-octanol and water [10]. It is used extensively in the chemical industry, pharmaceutical R&D, and environmental sciences [67, 54]. The octanol-water partition coefficient commonly has the symbol K_{ow} and is often represented in its logarithmic form as $\log_{10} K_{ow}$ [115]. K_{ow} is an important parameter as it can act as a surrogate for the accumulation potential of a chemical [15]. The relationship between accumulation metrics and $\log_{10} K_{ow}$ is well documented [187, 259, 163, 79, 17]. There are different methods for measuring K_{ow} including experimentally through the OECD Test Guideline 107, also known as the "shake flask method" [196]. A more reliable version of the shake flask method is OECD Test Guideline 123, referred to as the "slow-stirring method" [115, 198]. Both methods involve dissolving a chemical into octanol, adding water, and measuring the concentrations in each phase after shaking or stirring [115]. *In silico* predictions for K_{ow} are easily accessible because experimental data is available for a significant number of chemicals and predictions can be obtained from QSARs or free software [115]. A comparison between various prediction software highlighted that ADMET Predictor, ACDLabs,

and ChemSilico software predicted $\log_{10} K_{ow}$ within $\pm$ 0.5 log unit of measured values for 94.2%, 93.5%, and 93.5% of a 138 chemical test set, respectively [72]. This highlights the potential of utilising prediction software for K_{ow} predictions.

1.5.3.2 Ionisation – pKa, pH and distribution coefficient (D_{ow})

Researchers have become aware of ionising chemicals sensitivity to pH and the impact that has on toxicity and uptake in organisms [224]. The distribution coefficient (D_{ow}) is able to account for ionised and non-ionised forms of a chemical at a given pH [67]. The ratio of ionised and non-ionised forms is dependent on the pH and the pKa of the chemical [278]. According to the Brønsted-Lowry definition of acids and bases an acidic chemical is one that can donate a proton, while a base is capable of accepting a proton [28]. The acid-dissociation constant (pKa) of a chemical represents the pH of the solvent needed to obtain a 50/50 ionised/non-ionised state. Therefore, highlighting how acidic the environment must be so that the acid no longer dissociates protons. For example, if the functional group of a given chemical had a pKa of 9, at experimental pH 7, the chemical will tend to accept protons because the group itself has a smaller acidic behaviour (9) than water (7) and thus it accepts protons (receiving the acidic behaviour value of -1); alternatively, a functional group with a pKa of 3 will release protons in neutral solution (receiving thus an acidic behaviour value of +1) unless the water is more acidic than pH 3; finally, if the compound does not undergo ionisation in water, like alkanes, the chemical will be given an acidic behaviour value of 0. One example from the literature shows it is possible to determine the D_{ow} of a chemical at a given pH from the value of the K_{ow} and the pKa of the chemical using equation 1.7 [78],

$$log_{10} D_{ow} = \log_{10} K_{ow} - \log_{10}(1 + 10^{A(pH-pKa)}) \cdot Abs(A) \tag{1.7}$$

where A is either +1 for acids or -1 for bases and 0 for neutrals. This is a generalised equation that allows all three states to be accounted for and allows the chemical partition

coefficient predictions to revert back to K_{ow} when the chemical is not ionised. Multiple studies have shown the improvement of TK predictions for ionisable chemicals using the D_{ow} over the K_{ow} [99, 13]. Comparatively to K_{ow} most commercial software packages are able to predict D_{ow} using predictions of both K_{ow} and the pKa value of a chemical [73].

1.5.3.3 Lipid partitioning & protein binding

Many organic chemicals accumulate in the lipid fractions of aquatic organisms [278]. Bertelsen et al. (1998) showed that chemicals are concentrated in organisms and tissues with the highest lipid content [31]. Further studies have shown that concentration ratios are approximately proportional to K_{ow} with proportionality constants equal to the lipid fraction [187, 48]. Therefore, the steady-state concentration ratio depends not only the K_{ow} but also the lipid content as well [212]. Moreover, several studies have shown that the steady-state concentration ratio approximately doubles with the doubling of the lipid content [17, 75].

The steady-state concentration ratio is also determined by other biochemical components, such as protein. Debruyn & Gobas. (2007) suggest that if the lipid fraction makes up to less than 5% of the dry weight organic content, the absorption capacity of an organism will be dominated by the protein fraction [74]. This means investigation into protein chemical relationships are important for small aquatic invertebrates like *D. magna*, which have relatively small lipid fractions. Protein partitioning of chemicals can be described by the protein-water partition coefficient (K_{pw}), which highlights the distribution of a chemical between the protein fraction and environment [88]. These partition coefficients have been calculated for neutral chemicals using muscle protein from chicken, fish, and pig [88]. However, this limits the application of the calibrated model to neutral chemicals only. Additionally, as they are calibrated on species with large variations in physiology and biochemical content it may not be applicable to aquatic invertebrates, such as *D. magna*. While there is conservation at the cellular level where all animals will be similar

there might be potential physiological differences that drive variation in predictions. An alternative approach is by integrating protein binding, which is of particular interest to pharmacology R&D as only the unbound portion of a drug will have a pharmacological response [261]. In pharmacokinetics a drug binding to a protein in its simplest form is a rapid equilibrium and reversible process governed by the law of mass action [272, 261]. However, there is a lack of knowledge on the application of protein binding to aquatic invertebrates. A recent study was the first of its kind to demonstrate a mechanistic link between protein binding and TK modelling for the exposure of thiacloprid on *Gammarus pulex* [215]. Other key environmental chemicals of interest including PFAS have been shown to bind to human serum albumin in addition to protein-rich tissues [95, 160]. Most protein binding studies are chemical specific and further research is needed to establish the potential for a generalisable protein binding TK model for aquatic organisms.

1.5.3.4 Biotransformation

Biotransformation describes the process by which an organism transforming parent chemicals into biotransformation products (BTPs) through metabolic pathways [50]. Often biotransformation forms more hydrophilic compounds that can be excreted easier than their parent counterpart [255]. It was originally thought that all chemicals became less toxic after biotransformation, however, it is now known biotransformation can result in BTPs more potent than the parent chemical. For example, the insecticide parathion when converted to paraoxon is a more toxic chemical [50]. Additionally, EC_{50} values for growth inhibition of aquatic organisms were lower for ciprofloxacin (BTP) than the parent compound enrofloxacin [85, 278]. The biotransformation of the pharmaceutical diclofenac to diclofenac methyl ester resulted in a 430-fold increase in acute toxicity [98]. In relation to TK modelling, biotransformation is important to consider because it decreases the parent chemical concentration, which could lead to over predictions of the parent chemical. There are few available TK models that incorporate biotransformation processes for *D. magna*, mainly due to the lack of available reference standards for BTPs, which

hinders identification and absolute quantification of the internal concentration [58]. In the sections above (1.5.3) the influence of key parameters on TK predictions across human and ERA have been highlighted. Further research is needed to understand the impact of these parameters on TK predictions in *D. magna* and attempt to incorporate them into *D. magna* specific TK models, where '*D. magna* specific' refers to using biochemical compositions within the model to distinguish between organisms.

1.6 Methods for quantification of biotransformation products

1.6.1 Traditional quantification methods

Traditional quantitative methods for chemicals typically relies on the analytical technique mass spectrometry (MS). MS measures the mass-to-charge ratio (m/z) and the relative intensities of ionised molecules [8]. Advancement in MS instruments over the last decade has meant that MS has greater sensitivity in detecting small molecules like BTPs compared to other methods, such as nuclear magnetic resonance spectroscopy [56]. Generally, MS methods can be either untargeted or targeted. Untargeted methods aim to cover a broad range of small molecules, while targeted analysis aims to quantify one or a low number of pre-defined small molecules [260]. Untargeted analysis has greater scope in comparison with targeted from the perspective of its capability to discover and measure BTPs that are unreported [116]. There are two common types of quantification, which is relative or absolute quantification [152]. Relative quantification involves comparing the signal intensity of small molecules against reference or control groups [249]. As this does not require chemical standards it is commonly used for untargeted analysis of large numbers of analytes [270]. However, absolute quantification relies on authentic standards or isotope labelled standards to determine the exact quantity of the small molecule [152]. To achieve absolute quantification, a range of analyte concentrations are measured to create calibration curves to help define instrument linearity [125]. This is more applicable to targeted analysis where information about the identity of the small molecules is known *a*

priori [270]. The signal intensity of a small molecule is impacted not only by its concentration but its chemical structure and matrix [152]. Matrix effects and ion suppression can cause inaccurate quantification, especially in complex biological systems such as in pharmacokinetic studies [172, 152]. One major limitation of absolute quantification methods for BTPs is that standards are rarely commercially available or in reference databases [136]. Therefore, except in pesticide and drug toxicity studies, the absolute quantification of BTPs is rarely, if ever, conducted.

1.6.2 Methods for semi-quantification of biotransformation products

To overcome the limitations of available internal standards, new methodologies involving "semi-quantification" have been developed. Semi-quantification is utilised when reference standards are not available and under the assumption that a surrogate analyte with relatable characteristics can translate MS detector responses of a BTP into concentrations [249]. Over the last decade methods for semi-quantification of chemicals without reference standards have been undertaken, which are mainly based on structurally similar chemicals, closely eluting chemicals, using the response factor (RF) of the parent compound to predict the concentration of the BTP, related substructure, and electrospray ionisation efficiency [201, 1, 143, 167, 2]. The use of electrospray ionisation efficiency typically produces the best results compared to other methods [2]. Ionisation efficiency (IE) describes the extent to which analyte molecules in the liquid phase are converted to gas-phase ions and detected in the mass spectrometer [153, 201]. It is an important parameter because it varies dependent on the physiochemical properties of the chemical, chemical structure, and properties of the matrix and eluent of the study [167]. Multiple studies have identified key physiochemical properties that can be correlated with the IE of chemicals. Chalcraft et al. (2009) identified $\log_{10} K_{ow}$, molecular volume, effective charge and absolute mobility as key in ion evaporation of polar metabolites [55]. Alternatively, Oss et al. (2021) highlighted that IE by protonation is affected by basicity, molecular size in terms of molar volume or surface area, and hydrophobicity of the ion [202]. The most

common physicochemical parameters associated with a chemical's IE are related to the ionisability and hydrophobicity of the chemical [157]. However, most of the studies that have correlated parameters with IE are usually developed on a limited number of chemicals or the model is not able to describe the IE using the physiochemical descriptors of the chemicals [202].

A popular method for developing IE predictive models is using machine learning to train models on experimental measurements of IE values [167]. The most comprehensive analysis of IE data was conducted using a random forest regression analysis on 353 unique chemicals, including drugs, exogenous metabolites, amino acids, organic precursors, and lipids [156]. A random forest regression is an ensemble learning method that utilises multiple decision trees to improve the accuracy of predictions The algorithm creates subsets of the data through random sampling and a decision tree is trained on each subset of the data. After all the trees are created the model makes predictions based on an average of all the individual trees [234]. Liigand et al. (2020) used chemicals PaDEL descriptors, which are calculated molecular descriptors from open source software, to correlate with the IE data alongside multiple eluent composition descriptors, in positive and negative ion mode [273, 156]. The measured IE data was made relative to an "anchor compound" because the IE value can vary dependent on the instrument configuration [167].

Relative ionisation efficiency (RIE) predictions can then be used to estimate, or semi-quantify, concentrations of BTPs. Generally, RIE values are converted back to absolute IE values [156]. Then RF values are calculated for a set of calibration chemicals (internal standards) of known concentrations to make the RF instrument-specific by dividing each intensity measured in the MS by these known concentrations. These instrument specific RFs are correlated with predicted IE values for each calibration chemical through linear regression to enable the prediction of RFs. Finally, a predicted concentration for a BTP without a standard can be determined assuming the structure of the BTP is known, and an intensity is measured in the MS. The BTPs structure is used to predict an IE from a

predictive model, which can be used to predict an RF from the regression between the predicted IE and instrument specific RF values. Finally, a prediction of the concentration is calculated by dividing the measured intensity by the predicted RF. An overview of the workflow from predicting IE values to predicting concentrations of BTPs without standards can be seen in Figure 1.4.

Liigand et al. (2020) used a similar approach to predict the concentration of 35 pesticides and mycotoxins not included in the training dataset of the random forest regression model [156]. Overall, the root mean squared prediction error of the IE predictions were 2.2 and 2.0 times, in positive and negative ionisation mode, respectively. Moreover, the average quantification error was 5.4 times the experimental concentration, which is acceptable in relation to toxicology predictions [156]. This study highlighted the potential to predict IE, and therefore to semi-quantify the concentrations of chemicals, and also highlighted the importance of a robust chemical dataset and appropriate descriptors to obtain high quality results.

Mayhew et al. (2020) developed a machine learning approach that predicted RIE for 51 carboxylic acids, spanning a wide range of additional functionalities, to produce a model for predicting the RIE of chemicals without standards [173]. Rather than chemical descriptors, this study used molecular fingerprints, which turn chemical structures into binary bits (1s and 0s) to allow comparisons between chemicals [53]. For example, if a chemical has a hydroxyl group (-OH) then it would be represented with a 1, whereas if the chemical did not have this functional group it would be given a 0. This approach is applied across many different substructures and functional groups depending on the length of the fingerprints. The prediction results from Mayhew et al. (2020) were comparable with [201, 142, 141, 157]. A significant conclusion of this study was the ability to obtain comparatively good results while using molecular fingerprints instead of physiochemical descriptors.

Studies have shown the ability to predict RIE and concentrations relatively well, however,

there are few studies on the application of this approach to quantifying BTPs. A recent study measured the relative response factors of 26 pharmaceuticals and their BTPs [111]. Response factors were calculated by normalising the integrated peak area of each BTP with the peak area of the parent chemical. Results varied significantly even for parents and BTPs that had similar structures with relative response factors ranging from 70-fold lower relative response factors than the parent to 8.6-fold higher than the parent [111]. A drawback of using relative to parent concentrations is the assumed similarity in parent and BTP structure, but under biotransformation processes the structure of a BTP can change considerably, e.g. losing a functional group that impacts their IE [167]. Kruve et al. (2021) compared the BTP to parent method, with the closest eluting chemical method [209] and the ionisation efficiency method [156] for predicting concentrations for 341 chemicals [143]. The chemicals included pesticides, pharmaceuticals, and their transformation products in groundwater samples from Switzerland [143]. Evaluation of each method showed that the ionisation efficiency-based method had the best prediction accuracy, however, there were some major chemical outliers, and it was suggested that the IE predictive model training dataset should be updated with new chemical structures to achieve more accurate results [143]. Therefore, there is a clear need to obtain a wide variety of parent chemicals and related BTPs to enable the development of robust IE predictive models. This is important from a regulatory perspective as assessment schemes discuss including stable or toxic BTPs in risk assessment [90]. Additionally, if robust predictions of BTP concentrations using ionisation efficiency-based methods can be obtained then the predictions can be implemented into TK models as internal concentration compartments or to calculate biotransformation rates, which as previously described is essential for robust TK predictions. Having described the state-of-the-art in the semi-quantification of analytes, it is clear that the majority of research has focused on parents chemicals and further research into the semi-quantification of BTPs of industrial chemicals is required as a first step to predicting concentrations of BTPs for TK modelling purposes.

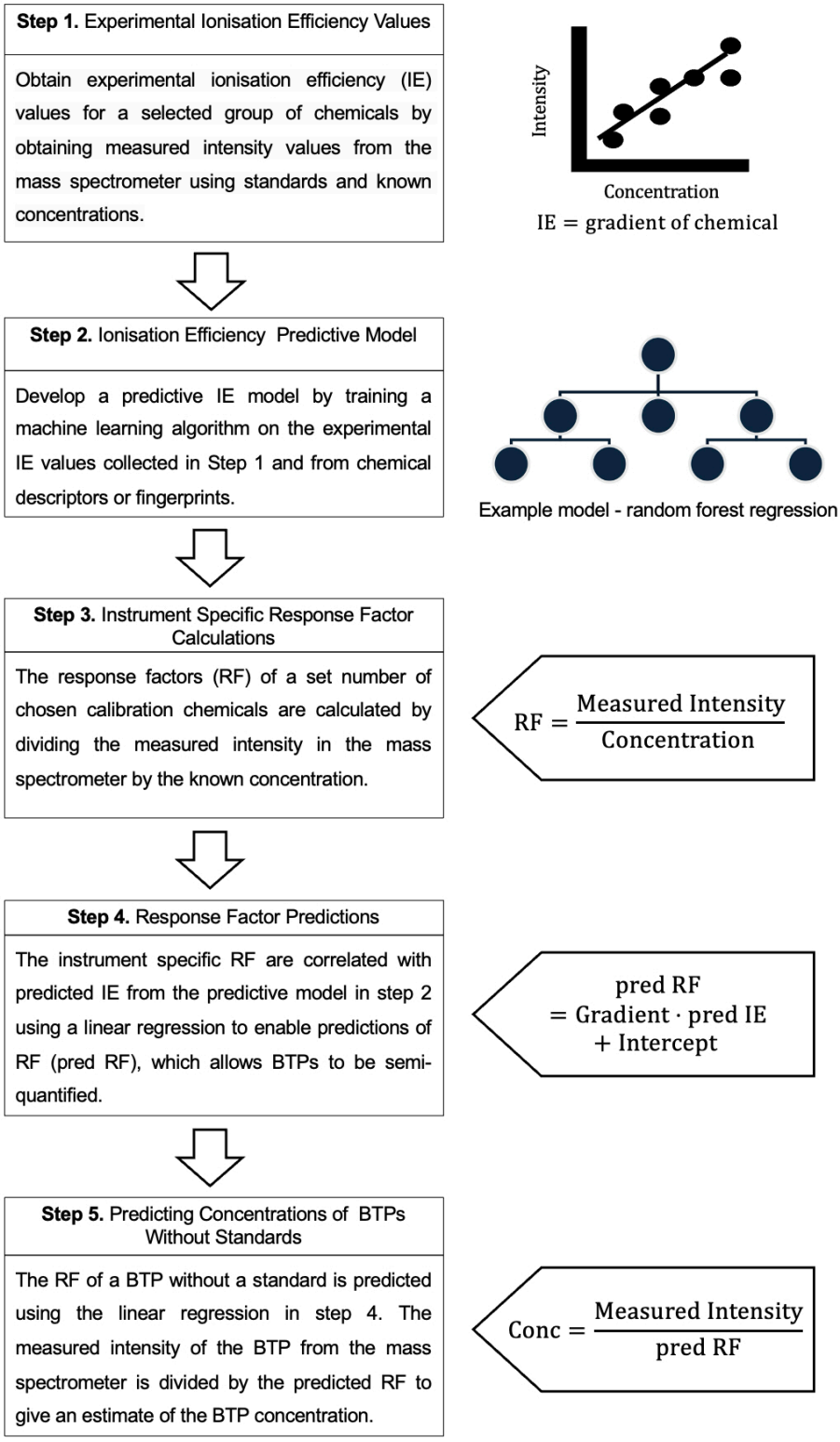

Figure 1.4: A 5-step workflow for predicting the concentrations of BTPs without standards.

1.7 Data considerations and acquisition methods for developing toxicokinetic models for *Daphnia magna*

1.7.1 Importance of data quality for the development of robust toxicokinetic models

The quality of the input data is essential to TK model predictions. The old adage is "garbage in = garbage out", where if poor data is input into the model it will give poor results [220]. A study extensively reviewed measured accumulation metric values for organic chemicals in aquatic organisms alongside the key factors that impact variance of the results [17]. A criterion was developed, which gave confidence scores ranking 1-3 (high to low score) for water analysis, radio-labelled chemical, aqueous solubility, exposure duration, and tissue analysis. The data quality assessment highlighted that 45% of steady-state concentration ratios had at least one major source of uncertainty leading to an underestimation of actual values [17]. This highlights the importance of reviewing study data and even applying criteria to establish uncertainty that could impact predictive model results.

1.7.2 *Daphnia magna* toxicokinetic data availability

The development of a TK model is dependent on the availability of experimental data. The US EPA created a large database of ecotoxicology data called "ECOTOX" [254]. This contained data for many toxicity endpoints for a wide range of species and chemicals. According to their website there is ecotoxicology data for over 12,000 chemicals spanning across more than 13,000 species. The data from ECOTOX is predominantly from peer-reviewed articles or gray literature, such as government documents and reports [199]. However, it is important to note that all data should be assessed for quality to assure confidence in the data before use in any model development or validation. Specifically for accumulation metrics there is a lack of raw data where the whole time-course profile

of the chemical is available [216]. Recent work has attempted to fill this data gap and create a publicly available database for accumulation study data [216]. This database is called $MOSAIC_{bioacc}$ and is available as a web application [184, 219]. $MOSAIC_{bioacc}$ contains more than 200 accumulation datasets for more than 50 genus across more than 120 chemicals [216].

Investigating these databases for *D. magna* data resulted in over 1,000 datapoints from the ECOTOX database, while the $MOSAIC_{bioacc}$ database returned 6 relevant time-courses for *D. magna*. The datapoints from the ECOTOX database lack the whole time-course data, which means important information about the uptake and elimination rates is lost. Conversely, the $MOSAIC_{bioacc}$ database provides whole time-course data but does not provide a substantial number of TK datasets for *D. magna*. Therefore, a collation of all available TK *D. magna* time-course data from the literature was required. Review of the available *D. magna* TK literature highlighted two major problems. Firstly, quantitative TK studies in invertebrates and *D. magna* specifically, are very limited especially in comparison to humans, rats, or fish. Secondly, even when quantitative TK data is available it is not fully accessible as it is restricted primarily to time-course plots, lacking the underlying data.

1.7.3 Methods for digitisation of toxicokinetic data

A significant amount of TK data is only available in plots. This has meant digitisation methods are essential to capture the historical quantitative internal concentration over time measurements. Digitisers can take images of TK time-course plots from available studies and extract each internal concentration and time datapoint with high accuracy. The accuracy of digitisation was illustrated during the development of a PBTK model where the error was shown to be less than 0.5% for biomarker concentrations following inhalation exposure [25]. Other key examples of the use of digitisation to obtain *in vivo* data include the $MOSAIC_{bioacc}$ database development, pharmacokinetic data, and parametrisation of a growth model for a mechanistic TK-TD model [231, 264, 216].

An example of how the digitisation process works and its accuracy is presented using TK data from in Figure 1.5 [87]. The TK time-course plot was screenshotted from the publication and uploaded to the free digitisation software [225]. The plot is calibrated by inputting the axes data. In this case the x axis would be set to 0 and 120 (hours) and the y axis would be set to -2 and 1 (log residue in *daphnids* μg g^{-1}). Focusing on the atrazine TK profile only (black outlined squares), each datapoint on the plot is manually marked with a cross ('X'). The digitiser includes the magnification of each data point to improve accuracy. Finally, the data is exported to CSV files. Figure 1.5 middle plot shows the exported data (orange crosses) for atrazine from the digitiser. The final plot on the right hand side shows the original plot overlayed on top of the digitised data plot with the axes matched [87]. It is clear that the atrazine digitised data near perfectly replicates the experimental data as the orange crosses can be seen in between each of the white squares representing the experimental data. Given the lack of readily available TK time-course data for *D. magna* there is a necessity for utilisation of the digitisation method to collect, extract, and consolidate publicly available data to fill quantitative data gaps and then compare it to other TK databases containing *D. magna* studies.

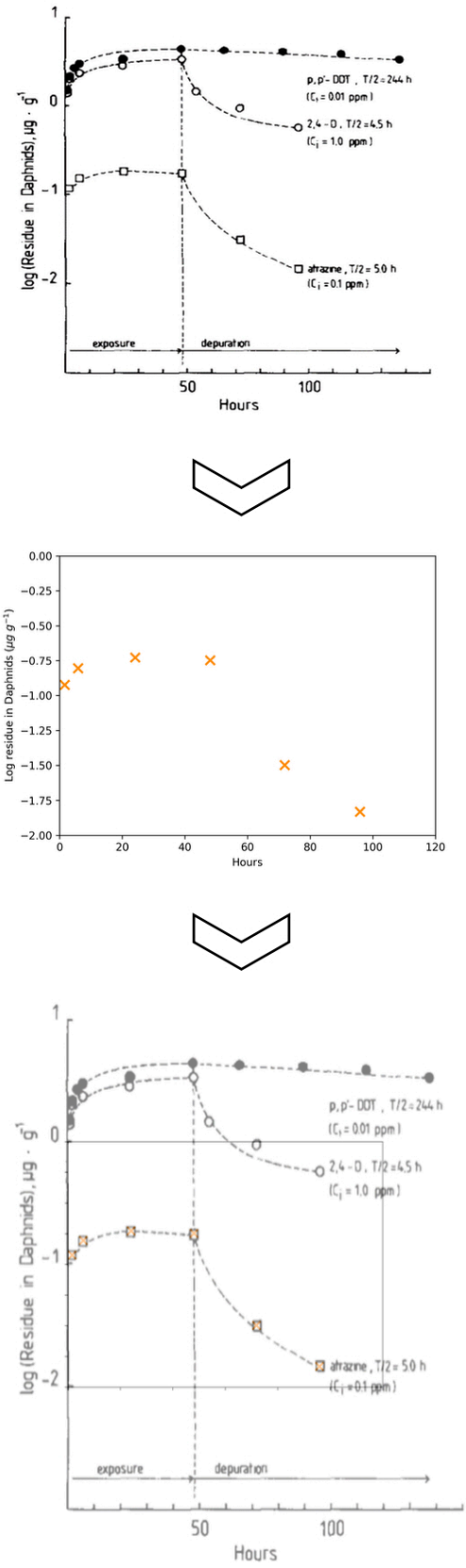

Figure 1.5: Digitisation process for extracting toxicokinetic data from [87].

1.8 Review of statistical modelling for inference and predictions of hierarchical data structures and its application to toxicokinetic data

1.8.1 A review of classical Frequentist approaches and their limitations

The foundations of frequentist theory were laid out by Fisher in the 1920s [97, 63]. These ideas were developed further by Neyman & Pearson whose school of thought is based on physical and ontological probabilities [188, 110, 243]. The frequentist approach to statistical inference has been around for over a century and became the dominant statistical theory in science [93]. Frequentist inference is based on conceptual repetition and the accumulation of data, with probability represented as the expected frequency of an event happening over many experimental repeats [134, 166]. The uncertainty related to unknown parameters is determined by confidence and significance levels in relation to hypothetical repetition [62].

There are two main limitations to the frequentist approach. The first limitation is the interpretation of confidence intervals used to define intervals over either model parameters or predictions. A common misconception is that once a confidence interval has been calculated (X%) that there is X% probability that observed intervals contains the true value. However, X% refers to the proportion of confidence intervals from repeated samples that would contain the true value. Although for a singular confidence interval you do not know if it contains the true value [181]. The second main limitation is the inability to build complex hierarchical models in a non-gaussian setting, corresponding to complex data and information structures like in TK modelling. Generalised linear models attempt to model this property but their interpretation is also non-intuitive e.g. link functions. In a Bayesian setting the distributions of parameters, observations, and error terms are directly defined in the model specification, which can be done for an arbitrarily complex system.

1.8.2 Modern hierarchical Bayesian modelling approaches and their benefits compared with classical frequentist methods

The Bayesian view of probability is that it represents a degree of belief about a given event, so it is possible to assign probability distributions to hypotheses (events) even if it is not possible to repeat the same event [150, 154]. For example, it is possible to assign probabilities to the event "the steady-state internal concentration of *D. magna* is between 10 and 1000 $\mu g\ kg^{-1}$". Moreover, one can assign distributions over parameters in statistical and mathematical models that represent the physical world.

A key benefit of Bayesian thinking is the common sense interpretation of the probability distributions associated with a given event [101]. For example, a Bayesian probability interval for an unknown quantity of interest, like the internal concentration in *D. magna*, can be directly interpreted as having that probability of containing the unknown quantity. Conversely, a frequentist interval confidence interval does not have an intuitive interpretation, which provides a strong motivation for Bayesian approaches [101].

The practical advantages of a Bayesian framework are its flexibility and generality to allow it to cope with complex problems. Another key benefit of Bayesian inference is the explicit and transparent quantification of uncertainty. Furthermore, Bayesian models have an advantage when modelling multi-tiered data structures. These structures can contain many parameters to describe group-level effects, but the hierarchical nature of Bayesian models allows for the use of partial pooling, which allows for variation and similarity information to be shared between groups [91]. This method is useful in reducing effective degrees of freedom of the parameter space to be somewhere within the range starting from the nominal number of groups (no pooling) to a single parameter (complete pooling). This helps to prevent overfitting of the model to the data. It is possible to build complex models in both frequentist and Bayesian frameworks. However, the ability to actually evaluate the posterior distribution of Bayesian models (i.e., make them actually useful) only truly emerged towards the end of the 20th century as advances in Markov chain Monte

Carlo (MCMC) techniques and desktop computing power began to make it possible; for many examples see [101].

These aforementioned advantages of Bayesian modelling have resulted in a recent increase in popularity in TK modelling as it allows the merging of *a priori* information and *in vivo* experimental data, while effectively managing the hierarchical structure of TK data [43]. TK models represent simplifications of complex biological systems with variability in populations, such as weight, biochemical compositions, biotransformation rate, or size. Therefore, to properly evaluate risk requires the incorporation of uncertainty into risk assessment [107]. Other sources of uncertainty come from not knowing parameter values altogether. Uncertainty analysis can implement MCMC methods to calculate data driven posterior distributions for all parameters within a Bayesian approach [107]. Hierarchical population models have been implemented in PBTK to determine uncertainty and to estimate large numbers of parameters [42, 43, 107]. Quantification of uncertainty in TK parameters has become important to regulatory bodies [61, 219].

An example of a hierarchical model for a TK model across multiple time-courses can be seen in Figure 1.6. It has two major components, across time-course variability and within time-course variability. Across time-course variability is a characterisation of how time-course specific parameters vary across the time-course space. Specifically, time-course specific parameters, such as θ, are assumed to be draws from some population distribution described by the mean (Loc) and standard deviation ($Scale$). Loc and standard deviation $Scale$ are hyper parameters and are assigned independent prior distributions. Within each time-course the time-course specific parameters θ deterministically generate the average internal concentration through the TK model. φ represents quantities that cannot be estimates from the data and therefore need to be specified *a priori*, which can be achieved through well-defined priors or as a constant. These can include specific physiological or physiochemical characteristics. Within time-course variability pertains to experimental measurements, which are assumed to be draws from some probability distribution whose

median is the prediction of the TK model and whose variance is defined by a common parameter σ.

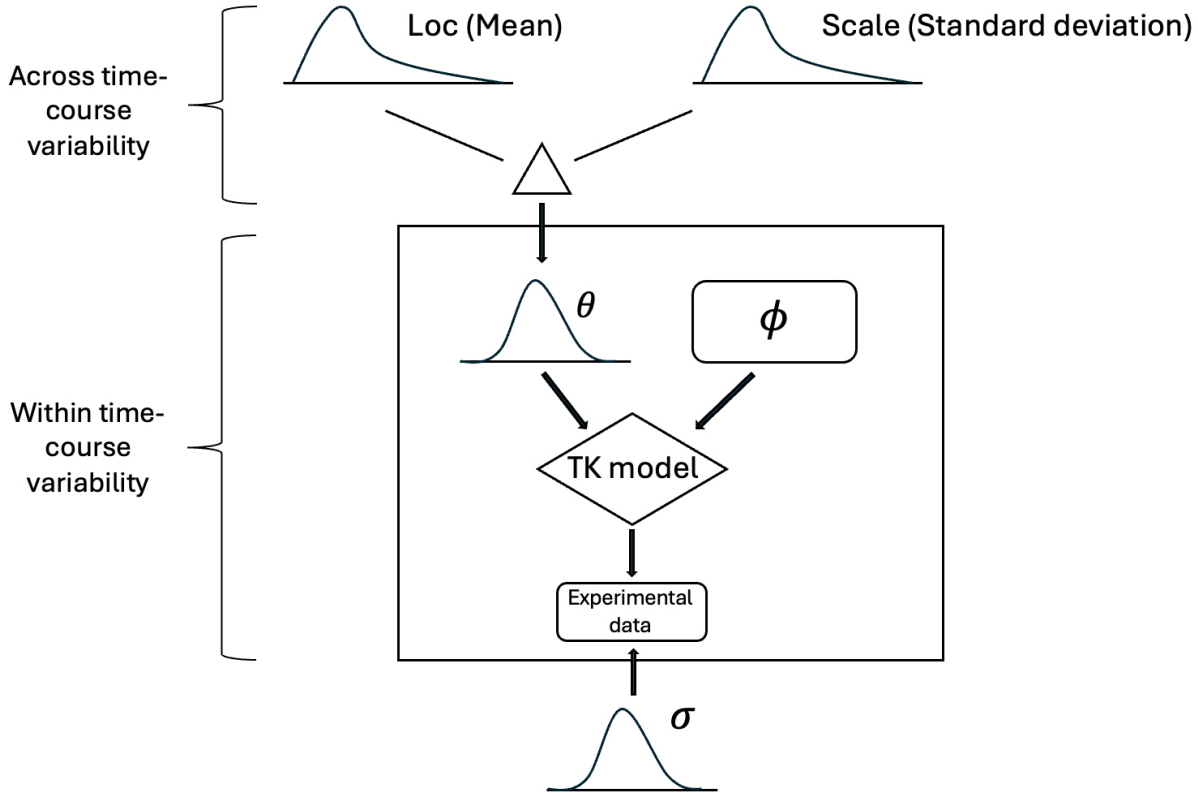

Figure 1.6: Example of a hierarchical model for a toxicokinetic time-course dataset. *Loc* is the mean uncertainty across the time-courses, *Scale* is the variance of the uncertainty across the time-courses, θ is the individual time-course uncertainty, φ represent quantities that are included in the TK model as a well-defined prior or as a constant, TK model is the deterministic toxicokinetic model, experimental data is the time-course data, and σ is the measurement or model error. Figure was adapted from [107].

1.9 Aims & objectives

The overarching aim of the research presented in this thesis is to advance theoretical and probabilistic methods for toxicokinetic predictions in *Daphnia magna* for use in ERA. The importance and limited knowledge of TK models in *D. magna* is illustrated in the previous sections. *D. magna* are a key model organism used in ecotoxicology and safety assessments, especially with the move away from the generation of new animal toxicity data. Even though there are TK models developed for other aquatic species, there is no comprehensive understanding of their application to *D. magna*, which is key to next-generation risk assessment. Any *D. magna* TK models are chemical specific, which restricts their application to the significant number of chemicals needed to be evaluated in ERA. Furthermore, there is insufficient *D. magna* TK data accessible in the literature, however, digitisation methods have shown the potential to collate available data from TK time-course plots. Additionally, the hierarchical data structure of TK data makes Bayesian modelling more applicable as the statistical framework alongside the need for direct quantification of uncertainty that can be applied to ERA. Several key factors that influence TK processes across human and ERA include ionisation of the chemical at environmentally relevant pHs, lipid and protein partitioning, and biotransformation. The impact of these factors on *D. magna* and their integration into TK models needs to be examined to obtain a holistic understanding of TK processes in *D. magna*. Therefore, to achieve the overarching aim of this research, four specific objectives have been developed. The results are presented in chapters 2-5:

1. Develop a proof-of-concept multi-tiered Bayesian approach for the integration of physiochemical properties and toxicokinetic time-course data for *Daphnia magna* with the longer term ambition to deploy this new tool to enable quantification of uncertainty associated with TK predictions in ERA.

2. Investigate the effect of ionisation on toxicokinetic predictions of neutral and ionisable organic chemicals in *Daphnia magna* within a Bayesian framework in order

to establish if the prediction performance can be improved for specific groups of chemicals and therefore establish the applicability domain of the model.

3. Develop a theoretical protein surface-binding model and apply it to *Daphnia magna* toxicokinetic predictions that differs from empirically derived models and enables the simulation of a broad range of external concentration scenarios without the need for *in vivo* data generation.

4. Develop a random forest regression model for predicting the ionisation efficiency values of parent and biotransformation products of industrial chemicals as a first step in a workflow for the semi-quantification of biotransformation products with the long term ambition of integrating biotransformation product concentration predictions into the toxicokinetic modelling of *Daphnia magna*.

CHAPTER 2 : A PROOF-OF-CONCEPT MULTI-TIERED BAYESIAN APPROACH FOR THE INTEGRATION OF PHYSICOCHEMICAL PROPERTIES AND TOXICOKINETIC TIME-COURSE DATA FOR DAPHNIA MAGNA

2.1 Introduction

Due to the large number and variety of chemicals currently in use worldwide, there is a drive towards enhanced safety assessment using the latest scientific knowledge, ensuring robustness and efficiency, as well as avoiding the generation of new *in vivo* animal toxicity data. One proposed strategy is to use *in silico* and *in vitro* methods commonly referred to as NAMs as hazard descriptors [129]. The main advantages foreseen are enhanced mechanistic understanding of toxicological processes leading to robust decisions, higher-throughput, and reduced cost while at the same time answering the societal desire towards a reduction and eventual avoidance in the use of animals for toxicity testing purposes, in line with the 3Rs [49].

The evolving NAMs-based safety assessment framework integrates increased *in silico* and *in vitro* information alongside more traditional alternative approaches (e.g. QSARs) as part of a weight of evidence approach, which is a framework that takes into account different lines of evidence to evaluate the potential risk when a single piece of evidence

is insufficient [119]. This is critical to enabling the use of *in vitro* derived points of departure (PoDs), which are determined from dose-response curves and correspond to a low or no effect concentration, as part of such safety assessments [229]. However, this creates an information mismatch. Traditionally, ERA is on the comparison between the evaluation of eco-toxicological effects, described by the PNEC, which represents the maximum water concentration at which no effect is anticipated in environmental species, with exposure, defined by the PEC, which estimates the concentration reached in the relevant environmental compartment [248]. *In vitro* data, however, generally reflect an internal effect concentration, e.g., tissue or cellular target site (intra and/or extracellular) of the organism. To re-align the two, it is fundamental to have the ability to relate internal and external concentrations, similar to qIVIVE, making the availability of robust TK mass-transfer models fundamental. Over the last few decades several such models have been developed, but they are mostly chemical-specific models, making the need for generic models ever more important, especially for environmentally relevant invertebrate species and those commonly used to support safety assessment of chemicals in the aquatic environment [195].

Generally, TK models for aquatic invertebrates provide estimates for the uptake and elimination rate values without accounting for uncertainty or variability [158]. To characterise uncertainty, a Bayesian statistical framework could be implemented to interpret available TK data more robustly [100, 218]. The application of Bayesian methods in TK modelling of chemicals to estimate parameters in aquatic invertebrates is becoming more common, as illustrated by the examples of the interspecies variation of imidacloprid, generalisable accumulation in benthic invertebrates, and in the evaluation of deltamethrin and its impact on *D. magna* survival [82, 218, 242]. Currently, there are no TK models that incorporate Bayesian inference for the key species, *D. magna* that can also be applied generally to a wide range of chemicals.

D. magna is an aquatic invertebrate species, used extensively as a model organism in

ecotoxicology and safety assessments, due to its ease of use under laboratory conditions and key ecological role, providing a link between primary production and higher trophic levels of the aquatic food web [84, 148, 9]. Yet to date, quantitative TK data for aquatic invertebrates, including *D. magna* is extremely limited, while most have been generated in humans, rodents and fish. As the robustness of developed TK models is evaluated by predictions built and benchmarked against currently available data, the lack of consolidated data is a fundamental obstacle for the development of robust TK and qIVIVE models for invertebrates [105].

The few available TK studies for *D. magna* are typically chemical specific and the developed models are fit to chemical data with limited applicability to other chemicals. The lack of access to the underlying data within these studies is also a constraint, as while it is possible to see the output of the currently available models in the form of time-course internal concentration plots, the exact internal concentration measurements at each time point are often not easily retrievable from the literature. Digitisers are a tool that have enabled the collection of historical quantitative internal concentration measurements from the literature. Digitisers can take images of time-course data plots from studies and calibrate the image axes so that numerical data for each time point can be extracted with high accuracy. Similar approaches have been used in pharmacology and physiologically based pharmacokinetic modelling with proven robustness [269].

The overall aim of this work was to propose a hierarchical Bayesian framework to infer steady-state concentration ratios from *D. magna* TK time-course data. This aim can be divided into three main objectives. The first objective was to collate publicly available *D. magna* TK data and make it readily available to the modelling community as as open-source R package and evaluate its uniqueness compared to other TK databases containing *D. magna* studies. The second objective was to develop a Bayesian framework using the collated TK to infer steady-state concentration ratios and associated uncertainty for each chemical from the uptake time-course data only to allow more data to be included in

the process. The third objective is to present a case study of atrazine to highlight the capabilities of the model to estimate the steady-state concentration ratio at different levels of precision within different data availability scenarios and its application to ERA.

2.2 Materials & methods

2.2.1 Extraction of *Daphnia magna* toxicokinetic data from literature

TK measurements for *D. magna* were extracted from available literature by searching for the following key terms: "*toxicokinetics*", "*bioconcentration*", and"*uptake*" and/or "*elimination*" in *D. magna* using the journal publication repositories Google Scholar (scholar.google.com) and PubMed (pubmed.ncbi.nlm.nih.gov). All literature databases have pros and cons, however, these were chosen, due to familiarity and easy access. The literature searches did not involve truncation or sets of logical operators. The US EPA ECOTOX database and the $MOSAIC_{bioacc}$ database were also evaluated for applicable TK measurements [184, 254]. Results were filtered by analysis of the whole text to ensure that only studies that measured the external and internal concentrations of at least one chemical over a range of time points (uptake and/or elimination) were considered for the following steps. The exposures were split into an uptake and elimination phase for *D. magna* where applicable. Further data quality requirements for a suitable study were:

1. Measurements recorded with units that enable standardisation
2. Chemical concentration measurements from whole body homogenates
3. Organic chemicals only (environmentally relevant)
4. *D. magna* specific study – no ecosystem, mesocosm or food chain studies
5. Static exposures only
6. Wet weight internal concentrations only

7. Study reduces impact of confounding factors, such as dissolved organic carbon or pH variations

This meant that some studies were excluded from the dataset because they did not contain enough measurement information for the unit conversion required. In total, there were 48 time-courses covering 30 chemicals from 17 studies that fit the defined criteria.

However, internal concentration measurements were often only available through plots in figures and were not recorded in tables, which made digitisation necessary for data extraction. Plots were uploaded to the Automeris WebPlotDigitizer (https://apps.automeris.io/wpd/) to manually extract the TK data [225]. A graphical representation of the time-course plot was uploaded to the digitiser tool with the axes calibrated to replicate the studies internal concentration units and values. Finally, each datapoint was added to the dataset by marking each individual point on the plot. The digitiser allows the magnification of specific data points before marking them to ensure accuracy. In respect to quality assurance multiple time-courses were checked by overlaying the plots as seen in Figure 1.5, however, no quantification of error was conducted. A qualitative method of error assessment was undertaken as the digitisation method is well established across pharmacokinetic modelling and there are low errors associated with the process, as highlighted by the digitisation error of biomarker concentrations following inhalation exposure, which was shown to be less than 0.5% [25].

The data from the extraction procedure were then collated into what is hereafter referred to as the *AquaTK* dataset containing information on a total of 48 time-courses for 30 unique chemicals from 17 studies. Chemicals in the dataset covered a broad chemical space including insecticide/herbicides, pharmaceuticals, surfactants, and flame retardants. Metadata for each chemical is available in Appendix A Table A.1.

In addition to internal and external concentration measurements, to support model development, other key experimental data were collected for each study including the K_{ow}, water temperature, water pH, organism age, number of replicates, and wet-weight,

where available. The K_{ow} was taken directly from the study, or alternatively, retrieved from other available databases, e.g., CompTox Chemical Dashboard (https://comptox.epa.gov) or PubChem (https://pubchem.ncbi.nlm.nih.gov) to fill data gaps [132, 253]. The study-collected values were also independently verified against currently available databases.

The internal and external concentration of the *AquaTK* dataset were standardised to $\mu g\ kg^{-1}$ and $\mu g\ L^{-1}$, respectively. Exposure times were standardised to hours and wet weights were converted to kg. The processing and standardisation scripts are captured in the R-package *AquaTK*, hosted on the author's website https://github.com/J-Collins1294/AquaTK.

2.2.2 Evaluating the uniqueness of the AquaTK dataset

The uniqueness of the *AquaTK* dataset was evaluated by comparing the number of studies contained in other available databases, such as the ECOTOX and $MOSAIC_{bioacc}$ databases. The ECOTOX database provides comprehensive ecotoxicology data for over 13,000 terrestrial and aquatic species. It contains single ecotoxicity data from over 52,000 references covering approximately 125,000 unique chemical CAS numbers [199]. Any relevant TK data was found by searching explicitly for *D. magna* accumulation data in the database. This resulted in 1,032 potential datapoints from 125 studies. Additionally, the $MOSAIC_{bioacc}$ database contains a TK (uptake and elimination specific) database with time-course data for 211 uptake and elimination individual datasets for 52 aquatic and terrestrial genera from 56 studies [217]. Additionally, this TK dataset includes 124 unique chemical substances with three exposure routes; water, sediment and soil [217]. However, this database only contains 6 relevant *D. magna* studies, due to the $MOSAIC_{bioacc}$ database needing both uptake and elimination data. Both databases were examined against the criteria set out in the previous section for its applicability to the *AquaTK* dataset.

2.2.3 Statistical analysis of time-course data

Each time course in the *AquaTK* dataset described above is modelled using a mechanistic, one-compartment, diffusion model. This model is then embedded within a statistical model that describes the distribution of experimental measurements conditional on the deterministic component. The joint model is defined within a Bayesian framework where each model parameter is treated as a random variable and the fitting of the model entails calculation of the joint distribution of model parameters after conditioning on the available data.

All of the chemical time-courses are modelled within a single statistical model. To keep track of variables, each unique time-course is assigned an integer index denoted using the subscript i. When i is present on a variable it indicates its association with time-course i.

2.2.4 Mechanistic component – Fickian diffusion model

The choice of mechanistic model is somewhat arbitrary, however, for illustrative purposes it can be assumed the diffusion follows a Fickian-like diffusion process where the rate of change of concentration is proportional to the concentration gradient between distinguishable regions or compartments [94].

Symbolically, $c_i(t)$, represents the median internal concentration of the time-indexed random variable, $C_i(t)$, in each time course i, with units $\mu g/kg$. A Fickian diffusion process corresponds to a first-order differential equation relating the rate of change of the median internal concentration to the internal/external concentration gradient,

$$\frac{\mathrm{d}c_i(t)}{\mathrm{d}t} = C_{\mathrm{water},i}k_{\mathrm{in},i} - k_{\mathrm{out},i}c_i(t) = C_{\mathrm{water},i}K_i k_{\mathrm{out},i} - k_{\mathrm{out},i}c_i(t) \tag{2.1}$$

where $C_{water,i}$, is the external water concentration with units $\mu g/L$, $k_{in,i} = K_i k_{\mathrm{out},i}$, and $k_{out,i}$, are the uptake and elimination rates with units $1/h$, expressed in terms of the

transfer rate ratio, K_i, and t is the exposure time measured in hours. The transfer rates represent the rate at which the chemical absorbs or eliminates.

Equation 2.1 has the particular solution,

$$c_i(t) = C_{\text{water},i} K_i \left(1 - \mathrm{e}^{-k_{\text{out},i} t}\right) \tag{2.2}$$

for the initial condition $c_i(t = 0) = 0$. For long exposures, which are typical in environmental risk assessments, the rate of change of internal concentration is negligible, corresponding to an approximately steady or constant internal concentration,

$$c_i(t) \approx C_{\text{water},i} K_i$$

The ratio of internal to external concentrations is then directly expressed as the ratio of transfer rates. The steady-state concentration ratio is an important quantity in TK modelling as it directly measures the amount of accumulation of a chemical. Figure 2.1 plots each individual time-points internal concentration against the external concentration for all chemical time-courses in the *AquaTK* dataset. This highlights the accumulation of chemicals showing that nearly all chemicals in the dataset accumulate in *D. magna* because the internal concentrations exceed the external concentrations in almost every study at nearly every time-point. Long-term exposure data highlights the effect of octanol-water partition coefficient on the concentration ratio. For example, DDT has a high $\log_{10} K_{ow}$ of 6.91, and after a two-day exposure on *D. magna*, it resulted in a significantly higher internal concentration than external concentration. In the context of ERA, the concentration ratio is the most environmentally relevant quantity as it provides insights into the long-exposure internal concentration of the chemical in the *D. magna*. If the concentration ratio >1 (equivalently the transfer rate ratio) the chemical accumulates in the *D. magna*, however if the ratio <1 it will not accumulate. The parameter $k_{\text{out},i}$ is defined as the *effective elimination rate* of chemical from the organism. The magnitude of

the elimination rate determines the amount of time required to reach an effective steady state. Finally, the parameter $C_{\text{water},i}$, is the concentration of chemical in water and is assumed to be constant for each time course.

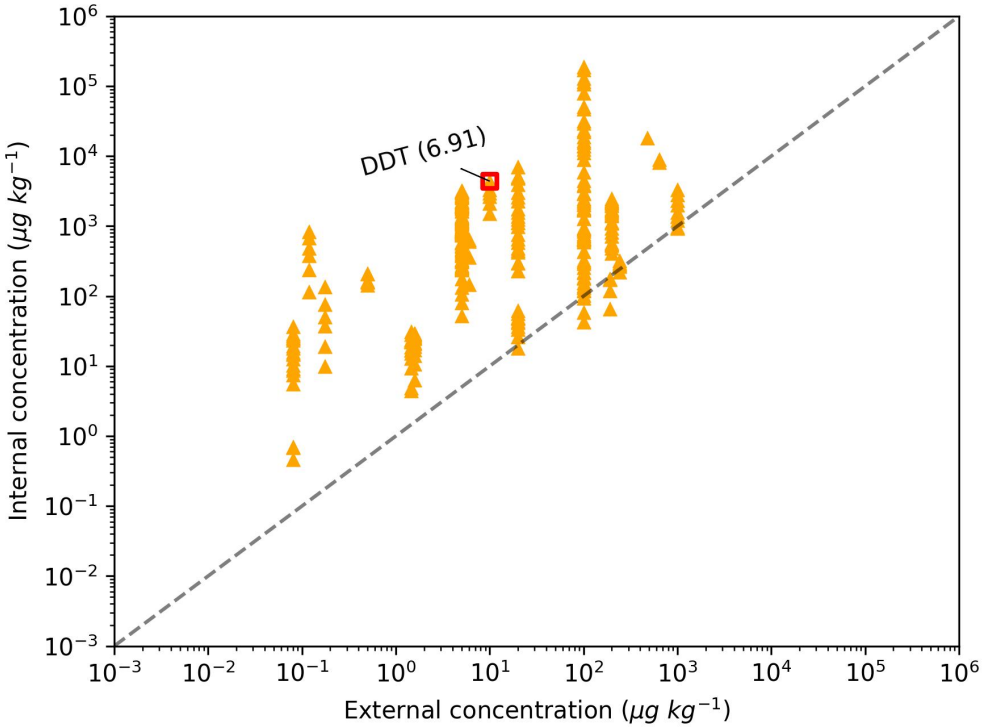

Figure 2.1: Internal concentration ($\mu g\ kg^{-1}$) of the *Daphnia magna* plotted against the external concentration ($\mu g\ kg^{-1}$) in the experiment for each individual time-point across the uptake phase for all 30 chemicals from 17 studies. Note that *D. magna* experiments are time-courses hence the single external concentration and increasing internal concentrations. The dashed bisection highlights chemicals that accumulate in the organism corresponding to a concentration ratio greater than 1.

2.2.5 Statistical component - Bayesian inference

To relate the deterministic model to experimental measurements requires us to specify the assumed sampling distribution of experimental measurements, conditional on a correctly parameterised mechanistic component. Let $y_{i,j}$ be the jth measurement in time course i with corresponding observation time $t_{i,j}$. The natural logarithm of this measurement is modelled as being drawn from a Gaussian distribution, such that

$$\log y_{i,j} \sim \text{Normal}\left(\log\left(c_i\left(t_{i,j}\right)\right), \frac{\sigma}{\sqrt{n_{i,j}}}\right) \tag{2.3}$$

where σ is a universal parameter governing measurement variability and $n_{i,j}$ is the number of replicates that measurement $y_{i,j}$ is an average over. The measurements are converted to a scale with unbounded support, i.e. from positive to a whole set of real numbers by logarithmic transformation. The normal distribution on that scale is the simplest assumption that can be made, assuming finite variance.

Parameter estimates of K_i and $k_{\text{out},i}$ are regularised using an adaptive prior structure. The distribution of $\log_{10} K_i$ across all time courses is assumed to be Gaussian, such that

$$\log_{10} K_i \sim \text{Normal}\left(K_{\text{loc}}, K_{\text{scale}}\right) \tag{2.4}$$

The elimination rate $k_{\text{out},i}$ determines the time for the internal concentration to reach 95% of the steady-state concentration, $t_{95,i}$, via the relationship

$$t_{95,i} = \frac{-\log(0.05)}{k_{\text{out},i}} \approx \frac{3}{k_{\text{out},i}}. \tag{2.5}$$

It is desirable to penalise very fast times to steady state (e.g., of the time-scale seconds or quicker) in the prior structure. An adaptive prior of the form

$$t_{95,i} \sim \text{InverseGamma}\left(t_{\text{shape}}, t_{\text{scale}}\right) \tag{2.6}$$

is assumed for the time to 95% steady state. This achieves the desired penalty since the inverse gamma distribution density rapidly approaches zero in its left tail.

2.2.6 Prior distributions

Prior distributions for σ, $K_{\text{loc}}, K_{\text{scale}}$, t_{shape} and t_{scale} are chosen to be weakly informative. Distribution choices are

$$\sigma \sim \text{HalfNormal}(0, 1),$$
$$K_{\text{loc}} \sim \text{Normal}(2, 1),$$
$$K_{\text{scale}} \sim \text{HalfNormal}(0, 1),$$
$$t_{\text{shape}} \sim \text{HalfNormal}(0, 1),$$
$$t_{\text{scale}} \sim \text{HalfNormal}(0, 10).$$

These prior parameters are fixed and updated conditional on the data, whereas, the $log_{10}K_i$ and $t_{95,i}$ parameters are hierarchal in nature. This means they represent distributions over the prior distribution parameters.

2.2.7 Using K_{ow} and the external concentration to predict the steady-state concentration ratio

The external water concentration and K_{ow} parameters are always available for any given study with robust tools available for predicting the K_{ow} parameter if experimental values are not available [73]. Both the external water concentration and K_{ow} are assumed to significantly affect the rate at which the organism accumulates a given chemical and exist as the predictive component in several popular models [187, 240, 163, 114, 16]. It is possible to include other covariates into the model like the water temperature, the organism wet weight, etc., but this simple two parameter predictive model is sufficient to illustrate its potential use in ERA.

A relatively simple model that includes information from both the K_{ow} parameter and the external water concentration replaces the prior distribution for K_{loc} and expresses it as a linear combination of the log-transformed external water concentration and K_{ow}

$$K_{\text{loc}} = \alpha_K + \beta_{K,C_{\text{water}}} \log_{10} C_{\text{water},i} + \beta_{K,K_{\text{ow}}} \log_{10} K_{\text{ow},i} \tag{2.7}$$

with weakly regularising priors:

$$\begin{aligned} \alpha_K &\sim \text{Normal}(2, 1), \\ \beta_{K,C_{\text{water}}} &\sim \text{Normal}(0, 1), \\ \beta_{K,K_{\text{ow}}} &\sim \text{Normal}(0, 1). \end{aligned}$$

This model is fit to the data, after which, we obtain posterior estimates of α_K, $\beta_{K,C_{\text{water}}}$, $\beta_{K,K_{\text{ow}}}$and K_{scale}. For each time course, the external water concentration and K_{ow} value for the chemical are combined with these parameter estimates, then $K_{\text{predicted},i}$ is sampled from the following equation

$$\log_{10} K_{\text{predicted},i} \sim \text{Normal}\left(\alpha_K + \beta_{K,C_{\text{water}}} \log_{10} C_{\text{water},i} + \beta_{K,K_{\text{ow}}} \log_{10} K_{\text{ow},i}, K_{\text{scale}}\right) \tag{2.8}$$

2.2.8 Computation

Data processing and figure generation was performed using the coding software Python 3.11, with packages Matplotlib 3.8, NumPy 1.26, pandas 2.1, PyStan 3.7, and SciPy 1.11. The Bayesian model was realised numerically in the probabilistic programming language Stan [52]. The posterior distribution of each model was approximated by sampling methods using a MCMC approach. Specifically, the software Stan implements a variant of the Hamiltonian Monte Carlo algorithm called the No-U-turn sampler (NUTS). For each model fit, 10 MCMC chains of length 2,000 iterations were run. The default choice of discarding 10,000 samples as burn-in was made, which discards earlier samples that have not yet converged towards the true values at the start of the MCMC simulation [258]. This left 20,000 samples from which to calculate summary statistics such as posterior

expectations and quantile estimates. The random seed was set to "42" to enable the reproducibility of the sample values.

2.3 Results

2.3.1 Uniqueness of the AquaTK dataset

The *AquaTK* dataset covers 30 unique chemicals covering pharmaceuticals, biocides, surfactants, and flame retardants and includes 48 different time-courses for each distinct chemical and external concentration. The range of $\log_{10} K_{ow}$ values is between -1.19 and 6.91 with a median value of 3.85. The studies in the *AquaTK* dataset were compared against the *D. magna* studies in the US EPA ECOTOX and $MOSAIC_{bioacc}$ databases to evaluate the uniqueness of the dataset. The ECOTOX and $MOSAIC_{bioacc}$ databases contained 125 and 6 *D. magna* relevant studies, respectively. However, only five of the ECOTOX and two of $MOSAIC_{bioacc}$ database studies are also contained within the *AquaTK* dataset. Only one of these MOSAIC TK studies was unique with the other study also contained in the ECOTOX database. Overall, when evaluating the uniqueness of the *AquaTK* dataset, there are 11 unique studies out of the 17 collected when compared with ECOTOX and $MOSAIC_{bioacc}$ databases. The comparisons between the *AquaTK* dataset and databases are visualised in the form of a Venn diagram in Figure 2.2.

2.3.2 Results of Bayesian analysis

A Bayesian hierarchical model was fit to the observed data using Stan to draw parameter samples from the posterior distribution [52]. Posterior estimates of the parameters for σ, $K_{\text{loc}}, K_{\text{scale}}$, t_{shape} and t_{scale} were compared against their respective priors in Figure 2.3. Recall, the parameters here correspond to parameters of distributions of parameters across all time-courses in the dataset. For each parameter, the posterior mass is concentrated away from the tails of the priors, indicating the absence of any prior-posterior conflict that might lead to the assertion that the priors are weakly informative being questioned. The

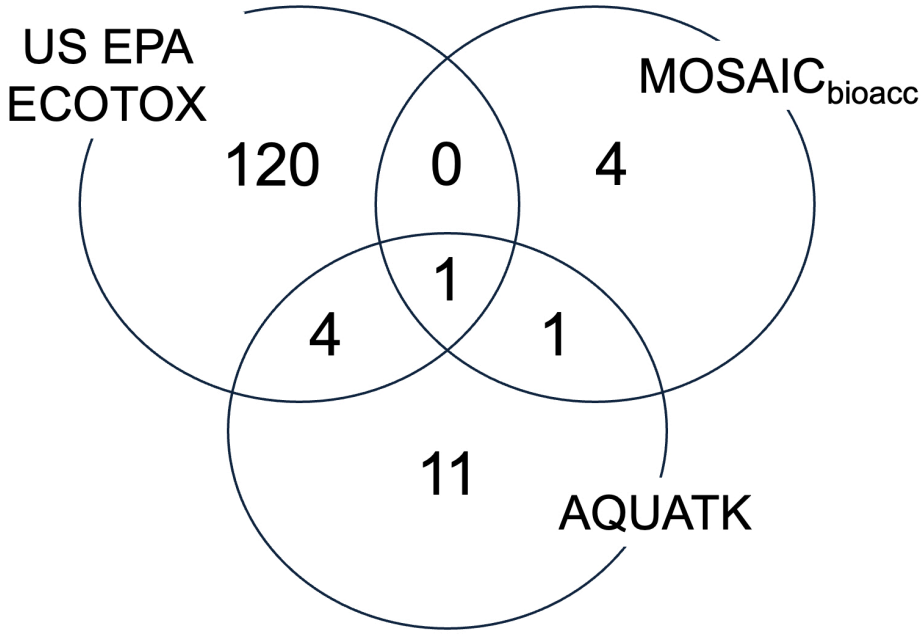

Figure 2.2: Venn diagram showing the overlap of relevant toxicokinetic *Daphnia magna* studies across the US EPA ECOTOX database, $MOSIAC_{bioacc}$ database, and the $AquaTK$ dataset collated within this study. The Venn diagram highlights that the AquaTK dataset has only 6 distinct studies found across the databases and therefore contains 11 unique studies overall.

parameter K_{loc} has a posterior expectation of 1.7 implying that the average steady-state concentration ratio K_i is around 55 (which implies chemicals accumulate significantly across the dataset). The posterior expectation of K_{scale} is close to 1, indicating that the steady-state concentration ratio varies over roughly four orders of magnitude across the dataset. The posterior distributions for t_{shape} and t_{scale} can be pushed through the distribution Inverse Gamma $(t_{\text{shape}}, t_{\text{scale}})$ to generate samples for the time to reach 95% of steady state. This distribution has a median of 7 hours and a mean of 92 hours. The 2.5^{th} and 97.5^{th} percentiles are 1.3 and 168 hours (1 week) respectively. From these estimates it can be concluded that 95% of time-courses take between than 1 hour and a week to reach steady state. Posterior predictive summaries of the mean internal concentration, as a function of time, are presented in Appendix A Figure A.1. There is no evidence from this plot that the model results in a poor fit of the data.

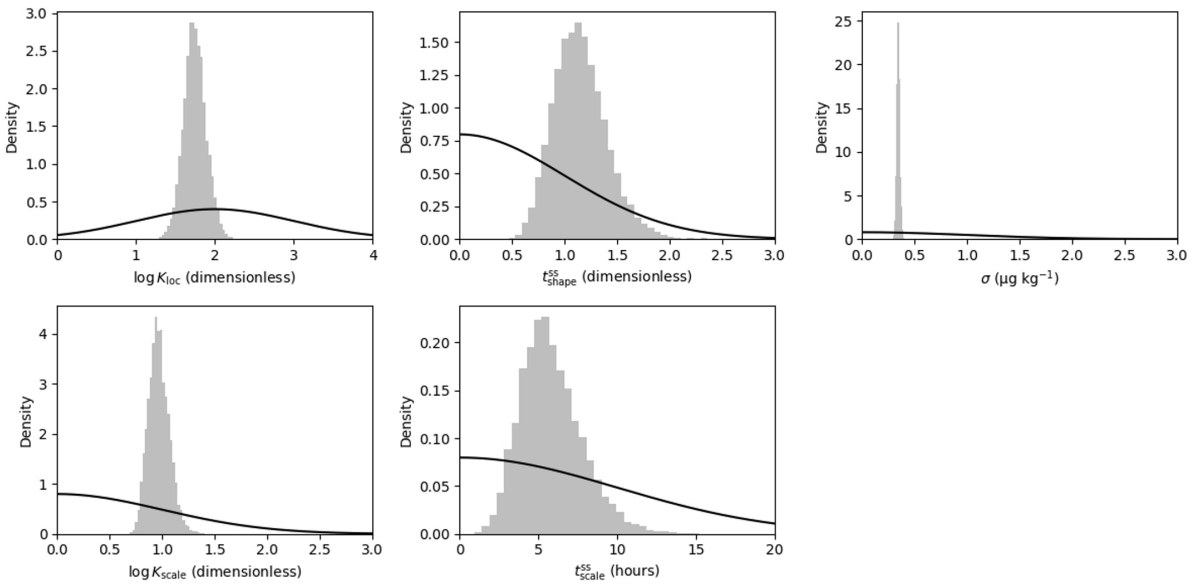

Figure 2.3: Comparison of posterior estimates (grey histograms) for the universal variability parameter (σ), the mean and standard deviation of the steady-state concentration ratio ($K_{\text{loc}}, K_{\text{scale}}$), and the mean and standard deviation of the time to 95% steady-state ($t_{\text{shape}}, t_{\text{scale}}$) versus prior density.

2.3.3 Estimates of the steady-state concentration ratio

For each time course, the estimated parameter K_i is the effective steady-state concentration ratio, and their distributions are represented as intervals in Figure 2.4. The chemicals propiconazole and prochloraz have limited uptake phase time-course data, therefore steady-state concentration ratio estimates for these chemicals are considerably more uncertain. Steady-state concentration ratio estimates for triphenyl phosphate (TPHP) and chlorpyrifos are also relatively uncertain. It is clear from Appendix A Figure A.1 that the time courses for these chemicals have not reached steady state. There are multiple instances where the steady-state concentration ratios for the same chemical differ by orders of magnitude. This variation is likely due to differences in external concentrations. For example, diazinon has steady-state concentration ratio values of 18 and 16 at external concentrations of 1.46 and 1.58 $\mu g\ kg^{-1}$, respectively. In contrast, propranolol has steady-state concentration ratio values of 25 and 7 at external concentrations of 5 and 100 $\mu g\ kg^{-1}$, respectively. This suggests that larger differences in external concentrations of the same chemical are associated with more significant variations in the steady-state

concentration ratio. In fact, the increase in the steady-state concentration ratio is almost proportional to the difference in external concentration, which may be related to a physiological process. Almost all steady-state concentration ratio estimates are greater than 1, indicating each of these chemicals are expected to bioaccumulate.

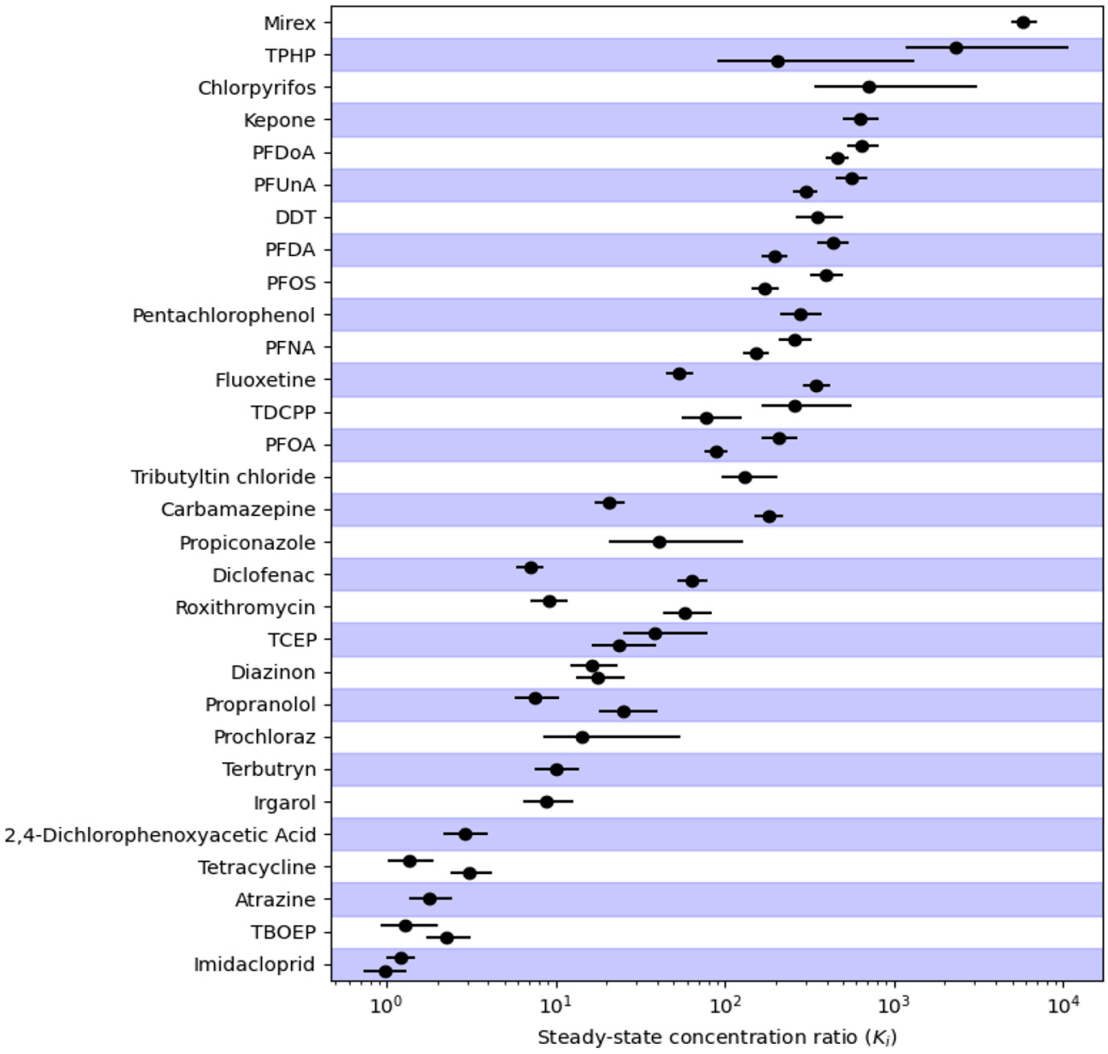

Figure 2.4: Estimates of steady-state concentration ratios (internal concentration / external concentration) for each time course of the 30 chemicals. Estimates are summarised using the distribution median (bullet) and a centred 95% interval. Some chemicals have multiple external concentration scenarios resulting in more than one set of time course data, hence the presence of multiple estimates for some chemicals.

2.3.4 Predicting the steady-state concentration ratio

Equation 2.8 represents the parameter K_{loc} as a linear combination of the log-transformed K_{ow} and external water concentration, which is fitted to the dataset and allows predictions of the steady-state concentration ratio ($K_{predicted,i}$). Estimates of $K_{\text{predicted},i}$ are overlayed on top of estimates of K_i obtained using the time course data in Figure 2.5, left panel. In the top right panel of Figure 2.5, the median estimates of $K_{\text{predicted},i}$ are plotted against the median estimates of K_i. However, as indicated by the width of the red bars, there is relatively large uncertainty in estimates of $K_{\text{predicted},i}$. Nevertheless, the uncertainty characterisation can be leveraged to provide a (mostly reliable) upper bound to K_i when estimated using only K_{ow} and the external water concentration. In the bottom right panel of Figure 2.5 the 95$^{\text{th}}$ percentile of $K_{\text{predicted},i}$ is, for most time courses, greater than the 95$^{\text{th}}$ percentile of K_i. On average, the 95$^{\text{th}}$ percentile of $K_{\text{predicted},i}$ is approximately 8-fold higher than the 95$^{\text{th}}$ percentile of K_i and exceeds it for 46 / 48 (96%) of the estimates. That 96% is close to the nominal percentile of 95% indicates that the uncertainty intervals are well-calibrated. Therefore, the 95$^{\text{th}}$ percentile of $K_{\text{predicted},i}$ may be considered a generally conservative, and potentially useful upper bound for the concentration ratio that may be used within a risk assessment framework.

2.3.5 Estimation of the steady-state concentration ratio in different data scenarios

The predictive model based on the $\log_{10} K_{ow}$ and the external water concentration in the previous section can be used to estimate the effective steady-state concentration ratio with different data scenarios of an ERA framework.

Consider the following scenarios of varying data availability:

- Scenario 1 – only the K_{ow} of the chemical and the external water concentration under consideration are known.

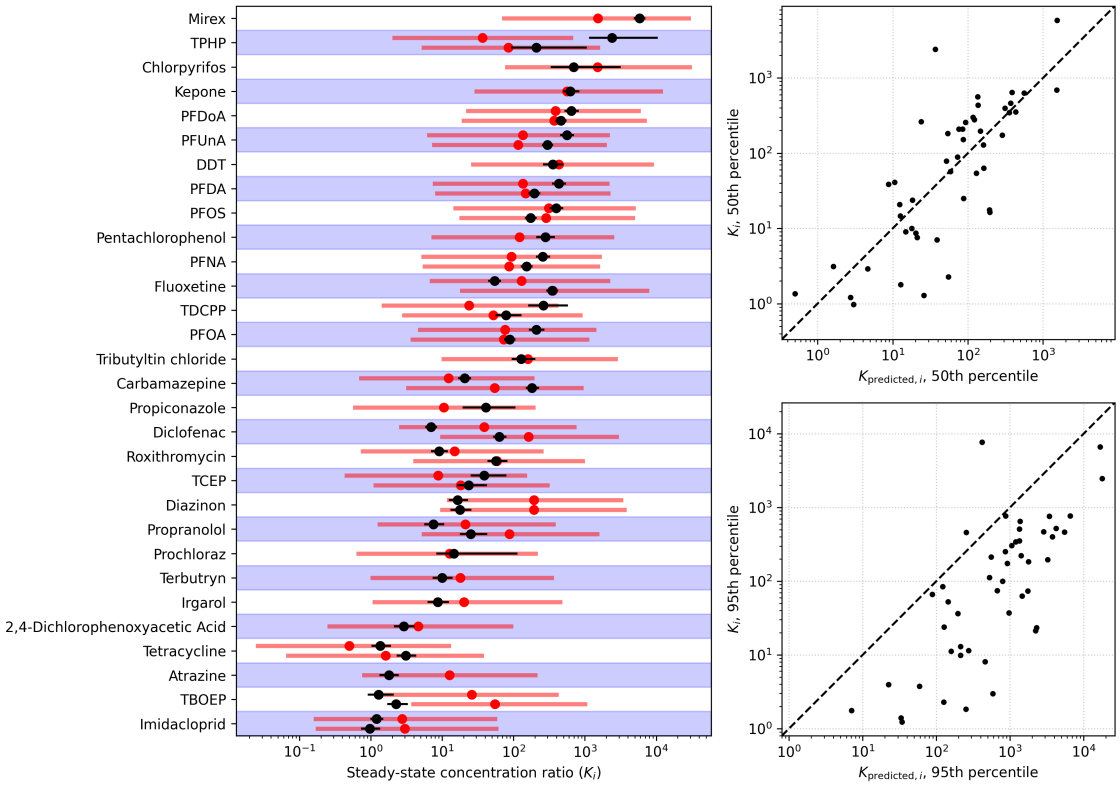

Figure 2.5: Left: Estimates of the steady-state concentration ratios K_i estimated using time course data are plotted in black. Estimates of the predicted steady-state concentration ratio $K_{predicted,i}$, which use only the external water concentration and the octanol-water partition coefficient K_{ow} are shown in red (95% centred interval). Right: 50^{th} percentile of $K_{predicted,i}$ against the 50^{th} of K_i and 95^{th} percentile of $K_{predicted,i}$ is, for most time-courses, greater than the 95^{th} percentile of K_i to highlight the differences in steady-state concentration ratio prediction methods.

- Scenario 2 – internal chemical concentration in *D. magna* at the end of the uptake phase has been measured (single time point TK data).
- Scenario 3 – full time-course data for the internal concentration.

The estimates of the atrazine steady-state concentration ratio within each of the three scenarios are presented in Figure 2.6 on the left-hand side. In the middle column of Figure 2.6, the time-course of internal concentration with 50 and 95% credible intervals are displayed with the additional time points included for each scenario. As previously stated, the concentration ratio is equivalent to the transfer rate ratios for long exposures, therefore plotting the relationship between the concentration ratio (K) and the elimination

rate (k_{out}) can illustrate the importance of the elimination rate in estimating the time to steady state resulting in reduced uncertainty of steady-state concentration ratio estimates. On the right-hand side of Figure 2.6 this relationship is plotted.

In Scenario 1, predictions of the steady-state concentration ratio are informed by the external water concentration and K_{ow} only. These predictions are the most uncertain out of the 3 scenarios with the 95% credible interval ranging two orders of magnitude, however, they require the least amount of data to predict the concentration ratios. In Scenario 1 there is no correlation between parameter estimates as there is no predictive information from time-course data used to estimate K_{out} other than its prior distribution. In Scenario 2, the inputs used in Scenario 1 are supplemented with the measured internal concentration at day-2 (end of uptake phase) time point. The updated steady-state concentration ratio estimate uncertainty is considerably reduced, the 95% credible interval for the steady-state concentration ratio ranges just less than an order of magnitude. The correlation between K and k_{out} indicates the challenge in predicting from a single time-point as the correlation reflects the tradeoff between the various combinations of K and K_{out} that could fit the observed data. For example, a single time-point can be fit to the model using either a small value of K and a fast k_{out} or a slow k_{out} and a high value of K. In Scenario 3, the full time-course is used, reducing the uncertainty in the estimate of the steady-state concentration ratio further still. There is relatively little correlation between K and k_{out} in Scenario 3 because the full time-course data provides sufficient information to constrain estimates of both parameters. It is clear from Figure 2.6 that the more time-course data is inputted into the model the higher the precision of the steady-state concentration ratios, due to the higher level of information that can be extracted.

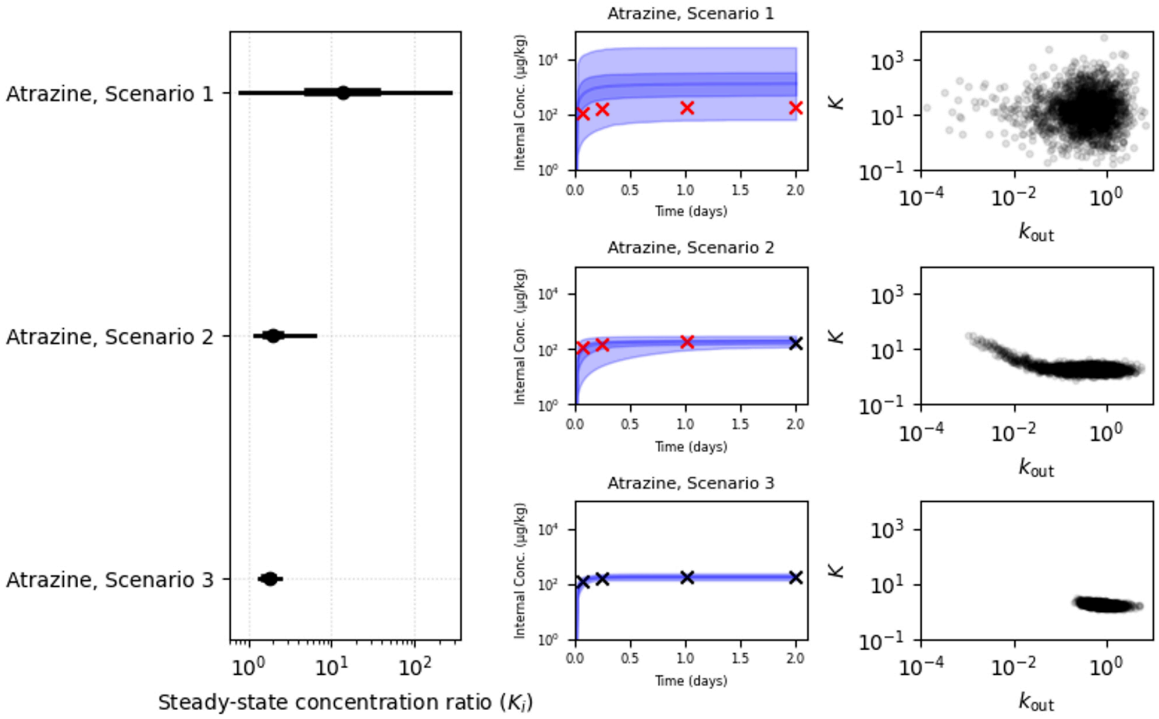

Figure 2.6: Scenario estimation of the steady-state concentration ratio for atrazine. In Scenario 1, only K_{ow} and the external water concentration are used to estimate concentration ratio. In Scenario 2, the internal concentration at the end of the uptake phase is used in addition to K_{ow} and the external water concentration. In Scenario 3, all time course measurements are used. Red is used in the plots in the middle column to indicate that the data point is not used within the scenario. The concentration ratio (K) plotted against the elimination rate (k_{out}) highlights the information obtained from each scenario and its importance to predicting the time to 95% steady-state.

2.4 Discussion

This work has extracted substantial high-quality TK uptake and elimination data for *D. magna* from available literature to fill quantitative data gaps. Moreover, the *AquaTK* dataset has been shown to be unique when compared to the ECOTOX and $MOSAIC_{bioacc}$ databases with 11 unique studies out of 17 encapsulated within the extracted dataset. The collated dataset is the largest time-course database for *D. magna* available, however, ECOTOX has the greatest number of studies for *D. magna,* but these were generally single time points. It was decided that adding singular datapoints from ECOTOX would introduce more covariates, due to their varied study design. Therefore, the decision was made to prioritise the utilisation of high-quality temporal data in this proof-of-concept study, in alignment with Arnot & Gobas. (2006) [17]. Providing high-quality and unique data is essential for the *in silico* modelling community to develop robust and precise models [213]. Furthermore, this signifies the importance of making the data readily available for use by the modelling community with the development of the R package.

Methods for analysing and utilising TK time-course data with other essential data types, such as chemical parameters, are limited. However, Bayesian modelling is the best approach to take advantage of time-course data by allowing the incorporation of prior knowledge, such as the steady-state concentration ratios and time to 95% steady state across the chemicals, handling of varying sampling sizes, incorporation of a given mechanistic model, which can be changed and adapted within the framework, and enables the integration of new time-course data without the need for redefining the model. An advantage of the Bayesian approach is that unlike current ERA it provides a range of values for the effective steady-state concentration ratio, which is important for risk assessment [26]. Previous Bayesian modelling to TK in aquatic invertebrates is limited with recent studies requiring robust time-course datasets, including both uptake and elimination data [218]. This reduces the number of studies that can be input into the model, specifically for *D. magna*, which already has limited TK data available. The Bayesian approach in this study

enables estimations of effective steady-state concentration ratios and the associated uncertainties from singular time-points of the uptake phase only, which means more unique TK study data can be analysed in comparison to other studies. Furthermore, this differs from other Bayesian studies in this space by utilising a predictive component that can predict steady-state concentration ratios in scenarios where TK data is limited (singular time point) or absent. Additionally, the time-course data is treated hierarchically so that inference and prediction information between time-courses can be shared.

Presenting uncertainties and results poses a challenge when incorporating Bayesian modelling into a probability-based risk assessment [69]. However, this study clearly illustrates the potential use of the Bayesian approach for a risk assessor conducting an ERA by providing estimates of the effective steady-state concentration ratios and associated uncertainties with different data availability scenarios. These scenarios range from a prediction of the concentration ratio using only the log-transformed K_{ow} and external concentration, singular time-course data points, and full time-course data. Increasing the number of TK data points input into the model reduces the uncertainty of the steady-state concentration ratio in most cases, due to more information about the uptake and elimination of the chemical being extracted from the data. An important aspect to highlight is that in a few cases the chemical had not reached steady-state meaning there were larger uncertainties. It is possible to leverage these uncertainty values to obtain conservative steady-state concentration ratio estimates for ERA. On average the 95^{th} percentile of the predicted steady-state concentration ratios was 8-fold higher than the concentration ratios from the experimental data for 96% of the data. Therefore, using the 95^{th} percentile in any of the data availability scenarios would be a conservative estimate and essentially avoids, in 95% of cases, underestimating the true concentration ratio values. Using the higher percentile provides a risk assessor with a calibrated uncertainty of the estimate and would be a more appropriate conservative estimate in comparison to using the arbitrary 10x safety factor usually applied to TK data. These conservative estimates can then be compared against known or predicted no effect concentrations to establish the impact of the chemical on

the environment at that external concentration scenario.

There are potential limitations associated with the mechanistic component of the one compartment, partitioning model, where for very long exposure scenarios the model is unable to capture the true uptake dynamics. A few potential sources of uncertainty in the current model and potential data that could be included to reduce these uncertainties are discussed. All experiments are subject to measurement error, associated with measuring the internal and external concentration or the chemical's partitioning coefficient. Previous studies have shown that partitioning measurement errors for some chemicals can be as large as one order of magnitude [30, 126]. It could be argued that the generic TK model used within this study may be too simple and not incorporate key TK processes of the *D. magna*. When analysing directly from time-course data the TK processes are already encompassed within the estimates of the concentration ratio. However, to enable more precise predictions of the concentration ratio the standard deviation (K_{scale}) needs to be reduced by including other predictors which can explain the variance in the steady-state concentration ratio across chemical space. For example, organisms biotransform xenobiotics by creating biotransformation products that often lack the toxic structures of the parent chemical or create products that have lower octanol-water partition coefficients that can be excreted more rapidly [123]. However, the Bayesian predictive model did not consider biotransformation as a predictor, despite evidence to suggest multiple chemicals within this study were in fact biotransformed. The presence of irgarol and terbutryn biotransformation products in *D. magna* have been identified [123]. Without consideration of these BTPs, the overall internal concentration of the parent products would be underestimated. This could potentially impact the ability of the model in this study to successfully predict the uptake dynamics of chemicals that are biotransformed when compared to the experimental data by overestimating the internal concentration of the parent chemical resulting in overpredicted concentration ratio values as well. Therefore, future TK modelling studies should focus on collecting and implementing parent and biotransformation product concentration data and potentially implement available

biotransformation models developed in fish [18].

Figure 2.4 shows that experiments using the same chemical but with different external concentrations have different steady-state concentration suggesting the relationship between the internal and external concentrations are not linearly proportional. Dependence between the steady-state concentration ratio and the external concentration has also been observed in other TK studies including [17]. Further investigation of the phenomenon is difficult, since it would leave underdetermined subpopulations of the data. The proposed "placeholder" model considers only the lipid partitioning, however, it has been argued that the chemical uptake of aquatic invertebrates is dominated by the protein fraction under the condition that the lipid fraction is relatively small (<5% dry weight) [74]. Future work could therefore incorporate protein partitioning into the mechanistic component of the model to improve model prediction accuracy. There is a clear distinction between the steady-state concentration ratio values of a given chemical at different external water concentrations suggesting future work could focus on identifying the relationship between exposure concentration and internal concentration further by exposing the *D. magna* to the same chemical but with more than two external water concentrations. This would help identify whether the nonlinear dependence between the external water concentration and steady-state concentration ratio truly exists. Further development of this idea could include creating a model that accounts for multiple external concentrations while linking the underlying physiological mechanisms in the *D. magna* that impact the steady-state concentration ratio.

Six of the chemicals in the study were PFAS. PFAS are a group of man-made chemicals often persistent in the environment. They generally have low pKa values, which means they are ionised (acidic) at environmentally relevant pH levels [190]. Acidic chemicals can be partially ionised with the neutral and ionic species exhibiting different polarities resulting in pH dependent partitioning [126]. Current approaches only model the neutral fraction, which can result in large errors in predictions for substance groups like PFAS.

Furthermore, tetracycline has multiple pKa values; 3.2, 7.7, and 9.3, meaning it behaves as both an acid and a base with the 3.2 fraction acting like an acid in the environment [131]. Consequently, the predictions are unable to account for the pH of the medium the study was conducted under. To account for the potential impact from ionisable chemicals, future work should focus on including estimates of the distribution ratio ($\log_{10} D_{ow}$) and using the pH of the studies to establish whether the performance is improved, i.e., uncertainty is reduced.

2.5 Conclusions

To conclude, this work has filled quantitative TK data gaps for *D. magna* and provided this in an available R package for use by the modelling community. The uniqueness of the *AquaTK* dataset has been evaluated against key databases and highlights the importance of this data with the increased unique data collection and accessibility essential for *in silico* modelling and the future of quantitative research in this area. This work has provided a Bayesian framework model that can predict the steady-state concentration ratio and its uncertainties from the log-transformed K_{ow} and external water concentration of a chemical. Furthermore, this work demonstrates how steady-state concentration ratio predictions can be generated for different data availability scenarios within a self-contained and consistent framework. These predictions are essential in enabling the implementation of high-throughput NAMs ERA. There is still a need to develop a *D. magna* specific TK model due to the large unexplained uncertainties within the results and the need for a specific aquatic invertebrate model within ERA. Future work could focus on improving the model for the *D. magna* by accounting for the potential ionisation of chemicals at environmentally relevant pHs.

CHAPTER 3 : INVESTIGATING THE EFFECT OF IONISATION ON TOXICOKINETIC PREDICTIONS OF NEUTRAL AND IONISABLE ORGANIC CHEMICALS IN DAPHNIA MAGNA WITHIN A BAYESIAN FRAMEWORK

3.1 Introduction

In Chapter 2 a proof-of-concept Bayesian analysis linking physiochemical properties and *D. magna* time-course data to estimate the steady-state concentration ratio at different levels of precision was presented. This work provided predictive capabilities based on a linear combination of log-transformed covariates, the external concentration and K_{ow}, but without further analysis of the applicability of the model. Despite the substantial step forward in the Bayesian approach there were multiple factors that potentially impact prediction error unaccounted for within the model. Ionisation was not considered, however, multiple chemicals had poor predictions and were likely ionisable at environmentally relevant pH levels, such as PFAS. The Bayesian approach developed was not able to distinguish between neutral and ionisable chemicals, which may have affected overall model performance. Therefore, in this study the ionisation of the chemical is going to be investigated within the Bayesian framework.

Ionisation is the process by which an atom or molecule gains a positive or negative charge

by either losing or gaining electrons, often in conjunction with other chemical changes. The resulting electrically charged atom or molecule is called an ion [168]. Ionisation potential is heavily dependent on external factors, including chemical structure, pH and pKa [222]. A substantial fraction of chemicals in use are ionisable, with over 55% of the chemicals registered under REACH and 60% of active pharmaceutical ingredients bearing a net charge at natural water and physiologically relevant pH's [20, 169]. This fact highlights the importance of considering ionisation potential when evaluating the relationship between concentration of chemicals in natural water bodies and the corresponding internal concentration in aquatic organisms for safety assessment purposes. Given the variation of pH between 6 – 9 in water bodies, and the ionisable nature of so many chemicals in current use, ionisation may substantially affect the partitioning affinity of these substance into organisms, as seen by the shifting of accumulation metrics of many substances with potentially large consequences in terms of their fate and toxicity to aquatic systems [223, 266]. To account for ionisation, the distribution coefficient (D_{ow}) can be used instead of the octanol-water partition coefficient (K_{ow}). D_{ow} is able to account for ionised and un-ionised forms at a given pH [67].

The overall aim of this study was to investigate the effect of ionisation on steady-state concentration ratio predictions in *D. magna*. This is achieved by comparing predictions generated with *in silico* estimates of K_{ow} and D_{ow}. The first objective was to examine the effect of ionisation on the deterministic non-lipid organic matter (NLOM) model steady-state concentration ratio predictions in *D. magna*. The second objective evaluates the impact of ionisation on the Bayesian model steady-state concentration ratio predictions in *D. magna*. The third objective was to test the sensitivity of the Bayesian model prediction parameters to the input data using a cross-validation methodology.

3.2 Materials & methods

3.2.1 *Daphnia magna* toxicokinetic data

TK measurements for *D. magna* were taken from the *AquaTK* R package, which was curated by taking readily available data from the literature and digitisation methods from Chapter 2. This package contained standardised TK data for the internal and external concentrations of 30 organic chemicals across 48 time-courses from 17 unique studies, which fit the high-quality data criteria set out in Chapter 2. The chemical space covered biocides, pharmaceuticals, surfactants, and organophosphates. The package contained other key experimental data important for model development, such as, the octanol-water partition coefficient (K_{ow}), experimental pH, water temperature, organism age, wet weight where applicable, and number of replicates. If the experimental pH was unavailable the average (pH = 7.55) of all the studies pH values was used to fill the data gap as it is comparable with the default pH (7.4) set by commercial software. For more details on the data collation process and the data available please refer to the package details that are available at https://github.com/J-Collins1294/AquaTK.

3.2.2 *Daphnia magna* biochemical composition metadata

The fractional tissue composition data for *D. magna* was obtained from [39], which provides the variation (upper and lower bounds) of the live form of the tissue components over a year-long observation period, summarised in Table 3.1. The measured values were transformed into fractional compositions by assuming the four key components (carbohydrates, lipids, proteins, and water content) make up 100% of the *D. magna*.

3.2.3 *In silico* predictions of partitioning coefficients

To enable comparison between partition coefficients it was decided that *in silico* software ACDLabs (https://www.acdlabs.com) would be used to predict K_{ow} and D_{ow} values to allow consistency across all the chemicals in the dataset. In the previous chap-

Table 3.1: Biochemical composition of *Daphnia magna* from [39] for carbohydrates, lipids, proteins, and water content. Normalised values are calculated assuming the four components make up 100% of the *Daphnia magna*. Normalised median values were used as model inputs for the non-lipid organic matter (NLOM) model.

Component	Lower	Upper	Median	Normalised
Carbohydrates	0.6%	4.0%	2.3%	2.3%
Lipids	0.1%	0.8%	0.45%	0.5%
Proteins	1.3%	5.4%	3.35%	3.4%
Water	86.4%	97.6%	92.0%	93.8%
Total	88.4%	104.2%	98.1%	100%

ter, the $\log_{10} K_{ow}$ values initially from the study and filled any resulting data gaps using available databases, such as, CompTox (https://comptox.epa.gov) or PubChem (https://pubchem.ncbi.nlm.nih.gov). However, this resulted in a combination of experimental and predicted values across the dataset, which would confound results when generating partitioning coefficients for ionisable chemicals. Additionally, many of the chemicals did not have experimental pKa values readily available. Overall, only 9 out of 29 of the unique chemicals had experimental $\log_{10} K_{ow}$ and pKa values available, which meant calculations of D_{ow} would be difficult and only be accessible through predictive commercial software. Therefore, chemical canonical SMILES for each of the 30 chemicals in the dataset were run in ACDLabs PhysChem suite alongside the pH value each of the studies were conducted under and resulted in predictions for $\log_{10} K_{ow}$, $\log_{10} D_{ow}$ and pKa values. D_{ow} is estimated in ACDLabs through predictions of K_{ow}, the predicted pKa value, and the pH of the experimental environment. Tributyltin chloride was excluded from the study because it is an organotin compound, which meant that ACDLabs PhysChem suite was unable to estimate the key chemical descriptors. This resulted in a final dataset with 47 time-courses for 29 chemicals from 16 studies.

Each chemical was either classified as neutral or ionisable to distinguish between the effect of ionisation on these defined groups. A chemical was classified as neutral if the predicted ACDLabs partition coefficients were equal ($K_{ow} = D_{ow}$) and ionisable if the partition

coefficients were not equal ($K_{ow} < D_{ow}$). A detailed overview of each time-courses chemical, source ID, external concentration, pH, ACDLabs $\log_{10} K_{ow}$ prediction, ACDLabs $\log_{10} D_{ow}$ prediction, and ionisable or neutral classification can be seen in Appendix B Table B.1.

3.2.4 Deterministic modelling - NLOM model

Previous work has highlighted improved model performance by accounting for ionisation with D_{ow} [13]. However, this was done for empirical fish data only. Therefore, the first step was to evaluate the performance of a state-of-the-art deterministic model on the *D. magna* dataset. The NLOM model developed by Arnot & Gobas. (2004) was chosen because of its extensive use within the ERA community [16]. The NLOM model was developed to predict the partitioning of organic chemicals between an organism and its environment through lipid, NLOM, and water where each of these parts can absorb and store the chemical [16]. Therefore, a daphnia-water partition coefficient (D_{bw}) can be calculated as,

$$D_{bw} = v_{lb} \cdot K_{ow} + v_{nb} \cdot \beta \cdot K_{ow} + v_{wb} \tag{3.1}$$

where v_{lb} is the lipid fraction, K_{ow} is the octanol-water partitioning coefficient, v_{nb} is the NLOM fraction, β is a proportionality constant corresponding to a sorption capacity and v_{wb} is the water content of the *D. magna.* Given the *D. magna* biochemical composition data in Table 3.1 a value of 0.005 can be assigned to v_{lb}. v_{nb} can be calculated by adding together the NLOM (protein and carbohydrate content), which equals 0.057. Other NLOM such as bio-ash were not considered significant for modelling [145]. Comparable to Arnot & Gobas. (2004) β is given as 0.035 [16]. v_{wb} is taken from Table 3.1 and is given as 0.938. The NLOM model performance was evaluated against inferred steady-state concentration ratio values from the Bayesian method developed in the previous chapter. Geometric mean residual error for the NLOM model were also calculated for comparison.

3.2.5 Statistical analysis of the time-course data

In Chapter 2, the TK time-course data in *AquaTK* dataset is modelled using a hierarchical Bayesian model. The average internal concentration of a chemical in *D. magna* is modelled using a one-compartment Fickian-like diffusion model involving two parameters to capture uptake and elimination rates. Parameter estimates are assumed to be time-course specific. The model assumes adaptive prior distributions to describe the variation in uptake and eliminations rates across the time-course space in the *AquaTK* dataset. Prior distributions act as a regularisation device when estimating these parameters from the dataset. This model is used to estimate the steady-state concentration ratio for each time-course. For convenience, this model, from here-on referred to as the reference model, is defined in Appendix B Table B.2

This model was extended in the previous chapter to utilise K_{ow} and external water concentration as a predictor for the steady-state concentration ratio in *D. magna*. For the purposes of investigating ionisability with respect to prediction errors, and furthermore to enable fairer comparisons against the NLOM model (which assumes steady-state concentration ratios are invariant to changes in the external water concentration), this model is modified slightly such that external water concentration is no longer used as a predictor. The modified model is defined in Appendix B Table B.3 and from here-on is referred to as the predictive model. A key difference between the NLOM and Bayesian predictive model is that the NLOM model is derived from fugacity concepts with a physical interpretation, while the Bayesian predictive model is data driven. The priors in the Bayesian inference model encode the mechanistic assumptions. This encoded information should be consistent with the NLOM model.

3.2.6 Computation

The coding software Python 3.11, with packages Matplotlib 3.8, NumPy 1.26, pandas 2.1, PyStan 3.7, and SciPy 1.11 were used for data processing and figure generation.

The probabilistic programming language Stan was used to develop the Bayesian model [52]. The software Stan uses a Hamiltonian Monte Carlo algorithm called the No-U-turn sampler (NUTS) to estimate posterior distributions. 10 Monte Carlo Markov Chains with 2,000 iterations were run for each of the models fit the data with 10,000 samples as burn-in leaving 20,000 samples.

3.3 Results

3.3.1 Effect of accounting for ionisation on NLOM predictions

The NLOM model has been shown to improve steady-state concentration ratio predictions when accounting for ionisation with D_{ow} in fish [13]. However, it was important to establish whether the model performed similarly on the *D. magna* dataset. From Figure 3.1, it can be seen that NLOM model predictions for ionisable chemicals using K_{ow} as an input result in an overprediction, on average. Conversely, NLOM predictions with D_{ow} consistently underpredict the steady-state concentration ratio for ionisable chemicals, even though the average magnitude of the prediction error is smaller. Predictions for neutral chemicals are comparatively unbiased; there are examples of both under and over predictions of similar magnitude.

The strong bias towards overpredictions using K_{ow} and underpredictions using D_{ow} presents an interesting opportunity to drastically improve prediction performance of the NLOM model. If both predictions with K_{ow} and D_{ow} are generated, then the geometric mean of predictions is considerably closer to the target steady state, on average, than either prediction in isolation.

3.3.2 Effect of accounting for ionisation on Bayesian model predictions

The Bayesian predictive model defined in Appendix B Table B.3 was used twice to generate predictions for the steady-state concentration ratio. Firstly, the model was trained

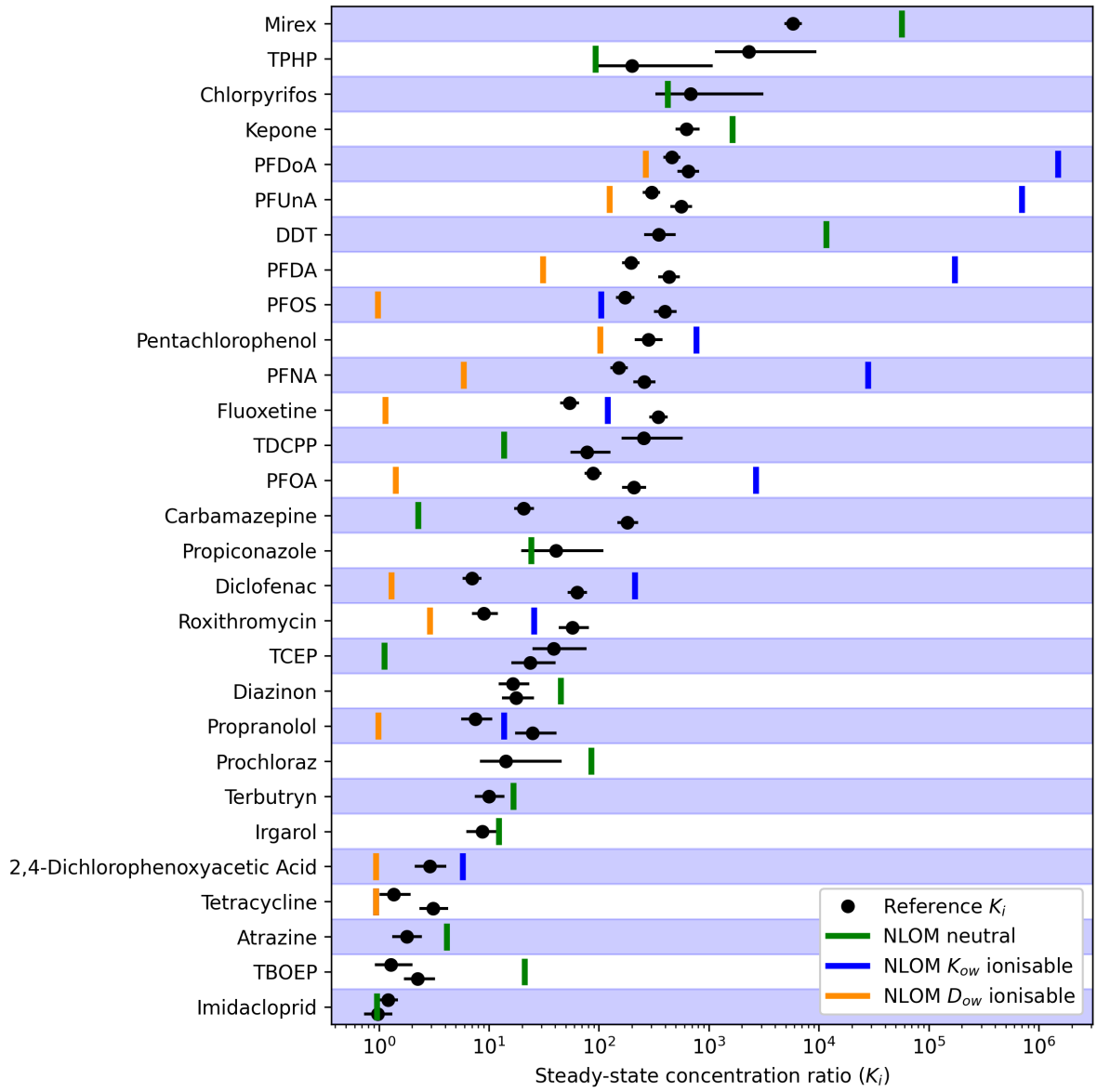

Figure 3.1: Estimates of steady-state concentration ratios K_i from the reference model are plotted in black with the bullet indicating the median and the horizontal line a centred 95% interval. The non-lipid organic matter model (NLOM) predictions using K_{ow} are plotted as blue vertical bars and orange bars indicated predictions using D_{ow}. Predictions for neutral chemicals are green ($K_{ow} = D_{ow}$). Some chemicals have two time-course datasets, with distinct external concentrations. Distinct steady-state concentration ratios are estimated from these data using reference. NLOM model predictions of the steady-state concentration ratio are independent of the external concentration, hence a single prediction is made for all time-course datasets for each chemical.

using K_{ow} as the data input P_i, followed secondly, by using D_{ow}. Median estimates of K_i^{pred} are plotted against the reference model estimates of K_i in Figure 3.2.

Predictions of the steady-state concentration ratio are generally within the range of values estimated using the reference model. The predictive model uses regression to relate the partitioning coefficient ($\log_{10} P$) to the steady-state concentration ratio, which results in shrinkage to the mean. This naturally results in under-prediction of the largest steady-state ratios in the dataset and overprediction of the smallest. This approach results in soft guarantees for the magnitude of the prediction error; the inferred prediction error standard deviation (summarised by the model parameter K_{scale}) is smaller than the standard deviation of the steady-state concentration ratio across the chemical space in the *AquaTK* dataset (see left plot, Figure 3.3). Furthermore, estimates of K_{scale} are smaller when using K_{ow} as the chemical partitioning coefficient than when using D_{ow}. That is, the model infers a stronger association between K_{ow} and the steady-state concentration ratio than with D_{ow}. Within the *AquaTK* dataset, not accounting for ionisation using the method to compute D_{ow} allows for more accurate predictions, on average.

Interestingly, prediction errors (as measured using absolute fold-error) for time-courses corresponding to the ionisable subset of chemicals are on average smaller than prediction errors for the neutral subset when using the predictive model. This occurs when using either K_{ow} or D_{ow} as the partitioning coefficient, see Table 3.2. However, this is not true for the NLOM model, which produces much more accurate predictions for neutral chemicals. Switching from K_{ow} to D_{ow} within the NLOM model reduces the average prediction error, but the average remains considerably higher than the average error for the Bayesian predictive model.

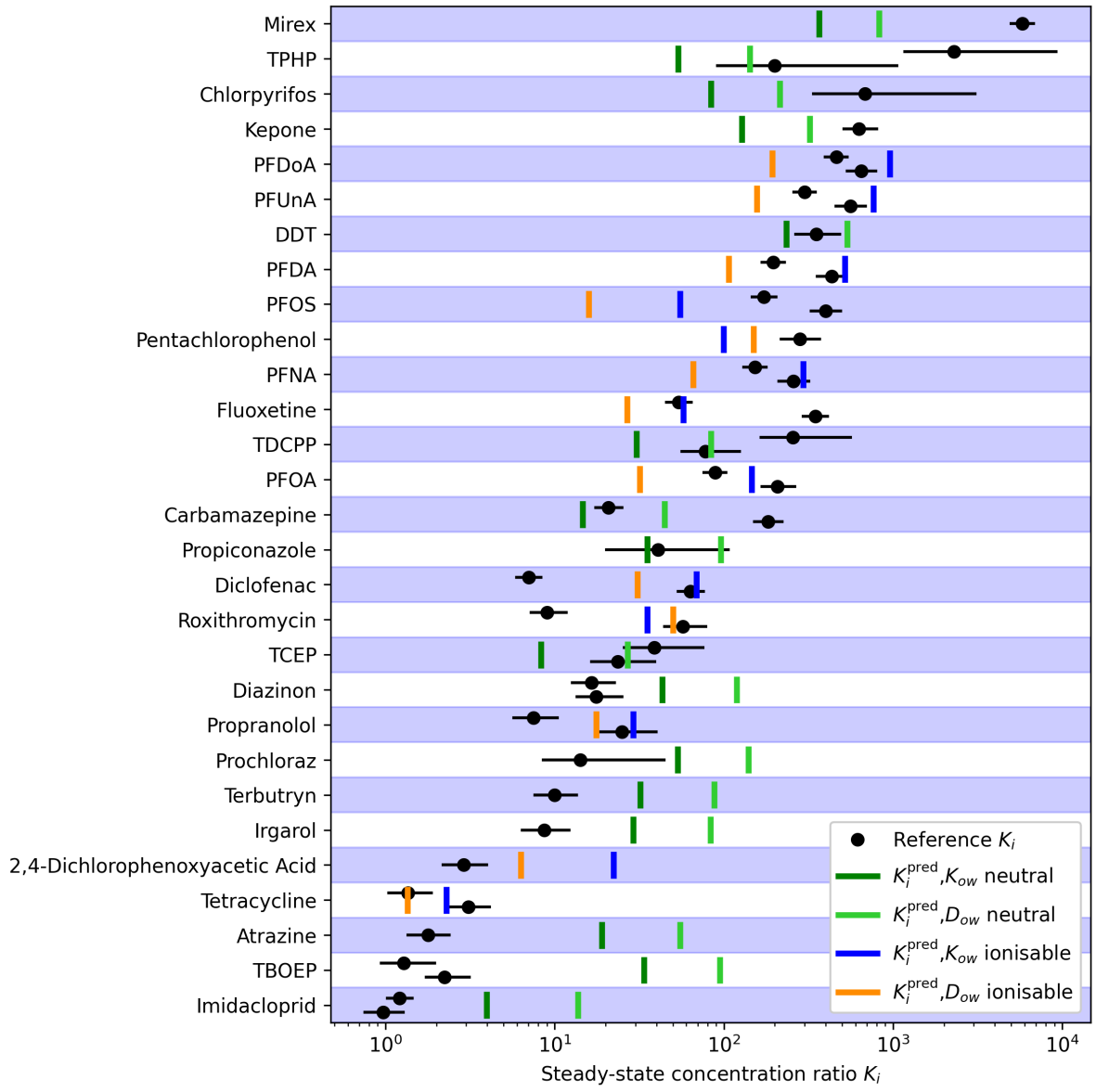

Figure 3.2: Estimates of steady-state concentration ratios K_i from the reference model are plotted in black with the bullet indicating the median and the horizontal line a centred 95% interval. Median estimates of predictions from the Bayesian predictive model using K_{ow} are plotted as blue vertical bars and orange bars indicated median predictions using D_{ow}. Unlike the NLOM model, for which predictions for neutral chemicals (green) are the same ($K_{ow} = D_{ow}$), the Bayesian predictive model parameters change for each input, hence predictions differ for neutral chemicals despite chemical-specific information being the same.

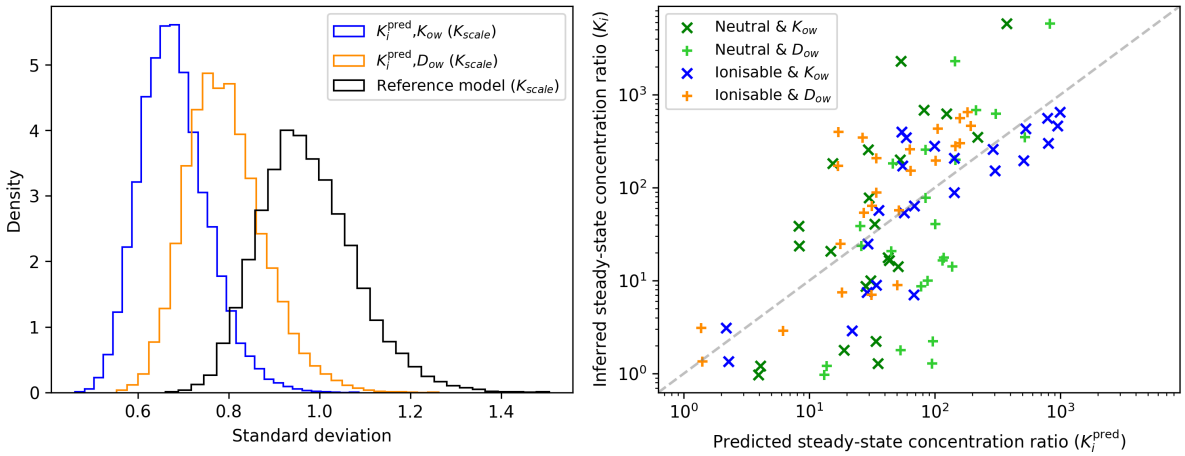

Figure 3.3: Left: Posterior estimates of the parameter K_{scale} estimated from the reference model (black) and the predictive model when trained with K_{ow} (blue) and D_{ow} (orange). Right: Median estimates of K_i^{pred} estimates from the predictive model (x-axis) plotted against median estimates of K_i (y-axis) from the reference model.

Table 3.2: Chemicals from the *AquaTK* dataset grouped by neutral and ionisable classification. Source ID, external concentration, inferred steady-state concentration from the reference model (K_i) and prediction errors, as measured by median absolute fold error (K_i^{pred}/K_i) for both NLOM and Bayesian predictive models with K_{ow} and D_{ow} as the partitioning coefficient.

Neutral chemicals	**Source ID**	**External concentration** ($\mu g\ L^{-1}$)	**Reference** K_i	K_i^{pred}, K_{ow} **fold error**	K_i^{pred}, D_{ow} **fold error**	**NLOM & K_{ow} fold error**	**NLOM & D_{ow} fold error**
Atrazine	d12	100	1.8	11	32	2.3	2.3
Carbamazepine	d08	5	180	12	4.1	80	80
Carbamazepine	d08	100	21	1.4	2.2	9.1	9.1
Chlorpyrifos	d05	0.08	680	8.3	3.3	1.6	1.6
DDT	d12	10	350	1.6	1.7	33	33
Diazinon	d02	1.5	18	2.4	6.6	2.6	2.6
Diazinon	d02	1.6	16	2.6	7.1	2.7	2.7
Imidacloprid	d04	190	0.97	4.1	14	1.1	1.1
Imidacloprid	d04	240	1.2	3.3	11	1.3	1.3
Irgarol	d03	200	8.7	3.2	9.2	1.4	1.4

Kepone	d15	0.17	630	5	2	2.6	2.6
Mirex	d15	0.12	5800	16	7	9.8	9.8
Prochloraz	d13	640	14	3.6	9.7	6	6
Propiconazole	d13	480	41	1.4	2.4	1.7	1.7
TBOEP	d14	20	2.2	15	42	9.4	9.4
TBOEP	d14	100	1.3	26	73	16	16
TCEP	d14	20	24	2.9	1.3	21	21
TCEP	d14	100	39	4.8	1.5	34	34
TDCPP	d14	20	78	2.7	1.3	5.7	5.7
TDCPP	d14	100	260	8.9	3.1	19	19
Terbutryn	d03	200	10	3.1	8.8	1.7	1.7
TPHP	d14	20	200	3.8	1.6	2.1	2.1
TPHP	d14	100	2300	44	17	25	25
Neutral geometric mean of fold error				5.1	5.7	5.8	5.8
Ionisable chemicals							
2,4-Dichlorophenoxyacetic Acid	d12	1000	2.9	7.6	2.2	2	3.1
Diclofenac	d09	5	63	1.2	2	3.3	49
Diclofenac	d09	100	7	9.7	4.5	30	5.4
Fluoxetine	d10	0.5	350	6.1	13	2.9	310
Fluoxetine	d10	5	54	1.2	2	2.2	47
Pentachlorophenol	d16	20	280	2.8	1.9	2.7	2.8
PFDA	d06	5	430	1.4	4.1	400	14
PFDA	d06	5	200	2.6	1.9	880	6.3
PFDoA	d06	5	650	1.6	3.4	2300	2.4
PFDoA	d06	5	460	2.1	2.4	3200	1.7
PFNA	d06	5	260	1.3	4	110	44
PFNA	d06	5	150	1.9	2.4	180	26

PFOA	d06	5	210	1.4	6.2	13	150
PFOA	d06	5	89	1.6	2.6	30	63
PFOS	d06	5	400	7.3	24	3.8	410
PFOS	d06	5	170	3.1	11	1.7	180
PFUnA	d06	5	560	1.5	3.6	1200	4.5
PFUnA	d06	5	300	2.6	1.9	2300	2.4
Propranolol	d01	5	25	1.3	1.5	1.8	25
Propranolol	d01	100	7.5	3.9	2.4	1.8	7.6
Roxithromycin	d01	5	57	1.6	1.3	2.2	20
Roxithromycin	d01	100	9	4	5.5	2.9	3.1
Tetracycline	d11	100	3.1	1.5	2.4	3.3	3.3
Tetracycline	d11	1000	1.4	1.7	1.7	1.4	1.4
Ionisable geometric mean of fold error				2.4	3.2	21	14
Overall geometric mean of fold errors				3.4	4.3	11	9

3.3.3 Sensitivity of predictive model parameters to the input dataset

The Bayesian predictive model infers its predictive performance from the *AquaTK* dataset. The robustness of these estimates for the purposes of assuming similar predictive performance on novel chemicals is therefore of interest. In this section a cross-validation methodology was used to assess how estimates of K_{scale} change following perturbations to the data used to train the model. The *AquaTK* dataset was partitioned by source ID and a leave-one-out method was used to re-estimate K_{scale} following iterative removal of all data corresponding to each source ID. Figure 3.4 highlights that the expected value of K_{scale} is similar in all cross-validation folds except for when source ID d14 is removed. Source ID d14 contains four organophosphate chemicals for two different external concentration exposures that were classified as neutral at the experimental pH. The two largest prediction errors for the predictive model using K_{ow} as the partition coefficient were TPHP and TBEOP, while the largest prediction error for the predictive model using

D_{ow} as partition coefficient were the two exposures of TBEOP. These chemicals were from source ID d14.

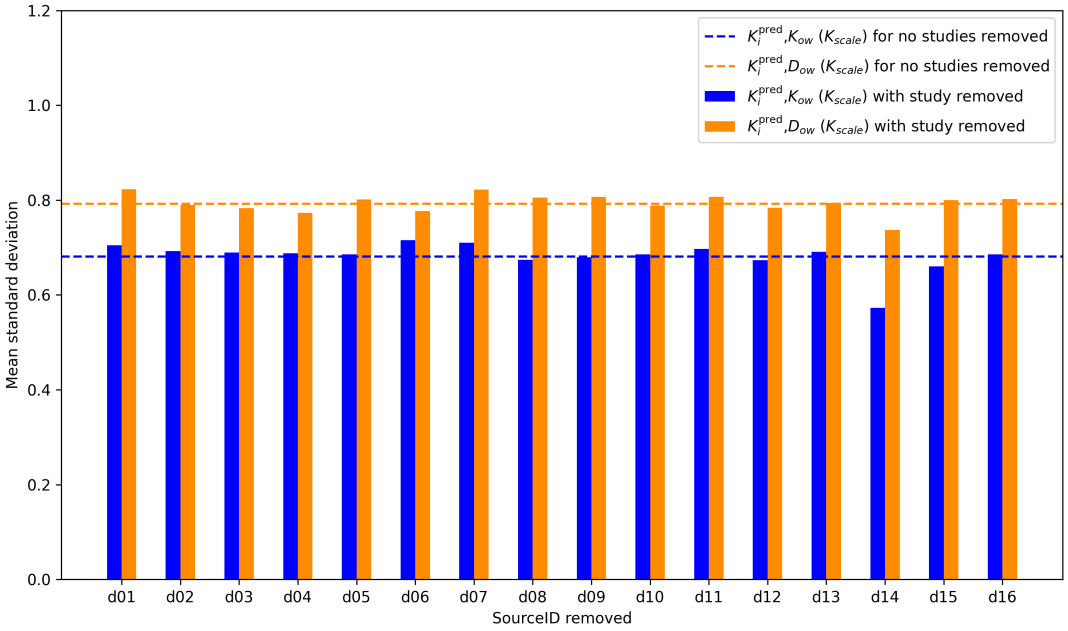

Figure 3.4: The mean residual standard deviation for each study (Source ID) removed using a leave-one-out method was plotted for the Bayesian predictive models optimised on K_{ow} (blue) and D_{ow} (orange). The blue and orange dotted lines represent the mean residual standard deviation when no studies were removed for K_{ow} and D_{ow}, respectively.

3.4 Discussion

This study examines the effect of accounting for ionisation when generating predictions of the steady-state concentration ratio in *D. magna*. Ionisation was accounted for in the NLOM and Bayesian predictive models through the use of the distribution coefficient (D_{ow}), which accounts for the pH the experiment is conducted under and the predicted pKa value of the chemical. The magnitude of errors for the NLOM model improves when accounting for ionisation, however, this leads to consistent underpredictions for the ionisable chemicals. Conversely, the magnitude of errors for the Bayesian predictive model gets worse when accounting for ionisation using D_{ow}. Overall, the prediction error in the Bayesian predictive models were smaller than those in the NLOM model. The Bayesian model predictions of the steady-state concentration ratio regress towards the mean of

the those inferred within the (reference model) *AquaTK* dataset. This has the effect of bounding the prediction error variance by the variance of steady-state concentration ratios in the *AquaTK* dataset.

The sensitivity of the predictive model parameters to the input dataset was tested using a cross-validation methodology to indicate whether predictive performance is consistent and can therefore be assumed for chemicals not found within the *AquaTK* dataset. This highlighted that the prediction error for all cross-validation folds was comparable apart from source ID d14. This study contained data for four neutral organophosphates, with TPHP and TBEOP the worst predicted chemicals for the predictive model with K_{ow} and D_{ow}, respectively. Overall, it is clear from the results that the worst predictions of the Bayesian predictive model were for the neutral chemicals, which suggests improvements to predicting this specific group of chemicals is crucial to improving prediction performance.

In the context of ERA, an estimate that is greater than the experimental steady-state concentration ratio would enable a conservative evaluation of the risk. Therefore, the use of the NLOM model with D_{ow} for *D. magna* may not be applicable, due to the underpredictions of the ionisable chemicals. Additionally, even though the magnitude of the prediction errors is smaller relative to NLOM, the Bayesian model significantly overpredicts the lowest steady-state concentration ratios and underpredicts the highest steady-state concentration ratios. Alternatively, a more conservative estimate for the Bayesian predictive model could be achieved by using the 95% credible interval values rather than the median, which has been applied in the previous chapter. However, the degree of conservativeness is correlated with the magnitude of the steady-state ratio being estimated, where the fold-difference of the overprediction is expected to increase for the smallest steady-state concentration ratios.

There are some limitations of this work in terms of the data itself and the modelling approach. Part of the prediction error may be as a result of inaccuracies in predictions

of K_{ow} and D_{ow} values. In future work the model predictions could be tested with alternative D_{ow} prediction methodologies through other prediction software. Alternatively, the Henderson-Hasselbach equation could be used to predict pKa values to establish the ionised and neutral forms of chemicals [222]. However, true utility of this method requires experimental data of the pH and pKa value. In this study only 9 of the 29 unique chemicals had experimental pKa and $\log_{10} K_{ow}$ values meaning there would have been a reduced dataset and reduced confidence in the interpretation of the results.

A more advanced mechanistic model using an adapted version of the NLOM was developed for ionogenic organic chemicals in fish using empirical steady-state concentration data [13]. This study was able to characterise the partitioning of ionisable chemicals to phospholipids by separating the overall lipid fraction into phospholipids and storage lipids [13]. This is important because ionisable chemicals are expected to accumulate in these phospholipids and even protein fractions but not the storage lipids [38]. However, in the scope of this study there is a lack of available empirical data for aquatic invertebrates [14]. While under current knowledge there is scattered data on the separation of *D. magna* lipid fractions into phospholipids / membrane and storage lipids. Studies have evaluated lipidomic profiles of the major lipid classes of *D. magna* [124]. However, the results were not presented as fractions relative to the overall biochemical composition, which means they cannot be used in the adapted version of NLOM. Future work could potentially conduct experiments to obtain these different fraction percentages and conduct a similar assessment set out in the study on the impact of membrane-water partitioning on the prediction error.

The Bayesian prediction model could be improved by accounting for other variables or processes. In this study the predictive model only has one covariate, the partition coefficient, but there is evidence that suggests external concentration needs to be considered. In Chapter 2 a log-transformed external concentration and K_{ow} was implemented in a linear combination to predict steady-state concentration ratios. However, the external concentration was not included in the predictive model of this study to isolate the effect

of accounting for ionisation in the estimate of the partitioning coefficient. Examining the results highlights multiple examples where external concentration may have an impact on the residual errors of the model. For example, diclofenac and roxithromycin had different predictions under the same experimental conditions but at different external concentrations. Furthermore, TPHP with an external concentration of 100 $\mu g\ kg^{-1}$ was ranked the worst on the predictive model optimised on K_{ow} with a fold error of 44. However, TPHP exposed at 20 $\mu g\ kg^{-1}$ had a fold error of 3.8 for the predictive model optimised on K_{ow}. As a neutral chemical this would suggest that variations in the prediction error are at least impacted by the external concentration. Future work could use the Bayesian framework to investigate the impact of the external concentration on steady-state concentration ratio predictions.

Another important process that is not encompassed within the Bayesian framework is protein binding. Out of the 24 ionisable chemical time-courses 8 of the 10 worst prediction errors based on fold error from the best predictive model (K_i^{pred}, K_{ow}) were either pharmaceuticals or PFAS. Pharmaceuticals are known to compete for binding sites of plasma proteins, which impacts the free concentration of the molecule and influences toxicokinetic processes [237]. Research suggests that PFAS interact with a number of different proteins within organisms [189]. Protein binding has a relationship with ionisation where anionic forms of acidic chemicals are less likely to be eliminated by an organism, due to being bound to plasma proteins [170]. Additionally, the pKa value that is a key component of ionisation predictions is known to play an integral role in protein binding [265]. Accounting for ionisation only resulted in some improvement of the fold errors for the NLOM model. Considerable work has been done to account for sorption of chemicals to NLOM including proteins within the proportionality constant of the NLOM model [103, 16, 17, 74]. However, this is through retrospective interpolation where experimental data is continually added to a database to widen its domain of applicability and then confidence intervals are added so that 95% of the values are within the desired range leading to an updated proportionality constant. However, for the prediction of novel chemicals

without experimental data it may be more advantageous to develop a generalisable protein binding model from first principles for use in ERA.

Multiple neutral chemicals had high prediction error within the Bayesian framework. For example, mirex was the third worst predicted chemical for the predictive model optimised on K_{ow}. Mirex is described as a chlorine box where chlorine has replaced every hydrogen atom attached to a carbon [66]. This degree of halogenation may impact metabolism as there is a relationship between halogenation and decreased metabolism [66]. In terms of the Bayesian model presented here there is potential to add structurally relevant parameters to help inform predictions. In the case of mirex it could be number of halogenated atoms. Further important structural features could be added, such as hydrogen bonding. Hydrogen bonding can affect chemical solubility and partitioning into different phases with QSARs simplifying hydrogen bonding into the number of hydrogen bond acceptors and donors [268].

3.5 Conclusions

To conclude, this work highlights the applicability of using a Bayesian framework to evaluate the effect of different parameters on model performance. The effect of ionisation was accounted for by using D_{ow} over K_{ow} and resulted in improved NLOM prediction errors but had a negative impact on the Bayesian model prediction error. However, overall the geometric mean prediction error of *D. magna* steady-state concentration ratio predictions were lowest in the Bayesian model with either partition coefficient suggesting the Bayesian model is able to capture the TK profile of *D. magna* more accurately than the deterministic NLOM model. It was clear from the results that the neutral chemicals in the dataset primarily had the largest prediction errors in the Bayesian model. Therefore, future work could work on implementing key chemical structures, such as hydrogen bonding, that may help explain the uptake dynamics of the chemical more effectively. Additionally, there were large differences in predictions error for the same chemical at different external con-

centration exposures, which suggests the external concentration needs to be considered in future TK models.

CHAPTER 4 : A THEORETICAL PROTEIN SURFACE-BINDING MODEL AND ITS APPLICATION TO DAPHNIA MAGNA TOXICOKINETIC PREDICTIONS

4.1 Introduction

Current general models utilised in TK modelling of uptake and elimination data are mostly based on empirical studies in fish. The empirical origin of the steady-state equation relating the concentration ratio with the product of the partition coefficient and lipid fraction was [187] with further contributions from [259], [163], [36], and [177]. A generalisable model for a range of aquatic organisms was developed that considers the partitioning of chemicals from the water to the lipid and non-lipid parts of the organism [16]. This model was developed for fish from empirical data and is known as the non-lipid organic matter (NLOM) model. In this model there is a proportionality constant that expresses the sorption capacity of non-lipid components relative to octanol [16]. Many studies have empirically updated this constant [103, 16, 17, 74]. This method works through retrospective interpolation where experimental data is added to the modelled dataset to expand the domain of applicability. This limits the application of these models to chemicals in the dataset and requires more experimental data to update the model parameters each time. Therefore, this makes applying these models to novel chemicals especially difficult.

It was argued that in fish the chemical affinity to non-lipid organic matter was lower compared to lipid [16]. However, it could impact partitioning in organisms with smaller lipid fractions, such as phytoplankton or aquatic invertebrates [16]. Further research suggests that if the lipid fraction makes up to less than 5% of the dry weight organic content, the absorption capacity of an organism will be dominated by the protein fraction [74]. Dry weight estimates of the lipid content in *D. magna* were 4.98% [39, 40]. This suggests that the protein fraction and the binding of chemicals into this specific component needs to be accounted for to understand the TK processes in *D. magna*.

Protein binding is an important aspect of pharmacokinetics, due to the significant number of proteins in plasma and the numerous drugs that can bind to them, impacting the effective drug concentration [41]. Chemical structure and physiochemical properties, such as lipophilicity represented by K_{ow}, the pKa that represents acidic or neutral characteristics, and hydrogen bonding can influence protein binding [265]. There are two types of protein binding, specific and non-specific. Specific binding can be described as a high affinity, saturable binding of a chemical to its target receptor resulting in a pharmacological response. Non-specific binding is described as non-saturable and low affinity binding with endogenous proteins without pharmacological response [74, 41]. Previous protein binding modelling has taken a stochastic approach where attachment to a binding site is based on a probability as a function of already attached binding sites [121]. Other popular approaches to quantifying protein partitioning is the creation of absorption isotherms, such as the Brunauer-Emmett Teller, Langmuir or Freundlich isotherm models [151]. These isotherms describe the protein absorbed to a surface as a function of the solution concentration [151]. Each isotherm has different applications and requirements for its utilisation. The Langmuir isotherm is based on the assumption that absorption occurs on a monolayer with no interactions between absorbed chemicals. Conversely, the Freundlich isotherm describes monolayer absorption but for heterogeneous surfaces with unique absorption sites in terms of absorption rates and energies. Finally, the Brunauer-Emmett-Teller isotherm differs from the other isotherms by considering multilayer absorption on different areas of a

surface [151]. These isotherms can reflect the binding of chemicals to different biochemical components. For example, it has been illustrated for the environmentally prevalent PFAS that their binding to proteins can be expressed in the form of the Freundlich isotherm suggesting that it is an absorption process rather than a linear process [271].

There is a lack of knowledge in terms of how protein binding theory can be applied to aquatic invertebrates, due to the lack of specific experimental data. However, a recent study of the impact of thiacloprid on the aquatic invertebrate *G. pulex* was the first to demonstrate a mechanistic link between protein binding and TK modelling [215]. Through parameterisation of the experimental data a receptor-binding TK model was developed. Irreversible binding of thiacloprid to membrane proteins was shown to cause elimination resistance, in addition to a dependence on external concentration, due to a maximum binding capacity and lack of elimination from the membrane protein compartment [215]. Additionally, it has been shown in previous chapters that the inferred steady-state concentration ratios from *D. magna* time-course data are dependent on the external concentration, which could be a result of chemical protein binding.

The overall aim of this study was to develop a theoretical model for chemical partitioning in lipids and surface-binding in proteins for determining a theoretical physiological upper bound for steady-state concentration ratios in *D. magna*. This can be divided into three key objectives. The first objective was to develop a theoretical framework for decomposition analysis of chemical partitioning and derive a new protein surface-binding (PSB) model. The second objective was to use the PSB model to define an external concentration and protein dependent theoretical upper bound for use in environmental risk assessment and evaluate the bound on historical *in vivo* measurements. The third objective was to use the PSB model to predict steady-state concentration ratios and compare the prediction performance with the NLOM model and on available aquatic invertebrate experimental data with several external concentration exposure scenarios.

4.2 Materials & methods

4.2.1 *Daphnia magna* toxicokinetic data

The R package *AquaTK* provided *D. magna* TK measurements. This package contains digitised internal and external concentration data from available literature for 48 time-courses across 30 organic chemicals from 17 studies. The data covers a broad chemical space with pharmaceuticals, biocides, organophosphates, and surfactants. The package also contained other key experimental data for model development with more details available at https://github.com/J-Collins1294/AquaTK.

4.2.2 *Daphnia magna* biochemical composition data

The biochemical composition of *D. magna* over a year-long observation period with upper and lower bounds of the live form taken from [39] is summarised in Table 4.1. Normalisation of the data was conducted assuming the carbohydrates, lipids, proteins, and water content made up 100% of the *D. magna*.

Table 4.1: Biochemical composition of *Daphnia magna* from [39] for carbohydrates, lipids, proteins, and water content. Normalised values are calculated assuming the four components make up 100% of the *Daphnia magna*.

Component	Lower	Upper	Median	Normalised
Carbohydrates	0.6%	4.0%	2.3%	2.3%
Lipids	0.1%	0.8%	0.45%	0.5%
Proteins	1.3%	5.4%	3.35%	3.4%
Water	86.4%	97.6%	92.0%	93.8%
Total	88.4%	104.2%	98.1%	100%

4.2.3 *In silico* partition coefficient predictions

Previous work has shown an improvement on steady-state concentration ratio predictions using the dissociation constant (D_{ow}) over the octanol-water partition coefficient (K_{ow})

[13]. The ACDLabs PhysChem suite commercial software was used to generate *in silico* predictions of $\log_{10} K_{ow}$ and $\log_{10} D_{ow}$. ACDLabs PhysChem suite predictions of K_{ow} have been shown to correlate with experimental values for a range of chemical classes with a coefficient of 0.992 [208]. It is one of the consensus best software for predicting pKa values [268]. D_{ow} is estimated by inputting the predicted K_{ow}, the predicted pKa value, and the pH of the experimental study. The pH was collected from the studies otherwise a mean pH of 7.55 from across the studies was applied. 30 chemical canonical SMILES were run through the ACDLabs PhysChem suite to obtain D_{ow} predictions. However, the time-course data for tributyltin chloride [92] was removed as the software is unable to predict organotin compounds. Therefore, the dataset was reduced to 47-time courses for 29 chemicals from 16 studies.

The chemicals within the dataset were labelled as either neutral or ionisable. A chemical was labeled neutral when,

$$D_{ow} = K_{ow} \tag{4.1}$$

and ionisable when,

$$D_{ow} < K_{ow} \tag{4.2}$$

A table containing a detailed overview of the 47 time-courses including the chemical name, source ID, external concentration, pH, ACD $\log_{10} K_{ow}$, and $\log_{10} D_{ow}$ predictions, and ionisable or neutral classification can be seen in Appendix B Table B.1.

4.2.4 Fickian-diffusion steady-state internal concentration

The exchange of a chemical between its environment, in this case water, and the organism can be described by a Fickian-like diffusion first-order ordinary differential equation (ODE),

$$\frac{dy(t)}{dt} = k_{in} \cdot y_w - k_{out} \cdot y(t) \tag{4.3}$$

where $y(t)$ is the organisms internal concentration at time t of exposure, and the y_w is

the external concentration, which is assumed to be constant throughout the experiment. k_{in} and k_{out} are transfer rates modulating the amount of chemical coming in and coming out of the organism.

Equation 4.3 has the following particular solution,

$$y(t) = y_w \cdot k_r(1 - e^{-t \cdot k_{out}}) \tag{4.4}$$

for an initial concentration of zero, $y(t = 0) = 0$, where $k_r = k_{in}/k_{out}$ is the transfer rate ratio. The concentration ratio, $z(t)$, is an important quantity for measuring the accumulation of a chemical in the organism.

$$z(t) = \frac{y(t)}{yw} = k_r(1 - e^{-t \cdot k_{out}}) \tag{4.5}$$

When $z(t) > 1$ the chemical accumulates, however, when $z(t) \leq 1$ then the chemical does not accumulate. Long exposures scenarios are the most relevant for ERA. For long exposures $t \cdot k_{out} \gg 0$, the internal concentration reaches an effective steady-state,

$$z_{ss} \approx \lim_{t \to \infty} z(t) = k_r; \tag{4.6}$$

For the special case where it is assumed the chemical mass contributions can be linearly decomposed into lipid-water, and protein contributions,

$$y = y^{(I)} + y^{(II)} \tag{4.7}$$

where y is the internal concentration, $y^{(I)}$ is the contribution to the internal concentration by the lipid-water partitioning, and $y^{(II)}$ is the protein surface-binding process contribution. It is important to specify that these two processes are modelled independently and their contributions can be added together. It is important to note that interaction terms have been avoided because of empirical evidence from the previous chapters that high-

light the dependency of steady-state concentration ratios on external concentration. It is possible that interaction terms exist but these are difficult to validate empirically without an experiment similar to Raths et al. (2023) [215]. Such validation would require an extensive number of chemicals across a range of K_{ow} values.

4.2.5 Generalised partitioning model

In Appendix C.1.3 a general expression is derived for the steady-state concentration ratio in terms of mass fractions ϕ_i and partitioning coefficients K_i,

$$z_{ss} = \frac{y_{ss}}{y_w} = \sum_i \phi_i \cdot K_i + 1 - \sum_i \phi_i \tag{4.8}$$

Note, the partitioning coefficients K_i are general and do not represent any particular compartment partition, that is they can be set to K_{ow}, D_{ow}, or their respective estimates from prediction software.

The partitioning model in Equation 4.8 predicts a steady-state concentration ratio,

$$z_{ss} = \nu \cdot K + \varphi \cdot K_{pw} + 1 - \nu - \varphi \tag{4.9}$$

where ν is the lipid fraction, K represents the lipid partitioning coefficient, which is K_{ow} for neutral chemicals and D_{ow} for ionisable chemicals, φ is the protein fraction, K_{pw} is the protein-water partition coefficient. For highly lipophilic chemicals $K \gg 1$ and negligible protein binding $K_{pw} \ll 1$, this model describes the empirical observations of [187]. Further studies by Debruyn & Gobas. (2007) derive estimates of the protein partitioning $K_{pw}(K)$ in terms of K [74],

$$K_{pw}(K) = \beta \cdot K \tag{4.10}$$

where $K = K_{ow}$, and β is empirically derived from experimental data. In this study $K = D_{ow}$ since the dataset for benchmarking contains multiple ionisable chemicals at environmentally relevant pH levels.

4.2.6 Protein surface-binding model

Current protein partitioning models cannot account for external concentration as a variable, however, recent studies in *D. magna* [68] and *G. pulex* [215] have shown that steady-state concentration ratios depend significantly on the external concentration. To address this limitation, a toy surface-binding model is proposed that relates the organism embedding water density, assumed equal to the external water concentration, to the protein surface density (Appendix C.2.4). The model predicts a protein surface density upper bound conditional on the chemical's protein affinity defined as,

$$P := \frac{p_b}{p_u} \tag{4.11}$$

where p_b is the probability the molecule binds, and p_u is the probability the molecules unbinds.

In Appendix C.2.6, it is shown that for chemicals with a protein affinity $P < (\frac{1}{\rho} - 1)$, the density on the protein surface cannot exceed the density of the enclosing volume (ρ), which can be interpreted as the internal water concentration. For $1 < P < (\frac{1}{\rho} - 1)$, the upper bound of the protein binding process can be defined as,

$$y^{(II)} = \rho_p^{(S)} \leq \left(\frac{1}{\varphi}\right)^{\frac{2}{3}} \sqrt{y_w} \tag{4.12}$$

where $\rho_p^{(S)}$ is the protein-chemical surface density, φ is the protein fraction, and y_w is the nominal external water concentration. The derivation can be found in Appendix C.2.7.

Combining the lipid partitioning and the protein surface-binding process yields the following concentration upper bound,

$$z_{ss} = \frac{y}{y_w} = \frac{y^{(I)}}{y_w} + \frac{y^{(II)}}{y_w} \leq \nu \cdot K + 1 - \nu + \left(\frac{1}{\varphi}\right)^{\frac{2}{3}} \frac{1}{\sqrt{y_w}} \tag{4.13}$$

The lipid-binding model can be interpreted as an upper bound on the amount of chemical in a two-phase system under the assumption there are no barriers limiting diffusion across a membrane. In reality the amount of chemical that binds to the lipid fraction would be less than or equal to the ideal two-phase system estimated by K_{ow}. The protein surface-binding model can also be interpreted as an upper bound due to the assumptions of the model, specifically that the chemical binds uninhibited to the surface. An expression for the amount of chemical bound to the protein surface is derived in Appendix C.2.

4.2.7 Determination of PSB model theoretical upper bound

A theoretical upper bound of z_{ss} can be determined for the proposed PSB model given the protein fraction set to the lower bound ($\varphi = 0.013$) and lipid fraction is set to the upper bound ($\nu = 0.008$) from Table 4.1, across a range of partition coefficient values. Since the steady-state concentration ratio is now a function of the external concentration, the upper bound must be computed for different values of y_w. The theoretical upper bound is compared to inferred median steady-state concentration ratios estimated from *D. magna* time course data using Bayesian methods from Chapter 3. To show that the theoretical upper bound holds for different exposure scenarios the upper bounds were calculated with external concentrations of 0.08 $\mu g\ kg^{-1}$ (minimum external concentration from the dataset), 1 $\mu g\ kg^{-1}$, 10 $\mu g\ kg^{-1}$, and 100 $\mu g\ kg^{-1}$, which made up the four external concentration scenarios. All steady-state concentration ratios were included in the first scenario with the minimum external concentration upper bound (≥ 0.08). However, for each of the subsequent scenarios the steady-state concentration ratio was only included if the external concentration was greater than the external concentration used in the theoretical upper bound calculation. The theoretical lower bound was calculated by using the upper bound for the protein fraction ($\varphi = 0.054$) and the lower bound of the lipid fraction ($\nu = 0.001$) from Table 4.1 with the maximum external concentration (1000 $\mu g\ kg^{-1}$) across a range of K_i values.

4.2.8 Benchmarking against the NLOM model

To evaluate the performance of the PSB model it was benchmarked against the non-lipid organic matter (NLOM) model developed by [16], due to its extensive use for ERA within the exposure modelling community. The NLOM model predicts the partitioning of chemicals from the environment into the organism into lipid, non-lipid organic matter and water, where each of these components can absorb and store chemical [16]. The mass decomposition analysis that was used to determine the lipid and non-lipid correction in the PSB model can be applied to the lipid, protein, and complement to obtain the NLOM model. The relationship between the steady-state concentration ratio and the partitioning by composition can be described as follows,

$$\frac{y}{y_w} = \nu \cdot K + \varphi \cdot \beta \cdot K + 1 - \nu - \varphi \tag{4.14}$$

where ν is the lipid fraction, K is the predicted distribution coefficient (D_{ow}), φ is the protein fraction, and $1 - \nu - \varphi$ is the NLOM fraction. β is a proportionality constant with a value of 0.035. The water content variable encompassed in the NLOM model is accounted for by assuming everything other than the proteins or lipids is water $1 - \nu - \varphi$. Median normalised biochemical composition data from Table 4.1 can be used for the lipid fraction ($\nu = 0.005$) and the protein fraction ($\varphi = 0.034$). The proportionality constant was taken from the NLOM model study with $\beta = 0.035$.

4.2.9 Evaluating the PSB model against experimental data

To further evaluate the applicability of the PBS model the steady-state concentration ratio predictions were compared against chemicals with experimental TK data across 3 or more external concentrations. There is a lack of available *D. magna* TK data with multiple external concentrations, however the best experimental data available was for six PFAS chemicals from [68]. In this study perfluorooctanesulfonic acid (PFOS) was chosen as the example chemical with *D. magna* exposed to 1, 5, and 10 $\mu g\ L^{-1}$ for 25 days [68].

Additionally, to examine the wider applicability to other aquatic invertebrates, the *G. pulex* TK data exposed to 0.05, 0.5, 5, 50, 500, 1500, and 5000 $\mu g\ L^{-1}$ of thiacloprid was collated from [215]. The PFOS data was digitised using the same method as in Chapter 2. While the thiacloprid concentration ratio data was available in the supplementary information [215]. The $\log_{10} D_{ow}$ of thiacloprid (1.22) was obtained using the same *in silico* methods as the other chemicals in this study.

4.3 Results

4.3.1 Theoretical upper bound for multiple external concentration scenarios

To test the theoretical upper bound of the PSB model the inferred steady-state concentration ratios were plotted against the predicted D_{ow} with the theoretical upper (red dashed line) and lower (green dashed line) bounds plotted for each of the external concentration scenarios ($\geq$ 0.08, 1, 10, and 100 $\mu g\ kg^{-1}$) (Figure 4.1). Generally, most of the 47 time-course inferred steady-state concentration ratios fit the theoretical upper bound across the four external concentration scenarios. While there are some datapoints that are close to and break the theoretical upper bound. It can be argued that this is attributed to experimental variation or the low protein affinity to concentration ratio, which is explored further in subsequent sections of the results. However, the greatest outlier across the $\geq$ 0.08, 1, and 10 $\mu g\ kg^{-1}$ scenarios is triphenyl phosphate (TPHP) with an external concentration of 100 $\mu g\ kg^{-1}$ highlighted by the red square in each relevant scenario. There are multiple steady-state concentration ratios below the lower bound, which are theoretically still plausible results but suggest that a physiological process or component is unaccounted for in the PSB model.

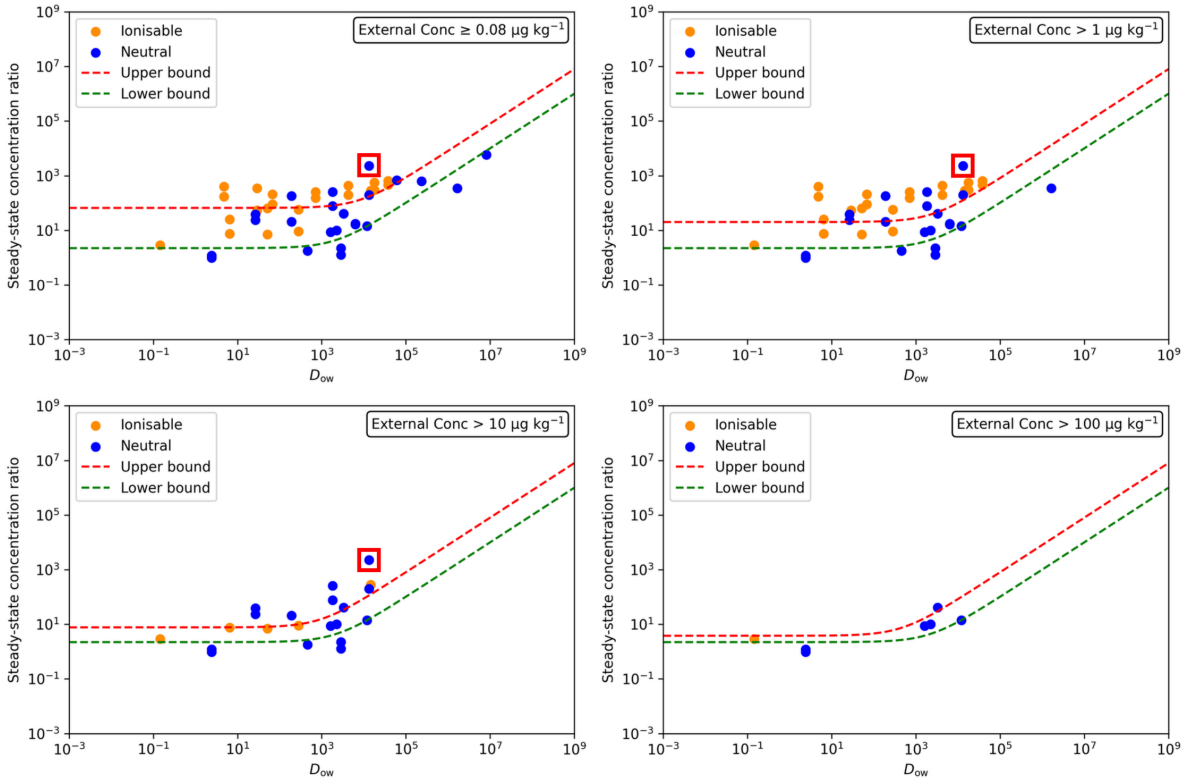

Figure 4.1: For 47 *Daphnia magna* toxicokinetic time-courses the inferred steady-state concentration ratio was plotted against the D_{ow} of the chemical with the theoretical upper (red dashed lines) and lower bounds (green dashed lines) of the protein surface-binding (PSB) model plotted for each external concentration scenario ($\geq$ 0.08, 1, 10, and 100 $\mu g\ kg^{-1}$). Steady-state concentration ratio data was only included if the external concentration met the threshold of the scenario. Ionisable (orange) and neutral (blue) chemicals were highlighted. Triphenyl phosphate (TPHP) was a clear outlier highlighted by the red square.

4.3.2 Benchmarking the PSB model against the NLOM model

The PSB model was benchmarked against the NLOM model by plotting the predicted steady-state concentration ratio against the inferred steady-state concentration ratio for the 47 time-courses (Figure 4.2). Both the PSB and NLOM model were run with D_{ow} as the partition coefficient. The 10-fold error bounds are highlighted with red dashed lines. The PSB model predicted 70.21% of time-courses (33/47) within 10-fold error while the NLOM model predicted 59.57% time-courses (28/47) within 10-fold error. Overall, it is clear that the NLOM model tends to underpredict the steady-state concentration ratio in comparison to the PSB model. The NLOM model underpredicts the 10-fold error margin

for 36.17% of time-courses (17/47) while the PSB model underpredicts for 25.53% of time-courses (12/47). Several of the underpredicted chemicals by the NLOM model are either pharmaceuticals or PFAS chemicals. Both the PSB and NLOM model overpredicts the 10-fold error margin for the neutral classified chemicals dichlorodiphenyltrichloroethane (DDT) and tris(2-butoxyethyl) phosphate (TBOEP) at 100 $\mu g\ kg^{-1}$.

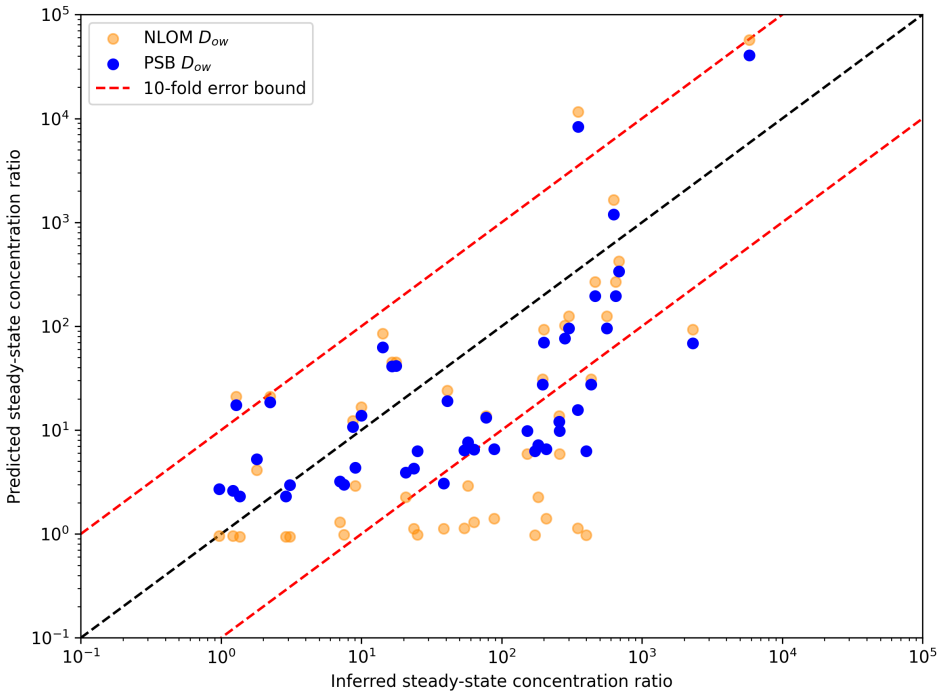

Figure 4.2: Predicted steady-state concentration ratio for the non-lipid organic matter (NLOM) model (light orange dots) and protein surface-binding (PSB) model (blue dots) plotted against the inferred steady-state concentration ratio for the 47 *Daphnia magna* toxicokinetic time-courses. The black dashed line shows the predicted and inferred steady-state concentration ratio are the same. A 10-fold error margin is highlighted with a red dashed line.

4.3.3 PSB model predictions for PFOS in *Daphnia magna*

To test the PSB model performance the steady-state concentration ratio predictions (solid red line) were compared with steady-state concentration ratios for PFOS experimental data for 1, 5, and 10 $\mu g\ kg^{-1}$ external concentration exposures from [68] (Figure 4.3). The NLOM model prediction was also highlighted for comparison (dashed black line). Based on the *in silico* $\log_{10} D_{ow}$ prediction of 0.68 the PSB model prediction clearly

underpredicts across all external concentration exposures. However, it is able to capture the decrease in steady-state concentration ratio as external concentration increases more accurately than the NLOM model. This underprediction suggests that some process or specific PFOS physiochemical property is unaccounted for in the model.

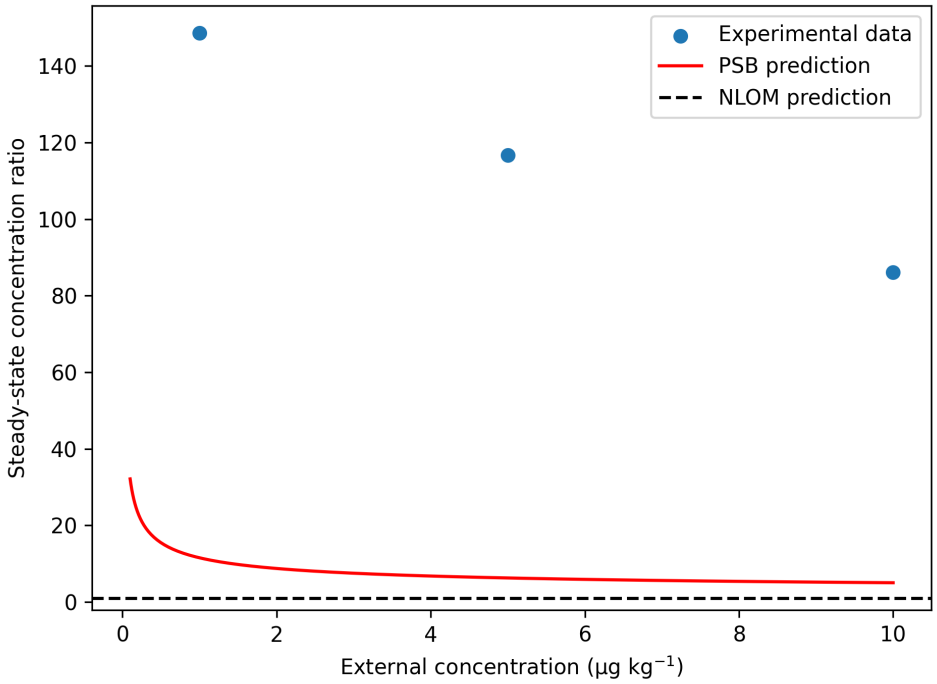

Figure 4.3: Experimental steady-state concentration ratios from [68] for perfluorooctanesulfonic acid (PFOS) in *Daphnia magna* across three different external concentrations (1, 5, and 10 $\mu g\ kg^{-1}$). The protein surface-binding (PSB) model (red line) and the non-lipid organic matter (NLOM) model (black dashed line) are highlighted to evaluate the performance of both models against the experimental data.

4.3.4 PSB model predictions for thiacloprid in *Gammarus pulex*

The PSB model prediction (red solid line) was plotted against the external concentrations alongside the experimental steady-state concentration ratios from [215] for thiacloprid on *G. pulex* at 2, 4 and 10 day exposures (Figure 4.4). The NLOM model prediction (black dashed line) was shown for comparison and to highlight the importance of accounting for external concentration changes for predicting the steady-state concentration ratio. It was important to test the PSB model against another aquatic invertebrate that had been exposed to external concentrations across several orders of magnitude to examine the

model's domain of applicability. Figure 4.4 illustrates the ability of the PSB model to account for a range of external concentration scenarios and capture the saturation effect seen in the experimental data. Conversely, the NLOM model steady-state concentration ratio predictions are the same across the range of external concentrations and does not capture the saturation effect in the experimental data.

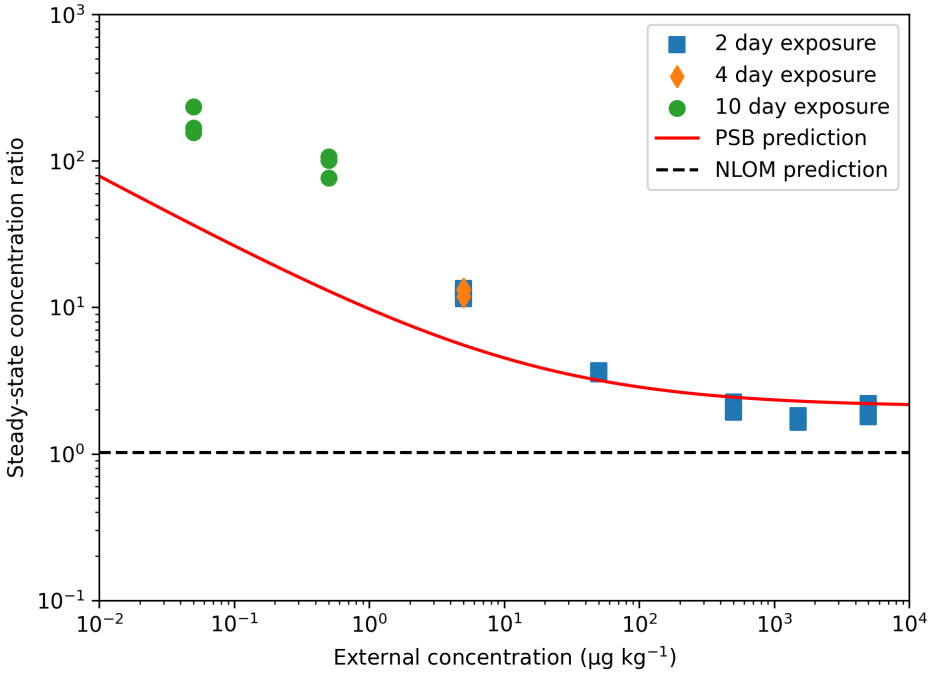

Figure 4.4: Experimental steady-state concentration ratios from [215] for thiacloprid in *Gammarus pulex* plotted against external concentrations from 0.05 – 5000 $\mu g\ kg^{-1}$ for 2 (blue squares), 4 (orange diamonds), and 10 (green circle) day exposures. The protein surface-binding (PSB) model (red line) and the non-lipid organic matter (NLOM) model (black dashed line) are highlighted to compare the predictive capabilities of the two models against the experimental data.

4.4 Discussion

This work has proposed a new theoretical protein-surface binding model that encompasses lipid fraction partitioning derived from decomposition analysis. Generally, TK models have been developed empirically, however, this requires large amounts of data and restricts the applicability of these model to chemicals within the training dataset.

This surface-binding approach taken in this study diverges from traditional empirical modelling and Fickian-like diffusion mass transfer by proposing a predictive model that accounts for protein-surface binding as a function of the external concentration and the protein fraction. However, the PSB model takes a more holistic approach to predicting the steady-state concentration ratio of the protein partitioning by assuming that the external concentration has the same density as the internal water concentration, which encloses the protein so that the chemical can bind to the surface of the total volume of the protein fraction. The protein surface-binding is a stochastic process where the probability of binding over the probability of unbinding is equal to the protein affinity. The theoretical upper bound scenarios highlight that PSB model predictions remain valid across external concentrations, which relates directly to the application of the PSB model to ERA with a risk assessor able to test different exposure scenarios. Furthermore, as the model does not require experimental data to widen the chemical space domain it can be easily applied to novel chemicals, which is essential to reducing animal testing and providing key toxicity data for ERA.

There is potential for chemical steady-state concentration ratio predictions to break the theoretical upper bound because either the chemical had a very high protein affinity or because the concentration relative to the affinity was low. As stated in Equation 4.11 the protein affinity is the ratio between the probability of a molecule binding to a probability of a molecule unbinding. Using the assumption that $P < (1/\rho - 1)$ will give you an upper bound for the protein affinity. In Appendix C.3.1 the protein affinity denoted by P can be interpreted as a relevant pharmacokinetic parameter or physiochemical property given the correct data, which could be evaluated against the same assumption to see if the physiochemical property breaks the model assumption and creates a tighter upper bound. The fup describes the extent of binding to plasma proteins and is a function of the protein binding affinity and the protein concentration [275]. This parameter can be estimated using *in silico* methods with a study showing that 87% of 441 chemicals predicted fup values were within 1-3 fold of empirical data [276]. Furthermore, fup data is not the only

physiochemical property that can be related to P. Developed QSARs have highlighted the correlation between protein affinity and lipophilicity [247]. A hydrophobic surface will bind 1000-fold faster than a hydrophilic surface, due to water creating a barrier to surface assembly [128]. Replacement of the protein affinity with a lipophilicity parameter, such as $\log_{10} K_{ow}$ would allow the applicability domain of the PSB model to be defined further. Future work could investigate the relationship between the protein affinity and key parameters, such as fup and $\log_{10} K_{ow}$ of the chemical to identify whether these parameters fit the model assumptions. Other steady-state concentration ratios above the theoretical upper limit could be attributed to experimental variation although one clear outlier was TPHP. However, this could be explained by the large standard deviation of the steady-state concentration ratio (890.32 ± 971.5) seen in the parent study [161]. The large uncertainties around the TPHP steady-state concentration ratio prediction were also highlighted through Bayesian analysis of the chemicals time-course data in Chapter 2.

The PSB model outperformed the state-of-the-art NLOM model in predicting the steady-state concentration ratio in *D. magna* with a higher percentage of predictions falling within 10-fold of the inferred steady-state concentration ratio values. The NLOM model tended to underpredict the steady-state concentration ratio, which can arise from the model not accounting for protein binding. Many of the underpredicted chemicals by NLOM were known to undergo protein binding. For example, fluoxetine and diclofenac as well as other pharmaceuticals are known to compete for binding sites of plasma proteins in humans [237, 44]. Additionally, there was no relationship between fluoxetine and lipid content in other aquatic invertebrates [176]. Pharmaceuticals such as diclofenac, propranolol, roxithromycin, and fluoxetine (at 5 $\mu g\ kg^{-1}$) were all underpredicted by the NLOM model, whereas, all four steady-state concentration ratios predictions by the PSB model were within 10-fold of the margin error. Fluoxetine exposed at 0.5 $\mu g\ kg^{-1}$ was still outside the 10-fold error margin for the PSB model, which might be a consequence of a low concentration relative to protein affinity. Fluoxetine has been shown to have

a high affinity to plasma proteins with 94.5% bound in pooled plasma [19]. However, more research would need to be undertaken to evaluate the relationship between protein affinity and steady-state concentration ratios. Therefore, accounting for protein partitioning through protein-surface binding processes in the PSB model can improve results for those chemicals, especially pharmaceuticals, that bind to proteins and further suggests that protein partitioning is an important process in accumulation of chemicals in aquatic invertebrates.

In the comparison between the NLOM and PSB models DDT and TBEOP steady-state concentration predictions were above the 10-fold error margin for both models. Both chemicals were neutral at the study pH and neither model predicted the chemicals well. DDT had a predicted octanol-water partition coefficient of 6.22, which makes it lipophilic and likely to accumulate in lipid components of organism. Previous work has shown that the total concentration of DDT partitioned into phospholipids is greater than the concentration bound to the protein of rat hepatic microsomes [23]. DDT is readily biotransformed into dichlorodiphenyldichloroethylene (DDE) and dichlorodiphenyldichloroethane (DDD), which can be present in greater concentrations than DDT in certain scenarios [162]. Similarly, it has been highlighted that TBOEP can be biotransformed into bis (2-butoxyethyl) hydroxyethyl phosphate (BBOEHEP) and bis (2-butoxyethyl) 3-hydroxyl-2-butoxyethyl phosphate (3-OH-TBOEP) [161]. This suggests that biotransformation of chemicals needs to be accounted for in the model and is the likely source of the overpredictions for these particular chemicals.

PFAS chemicals were a particular chemical group of interest in this study, due to their regulatory concern, widespread nature in the environment, and great uncertainty related to their exposure and effects in the environment [12, 174]. It is well established that PFAS chemicals bind to proteins [271]. Therefore, it was expected that the PSB model would be able to predict all the PFAS chemicals within a 10-fold error margin by accounting for protein binding. However, 7 out of the 12 PFAS time-courses were underpredicted

and outside the 10-fold error margin for the NLOM and PSB model. This suggests that the internal concentration is partitioning into a component unaccounted for in the model. It has been theorised that the biochemical component bio-ash could be relevant for the partitioning of ionisable chemicals, however, there are a lack of studies that focused specifically on this component [57]. This may be even more relevant for *D. magna* as the fraction of bio-ash is similar to the lipid fraction calculated in Bogut et al. (2010) [40].

While the theoretical PSB model holds for predictions of the steady-state concentration ratios in the *D. magna* dataset, there is limited chemical data for a range of external concentrations. Application of the PSB model to the PFOS steady-state concentration data from three different external concentrations in *D. magna* and thiacloprid across external concentrations spanning 5 orders of magnitude in *G. pulex* allowed the domain of applicability to be tested. A clear limitation in the *D. magna* example was that the external concentrations only covered one order of magnitude but the data available was limited. It highlighted a clear underprediction of the steady-steady concentration ratio across the external concentrations, however, the PSB model was able to marginally capture the change in steady-state concentration ratio over external concentration over the NLOM model. One explanation is the experimental variability of PFAS chemicals. Inter-laboratory studies have shown coefficients of variation of 125% for PFOS concentrations in fish tissues [257]. This suggests that internal concentrations can have large variations between experiments. Additionally, the two distinct PFOS internal concentrations in *D. magna* at 24 hours exposure time from the *AquaTK* dataset was 1995.1 $\mu g\ kg^{-1}$ [271] and 590.9 $\mu g\ kg^{-1}$ [68]. This constitutes a 3-fold difference in internal concentration at the same external concentration, which could account for some of the underprediction of the PSB model. Conversely, the PSB model predicts some of the experimental properties of the thiacloprid steady-state concentration ratios in *G. pulex* dependent on the external concentration. It is established that thiacloprid binds to proteins, which makes it a perfect example of the PSB models predictive capabilities across different external concentration

scenarios. Furthermore, this example suggests that the PSB model may not be limited to *D. magna* and could be applied to a wider range of organisms. Further work would involve using the PSB model for steady-state concentration ratios across a wider range of chemicals and organism to truly test the domain of applicability and its application to environmental risk assessment.

A key assumption of the PSB model is that the protein fraction is a smooth homogenous flat surface with a uniform distribution of chemical. Alternatively, protein surfaces are chemically heterogeneous and rough [4]. A non-smooth surface would increase the surface area enclosed in the protein volume resulting in a decrease in internal concentration. An empirical parameter than describes the heterogeneity of the surface and absorption intensity is the Freundlich constant $(1/n)$. Favourable absorption is indicated when $1/n < 1$ [252]. Comparatively, the PSB model exponent is 2/3, which is equivalent to the Freundlich constant and is assumed to be the ideal case across the dataset based on the assumptions of a flat surface. However, the exponent could be much larger if the surface is not flat and the surface contained different number and types of proteins to bind to on the surface. Conversely, the exponent could be much smaller if there there was only a percentage of the the surface of the protein to bind to or only a percentage of the chemical can bind to the protein. The percentage of chemical molecules available to bind could be impacted by the chemical concentration, which results in a change in the chemicals protein affinity. Two anaesthetics propofol and halothane [33], alongside the anticoagulant warfarin [207] were shown to bind to proteins with different affinities as the concentration of the chemical increased [138]. Further work could incorporate different empirical or *in silico* derived values of n to allow more chemical specific predictions of the surface-binding process.

4.5 Conclusions

To conclude, this work has developed a new TK model that incorporates lipid partitioning and protein surface-binding processes through decomposition analysis for TK predictions in *D. magna*. The distinction between the PSB model and other TK models is that it is not necessary to calibrate on experimental data and integrates protein surface-binding as a function of the protein fraction and external concentration. A theoretical upper bound was validated using available steady-state concentration ratios to increase confidence in the use of the PSB model. This model will have a positive impact on NAMs-based ERA as it can provide predictions of the steady-state concentration ratio across external concentration scenarios, which removes the restrictions of risk assessors to available experimental data only and reduces the need for further animal testing. Furthermore, this will enable novel chemicals outside the training dataset of the model to be evaluated. Additionally, the PSB model predicted steady-state concentration ratios more consistently within a 10-fold error margin than the state-of-the-art NLOM model, which is relevant to ERA as a 10-fold safety factor is usually applied to TK data. Further work could focus on expanding the domain of applicability by testing the PSB model predictions against a wider chemical space and other available aquatic invertebrate experimental data with the example of *G. pulex* showing good correlation with the PSB model predictions.

CHAPTER 5 : A RANDOM FOREST REGRESSION MODEL FOR PREDICTING THE RELATIVE IONISATION EFFICIENCY OF PARENT CHEMICALS AND THEIR BIOTRANSFORMATION PRODUCTS

5.1 Introduction

To understand the potential risk on the environment it is important to take into account both the parent contaminant and its transformation products [136]. BTPs can be more prevalent than their respective parent chemical and can contribute to the risk of the parent if they are formed with high abundance, have a high toxicity, or are more mobile and persistent [90]. Regulatory documents highlight the need to include BTPs in risk assessment but differ significantly in guidance and applicable tools [90]. Traditional analytical methods to quantify the concentrations of chemicals employ MS with an internal standard (for each pre-selected chemical), which limits these studies to known substances [106]. However, many BTPs have not been identified yet, which leaves a significant number of chemicals and their impacts on the environment unknown [90, 106]. Moreover, even if BTPs are known, the ability to quantify them is restricted by the lack of authentic standards [122]. This problem spans not only ERA but human risk assessment as well, as only 18.5% of the ca. 114,100 largely synthetic chemicals have been detected and identi-

fied in the Human Metabolome Database [157]. This has considerable consequences for TK model development where relevant BTPs cannot be characterised in compartments or in terms of biotransformation rates to understand the fate of the chemical in an organism [147].

Due to the shortage and typically high cost for synthesising standards, alternative methodologies are being developed that involve "semi-quantification" without standards. While there are several emerging methods available, one of the most popular is the use of IE, which has been shown relative to other methods to produce the most accurate semi-quantification of chemicals compared to experimental values [143, 2]. IE values are dependent on the configuration of the MS instrumentation and therefore many studies calculate the IE relative to an "anchor compound" to make the values generalisable across labs [167]. Multiple IE predictive models have been developed from multiple linear regressions to a variety of machine learning algorithms [156, 173, 204]. However, the most generalisable and extensive model is derived from the random forest regression developed by [156]. This model included experimental IE data from 353 unique chemicals in positive ion mode and 109 unique chemicals in negative ion mode across a range of liquid chromatography eluent compositions [156]. The model is able to predict IE of chemicals using PaDEL descriptors, which are predicted molecular descriptors for each chemical. Reliable predictive IE models are the first step to predicting the concentration of chemicals without standards including BTPs. See section 1.6.2 for an introduction to this approach, and Figure 1.4 for an overview of the modelling strategy.

Multiple studies have used the IE prediction model from Liigand et al. (2020) to predict the concentrations of chemicals (i.e. semi-quantify the chemicals) [156], such as hydroxylated polychlorinated biphenyls that resulted in a mean quantification error of 4.4x the actual value [130]. More relevant is the use of the model to predict concentrations of 60 BTPs with a mean error of 2x and a maximum error of 11x of the actual value when compared to standards [143]. However, this is the full extent to which models have been

applied for predicting the IEs of BTPs and hence their concentrations, with the latest research showing no specific model developed for parents and BTPs, which is a significant knowledge gap that needs investigating further especially for ERA TK modelling purposes. Other methods have tried calculating RF of BTPs relative to parent chemicals for 71 drugs and BTPs, focusing on a limited number of transformation pathways, resulting in large errors in prediction accuracy with fold responses varying from 70-fold lower to 8.6-fold higher than the parent drug [111].

Therefore, the overall aim of this research was to design a study comprising the generation of a new MS dataset and subsequent development of a random forest regression model to predict the RIE values of parents and BTPs from the new experimental IE data. This aim can be divided into three specific objectives. The first objective was to perform a chemical criteria assessment where commercially available parent and BTPs were identified (for purchase) and assessed against a criterion for selection in IE experiments. The second objective was to collect experimental IE data using MS (conducted by the Phenome Centre Birmingham, University of Birmingham). The third objective was to develop a random forest regression model using the experimental IE values and evaluate its performance using different validation methods.

5.2 Material & methods

5.2.1 Chemical selection

As standards for parent chemicals and BTPs are limited, the first step was to identify what standards were available to buy commercially. A comprehensive analysis of parent and BTP pairs in the Eawag parent-BTP pair [232] and MetxBioDB [83] databases were checked for availability from online vendors, such as Sigma-Aldrich (https://www.sigmaaldrich.com) and LGC Standards (https://www.lgcstandards.com), which resulted in 213 parent chemicals and BTPs. Specifically, each parent and BTP was given a specific ID, where each chemical is given a number and a letter. Each related parent and

BTP is given the same number, while an A represents the parent chemical of the group, any other letter represents a biotransformation of A. For example, atrazine was labelled 16A, while its BTPs atrazine-desethyl and atrazine-desisopropyl were labelled 16B and 16C. The number of letters is dependent on the number of available BTPs, which varies between chemicals.

Based on resource constraints, the first list of chemicals needed to be reduced. A criteria list made up of seven key questions was created with answers given as "Yes" or "No", which allowed a binary scoring system (1,0):

1. Is the molecular weight of the chemical > 100 Da (chemicals with larger molecular weights are easier to detect in the MS)
2. Is the mass of the available chemical standard > 10 mg (important to have a high enough quantity to create calibration curves)
3. Are the number of BTPs linked to parent ≥ 2 (attempt to include as many related BTPs to a given parent as possible)
4. Is the price of the chemical < £200 (cost effective to include lower priced chemicals)
5. Is the source MetXBioDB (considered the most reliable database)
6. Is the BTP pathway known (enables as many transformation pathways to be included as possible)
7. Is the purity of the standard known and above 95% (more reliable and reproducible calibration curves)

The total scores (out of 7) for each chemical standard were inspected and low scoring chemicals were removed. This resulted in the chemical dataset being reduced from 213 to 125. Further inspection of the chemical list highlighted multiple further chemicals that needed to be removed as they were either known to not ionise well or were relatively high cost. Some chemicals that had lower criteria scores were able to be included, such as

imidazole, that had low mass (68 Da) but have been known to ionise and be detected with wider m/z ranges [175]. After manual review, the chemical list was reduced to 109 parent and BTPs. Finally, the chemicals were given CLP Toxicity ratings subdivided into decreasing toxicity groups from 1 to 4 [71]. Twenty-two chemicals were characterised as either group 1 or 2, which meant they were some of the most toxic and of particular concern. To avoid any safety concerns for the lab team in the Phenome Centre Birmingham, these 22 chemicals were removed from the dataset resulting in a final chemical dataset of 88 parent and BTPs plus an anchor chemical (tetraethylammonium chloride) that were purchased from vendors.

5.2.2 Experimental data collection and results of pilot studies

The standard preparation, preliminary and primary experiments, and initial processing of the raw flow injection-mass spectrometry (FI-MS) data was conducted by the Phenome Centre Birmingham.

5.2.2.1 Standard preparation

All standards were prepared in 100% methanol (LC-MS grade, VWR). Serial dilutions were conducted with 100% methanol. Standard solutions in methanol were loaded into 96-well plates (TwinTec, Eppendorf) which were placed in the autosampler (Transcend Vanquish Flex Duo LX-2 system Vanquish LC system, ThermoScientific).

5.2.2.2 Overview of flow injection-MS method for preliminary and primary experiments

Samples were analysed as triplicate injections, one compound at a time with the analysis order going from most dilute to the most concentrated. The FI-MS assay was conducted on a Transcend Vanquish Flex Duo LX-2 system Vanquish LC system (Thermo Scientific) using a metal coupler instead of an LC column. Data were acquired on an Orbitrap Exploris 120 mass spectrometer (Thermo Scientific) in positive ion mode only. Ion source parameters: Sheath gas = 40 arbitrary units, Aux gas = 8 arbitrary units, Sweep gas =

1 arbitrary unit, Spray Voltage = 3.2 kV, Capillary temp. = 300°C, Vaporiser temp. = 300°C. The raw FI-MS acquisition data was processed using the R package proFIA [76] to identify peaks for each chemical under consideration and its corresponding concentrations. The resulting peak matrix was then matched to the mass-to-charge ratios of the chemicals.

5.2.2.3 Preliminary experiments

Preliminary experiments were undertaken to identify the optimal concentrations to achieve a linear gradient (of detector response versus concentration), the most relevant m/z range to allow detection but minimise loss of accuracy, and to identify chemicals in the list that do not ionise at all. In preliminary experiment 1, six serial dilutions of a 100 μM standard were prepared with a 1 in 5 dilution factor resulting in 100, 20, 4, 0.8, 0.16, and 0.032 μM with a m/z range set to 150 and above with each sample analysed with 10 μL injections. In preliminary experiment 2, six serial dilutions of a 200 μM were prepared with a 1 in 30 dilution factor resulting in 0.00000823, 0.00025, 0.0074, 0.22, 6.67, and 200 μM with a wider m/z range set to 60 and above. These samples were analysed with 5 μL injections. Three chemicals were also removed from the study because of solubility issues in standard preparation. Across the two preliminary experiments there were chemicals that had no peaks detected and also had low R^2 values. Therefore, these chemicals were also removed from the study and this resulted in a final dataset of 67 parent and BTPs, including the anchor chemical. This was comprised of 29 parent chemicals (including the anchor) and 38 BTPs. Based on the preliminary experiment results each chemical was assigned a concentration group from A (best ionisers) to D (poor ionisers).

5.2.2.4 Primary experiment

The primary experiment used the final 67 chemicals and the assigned concentration groups to obtain linear gradients of MS detector response verse concentration of standards. Each chemicals concentration group had six serial dilutions of the highest concentration and

were prepared with a 1 in 5 dilution factor, which generated a maximum number of 18 intensities across six concentrations. A table illustrating the concentrations of each group can be seen in Table 5.1. The MS acquisition range was over 2 x m/z ranges (range 1: 60-205 m/z; range 2: 200-2000 m/z) defined as separate MS scan events within a single assay. Each sample was analysed with triplicate 10 μL injections in the MS.

Table 5.1: Concentration values for each of the assigned concentration groups A, B, C, and D.

Concentration group	Conc 1 (μM)	Conc 2 (μM)	Conc 3 (μM)	Conc 4 (μM)	Conc 5 (μM)	Conc 6 (μM)
A	0.00128	0.0064	0.032	0.16	0.8	4
B	0.0064	0.032	0.16	0.8	4	20
C	0.032	0.16	0.8	4	20	100
D	0.064	0.32	1.6	8	40	200

Calibration curves between the concentration and intensity were created with linear regressions fit for each specific chemical forcing the intercept through the origin. The calibration curves with R^2 values can be seen for all chemicals in Appendix D Figures D.1 - D.6. The linear regressions of the calibration curves and figure generation were conducted in R. As a result of the calibration curves and R^2 values it was evident for some chemicals that the detector was becoming saturated at higher concentrations and not exhibiting a linear relationship between the intensity and the concentration. Therefore, initially the highest concentration (concentration 6) and associated intensity was removed from the dataset for all chemicals and the linear regression lines refit. The new R^2 values for the reduced dataset were compared with the values from the original dataset with all concentrations included. If the R^2 value improved by greater than 2.5% when removing the highest concentration from the dataset then the new R^2 value was accepted and the highest concentration was removed for that chemical. This improved the R^2 values of 11 chemicals. Moreover, there were still some low R^2 values and it seemed that multiple chemicals were saturated beginning at the second highest concentration (concentration 5). Therefore, the same method with a 2.5% improvement threshold was also implemented but when removing the two highest concentrations (concentration 5 and concentration

6). This improved R^2 values for a further 5 chemicals and especially the chemicals with low R^2 of concern. Overall, all 67 chemicals had R^2 values > 0.97. Finally, RIE values were calculated by dividing the gradient of each chemical by the gradient of the anchor chemical,

$$RIE = \frac{\text{gradient of chemical}}{\text{gradient of anchor}} \tag{5.1}$$

The RIE values were log-transformed ($\log_{10}$RIE) to reduce the variance as the RIE spanned 5 orders of magnitude. An overview of the chemical ID, chemical name, concentration group, gradient of linear regression, R^2 value, RIE, and $\log_{10}$RIE for each chemical can be seen in Appendix D Table D.1.

5.2.3 Model development and evaluation

All pre-processing of the $\log_{10}$RIE and PaDEL descriptor data, model development, and figure generation were conducted in Python 3.11 with packages Matplotlib 3.8, NumPy 1.26, PaDELPy 0.1, pandas 2.2, scikit-learn 1.4, SciPy 1.13, Seaborn 0.13, and Shap 0.45. Prior to developing the $\log_{10}$RIE predictive model chemical SMILES were used to calculate 2D PaDEL descriptors (1444 descriptors) for the 67 chemicals from the PaDELPy package. Pre-processing steps were taken to reduce the potential of the model capturing noise or increasing complexity of the model as a result of too many descriptors. Firstly, any PaDEL descriptor that was unable to be predicted was removed. After that descriptors that had the same value for 60% or more of the chemicals were identified, which resulted in 541 descriptors being removed. Finally, descriptors that correlated highly with each other but less so with the $\log_{10}$RIE value were removed to address multicollinearity as it could impact the reliability of model estimates [7]. If a descriptor had an R^2 correlation greater than 0.6 with another descriptor then the descriptor with the lower correlation with the $\log_{10}$RIE value was removed. The final number of descriptors was reduced to 113.

The random forest regression algorithm has been shown to consistently predict the RIE

of chemicals [156, 143], therefore, it was chosen as the model for this study. The dataset was split into X (PaDEL descriptors) and y ($\log_{10}$RIE) variables. Initially the 67 parent and BTPs (including the anchor chemical as a parent chemical) were randomly split into training and test datasets with 70% of the chemicals used for training the model and 30% as a validation test set. However, this split resulted in overfitting where the model was learning too well on the training set. Therefore, a 60% (40 chemicals) and 40% (27 chemicals) training to test split was implemented. The test set is independent and only used for assessing the predictive power of the model, which is considered the gold standard [6]. A random seed of "2024" was assigned to allow reproducibility of results. The descriptors were scaled using the standard scaler function from scikit-learn, which standardises each PaDEL descriptors by making the mean and variance equal to 0 and 1, respectively. This makes sure that any descriptors with large means or variances are reduced and not dominating over other descriptors, which would skew the model decision-making.

A GridSearchCV procedure from scikit-learn was carried out to identify the appropriate parameters for the random forest regression model, including the number of decision trees, the criterion to decide the quality of each split in the decision tree, depth of each individual decision tree, and the maximum number of features (PaDEL descriptors). A 5-fold internal cross validation on the training set was undertaken to tune the best parameters, with the lowest scoring parameters scored by mean squared error chosen. The best parameters were 10 decision trees, squared error as the criterion, maximum depth of 4, and maximum number of descriptors of 4. The random forest regressor package from scikit-learn with the best parameters was fit to the training dataset.

The model performance was evaluated on both the training and test sets using the RMSE and Pearsons R^2 values and comparing the ratio of the standard deviation of $\log_{10}$RIE values and RMSE to show the model performance relative to a random prediction. Additionally, an internal validation on the training set was undertaken using a 10-fold cross

validation to investigate the generalisability of the model on unseen data. The training set containing the random 40 parent and BTPs was split into 10 subsets (4 chemicals each). For each of the 10 iterations the optimised random forest regression model was fit to nine of the subsets, acting as a training set and the other subset was used as the validation set and evaluated with the RMSE. The RMSE was chosen as the main performance metric as it is typically more practically important than R^2 [6]. It has been highlighted that models with low R^2 values can still be useful with low RMSE values [6]. The robustness of the model was tested further by performing a y-scramble, which is a test widely used in machine learning evaluation [178]. The y-scramble breaks any meaningful link between the X variable (PaDEL descriptors) and the y variable ($\log_{10}$RIE) by randomly assigning the y variables to different X variables [178]. This was done 1000 times and the performance metric for each iteration using RMSE and Pearsons R^2 was compared.

The most influential descriptors on predicting $\log_{10}$RIE values from the model were obtained using the SHAP: Shapley additive explanation, which is a method based on game theory to enable the contribution of important descriptors to be interpreted [159]. Recently, SHAP values have been used in metabolomics studies to identify the key features for metabolites [34]. The relative importance of each PaDEL descriptor is assessed by removing a descriptor and assessing the impact on the unique SHAP values of the other descriptors. Each unique SHAP value determines the contribution of the descriptor on the target variable [27]. The SHAP values were applied to both the training and test dataset to evaluate whether the same influential descriptors appear, in this way it is possible to make sure the model makes the same decisions when seeing chemicals from the training and test set.

5.2.4 Chemical similarity

The chemical space that the final dataset of 67 parent and BTPs (including the anchor chemical) captured was evaluated against the original dataset collated in 5.2.1 containing 213 chemicals and then the Eawag parent-BTP pair and MetXBioDB databases, which

included 1530 chemical entries across the two databases by comparing the molecular similarity. Molecular similarity between the different datasets was compared by using the molecular fingerprints of each chemical and visualising using principal component analysis (PCA). Morgan fingerprints, are one of the most popular molecular fingerprints and represent chemical structural characteristics as a vector. These structural characteristics are encoded as binary bits (1s and 0s). For example, if a chemical contains a hydroxyl group, it is represented by a 1, but if it does not contain that functional group, it is represented by a 0. Morgan fingerprints typically have a length of 2048 bits and are generated using the RDKit library package 2023.03.2 [182, 51]. This allows the structure of chemicals to be compressed and simplified to be used for comparing structural similarity. A PCA was applied to the collated Morgan fingerprints to explain the similarity or variation in molecular structures.

5.3 Results

5.3.1 Model performance

To understand the prediction performance of the random forest regression model the predicted $\log_{10}$RIE were compared against the experimental $\log_{10}$RIE for the 67 parent and BTPs divided into a 60% training set (blue) and a 40% test set (orange) (Figure 5.1). Overall the experimental $\log_{10}$RIE values span approximately 5 orders of magnitude, which illustrates the variability in $\log_{10}$RIE across the physiochemical and structural characteristics of parent and BTPs. From a qualitative perspective it can be argued that the model tends to overpredict the lowest $\log_{10}$RIE values (poorest ionisers) while underpredicting the highest $\log_{10}$RIE values (best ionisers). The training set showed strong prediction performance with RMSE of $\log_{10}$RIE and Pearsons R^2 values of 0.37 and 0.92, respectively, which is expected as the model is optimised on this set of chemicals. Alternatively, the test set showed worse prediction performance with RMSE of $\log_{10}$RIE and Pearsons R^2 values of 0.77 and 0.64, respectively. A worse prediction performance on the

test set is expected as the optimisation of the model did not include these chemicals.

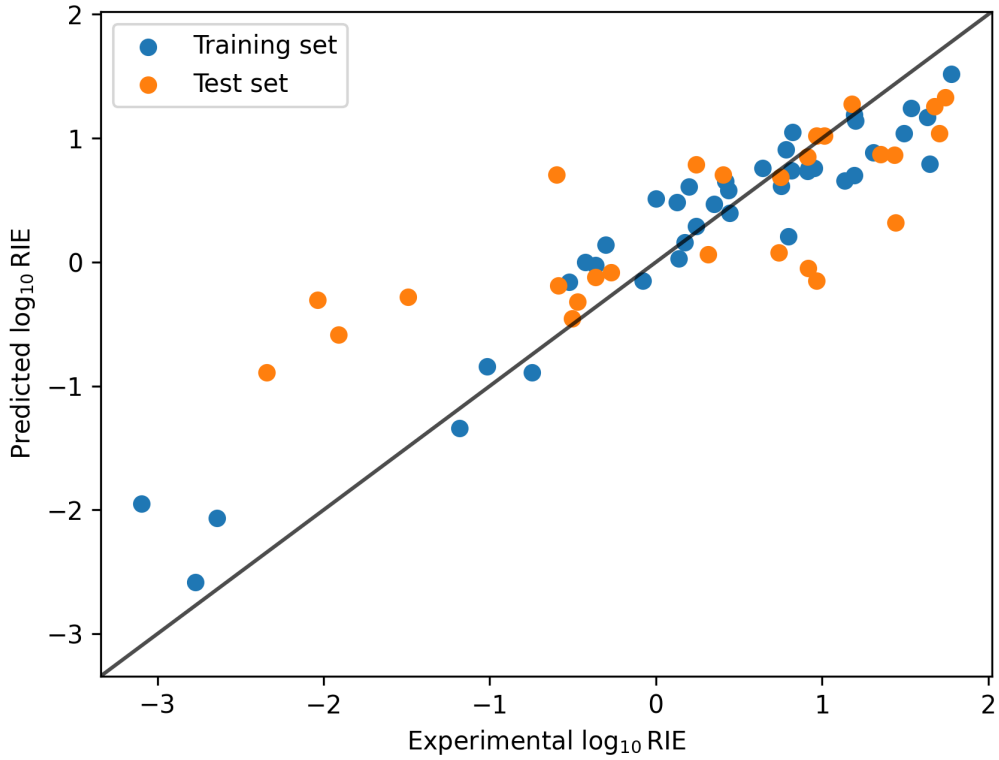

Figure 5.1: Predicted log-transformed relative ionisation efficiency ($\log_{10}$RIE) values from the random forest regression model against the experimental $\log_{10}$RIE values for the training set (blue) and the test set (orange) for 67 parent and biotransformation products split into a 60% and 40% training and test split. The root mean square error (RMSE) of $\log_{10}$RIE for the training and test sets were 0.37 and 0.77, respectively. The R^2 values of the training and test set were 0.92 and 0.64, respectively.

To estimate the generalisability of the model across different subsets of parents and BTPs a 10-fold cross-validation was performed on the training set of chemicals. For the remaining analysis the RMSE of $\log_{10}$RIE is used as the metric performance and is hereafter referred to as RMSE for brevity. The distribution of the RMSE of $\log_{10}$RIE of each fold of the cross-validation can be seen in Figure 5.2 with the mean (red cross), median (orange line), RMSE test set (green dashed line), and RMSE of the training set (blue dashed line). The minimum RMSE is 0.34 and the maximum is 1.54 with a range of 1.2. It could be argued that the random forest regression model struggles to predict certain subsets of the chemicals. The mean RMSE across each cross-validation fold is 0.768, which is

relatively central in relation to the minimum and maximum RMSE values and therefore representative on average of the model performance. The mean RMSE value is greater than the training set RMSE (0.37) and rounded to 2 decimal places is the same as the test set RMSE (0.77). Additionally, there is variance of RMSE across cross-validation folds, which will be driven by the fact RMSE for each fold is calculated from 4 chemicals, as opposed to 27 chemicals in the test set. Furthermore, as the average RMSE and the RMSE of the test set are similar it highlights the consistency of the model in predicting unseen data. The RMSE of $\log_{10}$RIE can be interpreted as a multiplicative factor, see Liigand et al. (2020) for a similar interpretation [156]. Therefore, the cross-validation folds mean RMSE of $\log_{10}$RIE value of 0.77 can be interpreted as a multiplicative factor of 6. This means that any future predictions on average will be 6 times away from the true value. Additionally, as previously stated the cross-validation RMSE of $\log_{10}$RIE ranges from 0.34 to 1.54. Therefore, the cross-validation RMSE of $\log_{10}$RIE ranges from 2 times and 35 times away from the true value. This highlights that the performance can differ significantly depending on each of the 4 chemicals contained within a cross-validation fold.

To assess the robustness of the random forest regression model a chance correlation analysis in the form of a y-scramble where $\log_{10}$RIE values were randomly shuffled across the unchanged X variables (descriptors) 1000 times was performed on the test set of parent and BTPs (Figure 5.3). This was performed to establish whether the model performance is dependent on the specific arrangement of X and y variables input into the original model. The RMSE of $\log_{10}$RIE i.e. multiplicative factor (top plot) and Pearsons R^2 (bottom plot) performance metrics were calculated for every iteration and compared with the original models RMSE test set (dotted red line - top plot) and Pearsons R^2 test set (dotted red line - bottom plot). The y-scramble results highlight that through 1000 random permutations the calculated RMSE value is never less than the RMSE of the test set (0.77). Similarly, the Pearson R^2 values through the permutations are never greater than the R^2 values of the test set (0.64). This strongly suggests that the model is able to

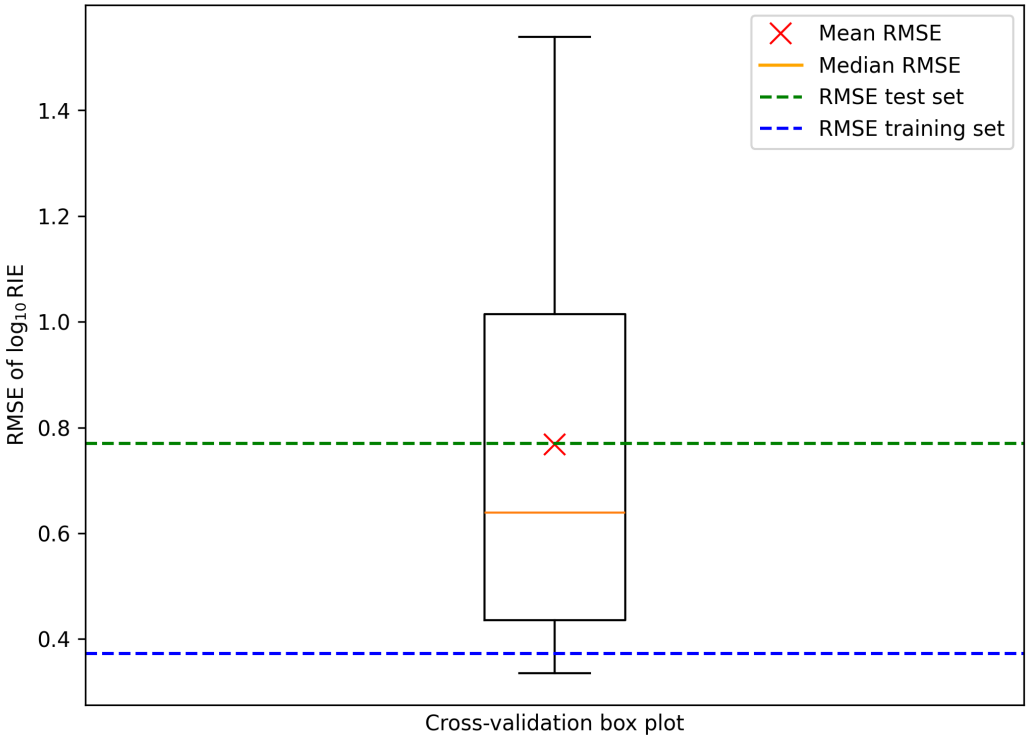

Figure 5.2: Box plot showing the distributions of the root mean squared error (RMSE) of $\log_{10}$RIE values i.e. multiplicative factor from a 10-fold cross-validation of the training set of parent and biotransformation products with representations of the mean (red cross), median (orange line), RMSE test set (green dashed line), and RMSE of the training set (blue dashed line).

capture a relationship between the $\log_{10}$RIE and the chosen descriptors, which is not as a consequence of randomness. Following the method highlighted in the previous section the RMSE of $\log_{10}$RIE test set value of 0.77 can be interpreted as a multiplicative factor of 6 times. This suggests that any future prediction will on average be 6 times away from the true value. Furthermore, the average scrambled model had a RMSE of $\log_{10}$RIE of approximately 1.2, which has a multiplicative factor of 16. Overall, the model performance on the test set was approximately 2.7 times better than the average scrambled model.

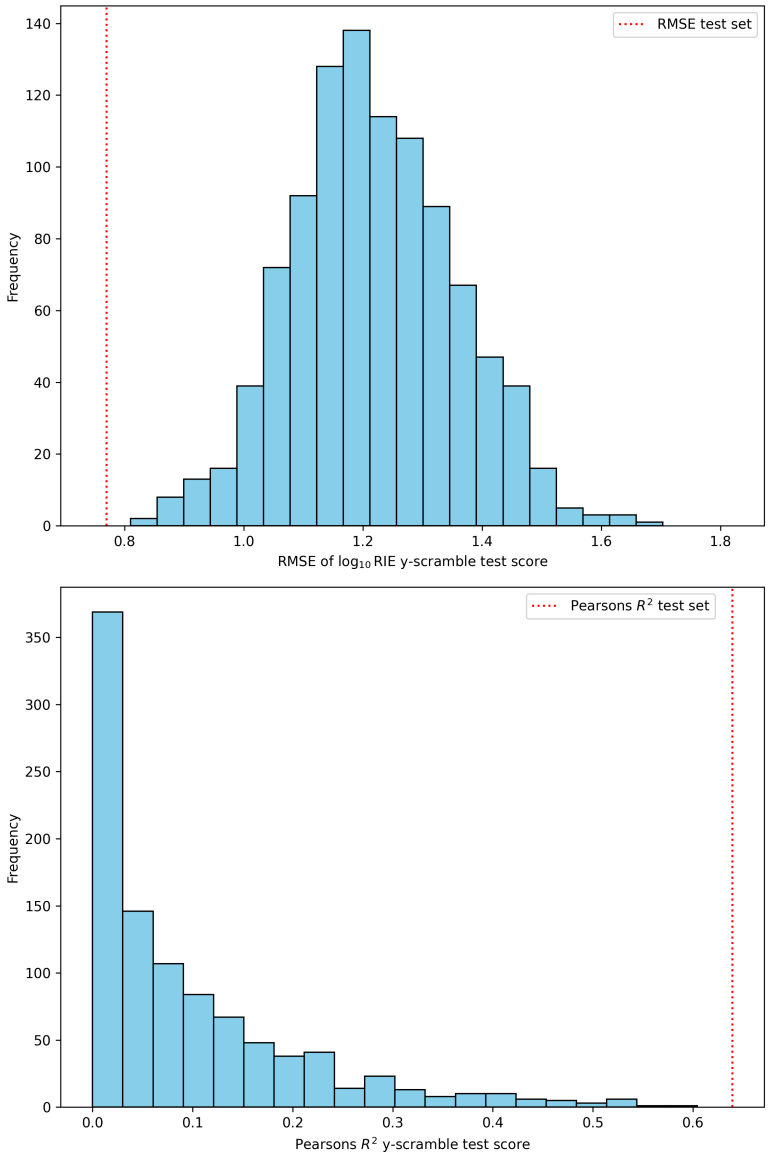

Figure 5.3: Histograms showing the frequency of the root mean square error (RMSE) of $\log_{10}$RIE i.e. multiplicative factor (top plot) and Pearsons R^2 values (bottom plot) as a result of a y-scramble where the y-variable (log-transformed relative ionisation efficiency - $\log_{10}$RIE) is shuffled 1000 times across the unchanged X variables (PaDEL descriptors). The RMSE test set and Pearsons R^2 values are highlighted with dotted red lines on the top and bottom plot, respectively.

5.3.2 Most influential model parameters

To establish the most influential PaDEL descriptors the top 10 highest SHAP values summarising the impact of descriptors on model predictions of the $\log_{10}$RIE were calculated for the training set (top plot) and test set (bottom plot) in Figure 5.4. The SHAP values

were repeated for the training and test sets to see if the most influential parameters were the same in the test and training set. The greater the colour and value separation the more influential the descriptor is to the model output. If a datapoint is blue it represents a negative value for that descriptor whereas a red colour represents a positive value. A negative SHAP value represents a negative effect on the $\log_{10}$RIE value and a positive SHAP value represents a positive effect. There are multiple influential descriptors that impact the models prediction of the training (top plot) and test set (bottom plot) including "ATSC8m", ranked 1st (training) and 3rd (test), where really high values of the descriptor have negative effects on the predictions. Alternatively, "AATSC3s", ranked 3^{rd} (training) and 2^{nd} (test), "ATSC8p", ranked 5^{th} (training) and 5^{th} (test), "C1SP2", ranked 7^{th} (training) and 7^{th} (test), "MaxHother", ranked 10^{th} (training) and 10^{th} (test), have clear distinction of colours and values where negative descriptor values have a negative impact on the predictions and vice versa for positive values of the descriptor.

5.3.3 Visualisation of chemical similarity

A PCA of chemical molecular fingerprints from the 67 final selected parent and BTPs (green) compared to the 213 commercially available chemicals (black circles) to visualise the chemical similarity (Figure 5.5 PCA (A) - top plot). To further understand the chemical space captured in this study a PCA was conducted on the wider Eawag parent-BTP pair and MetXBioDB databases that contained 1530 chemicals (Figure 5.5 PCA (B) - bottom plot). In terms of the chemical similarity comparisons of the commercially available parents and BTPs the selected chemicals capture most of the chemical structures out of the 213 parent and BTPs. This highlights that the model includes structures that cover most of the commercially available parents and BTPs that were identified. However, when the chemical similarity of the selected chemicals is compared with the Eawag and MetXBioDB databases there are substantial numbers of chemicals that are not captured. While the main cluster of chemicals are similar in structure there are a considerable number of structures unique to the databases.

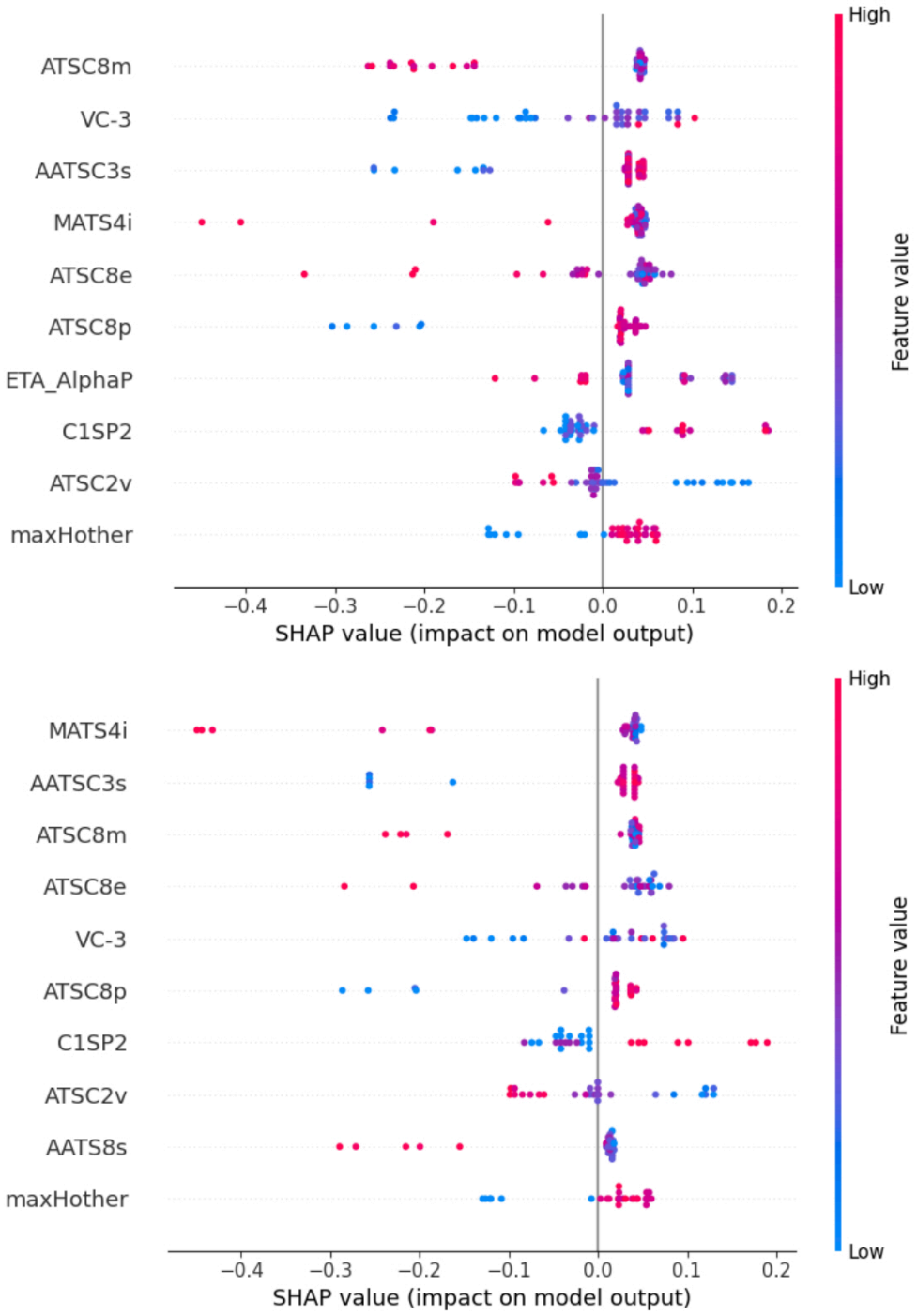

Figure 5.4: The PaDEL descriptors with the top 10 highest SHAP values summarising each descriptors contribution to the random forest regression models prediction of log-transformed relative ionisation efficiency ($\log_{10}$RIE) values for the training (top plot) and test set (bottom plot). A datapoint with a red colour highlights a high value for the descriptor and a blue colour represents a lower value for the descriptor. Negative SHAP values represent negative impacts on the $\log_{10}$RIE values, while positive SHAP values represent positive impacts on the $\log_{10}$RIE values.

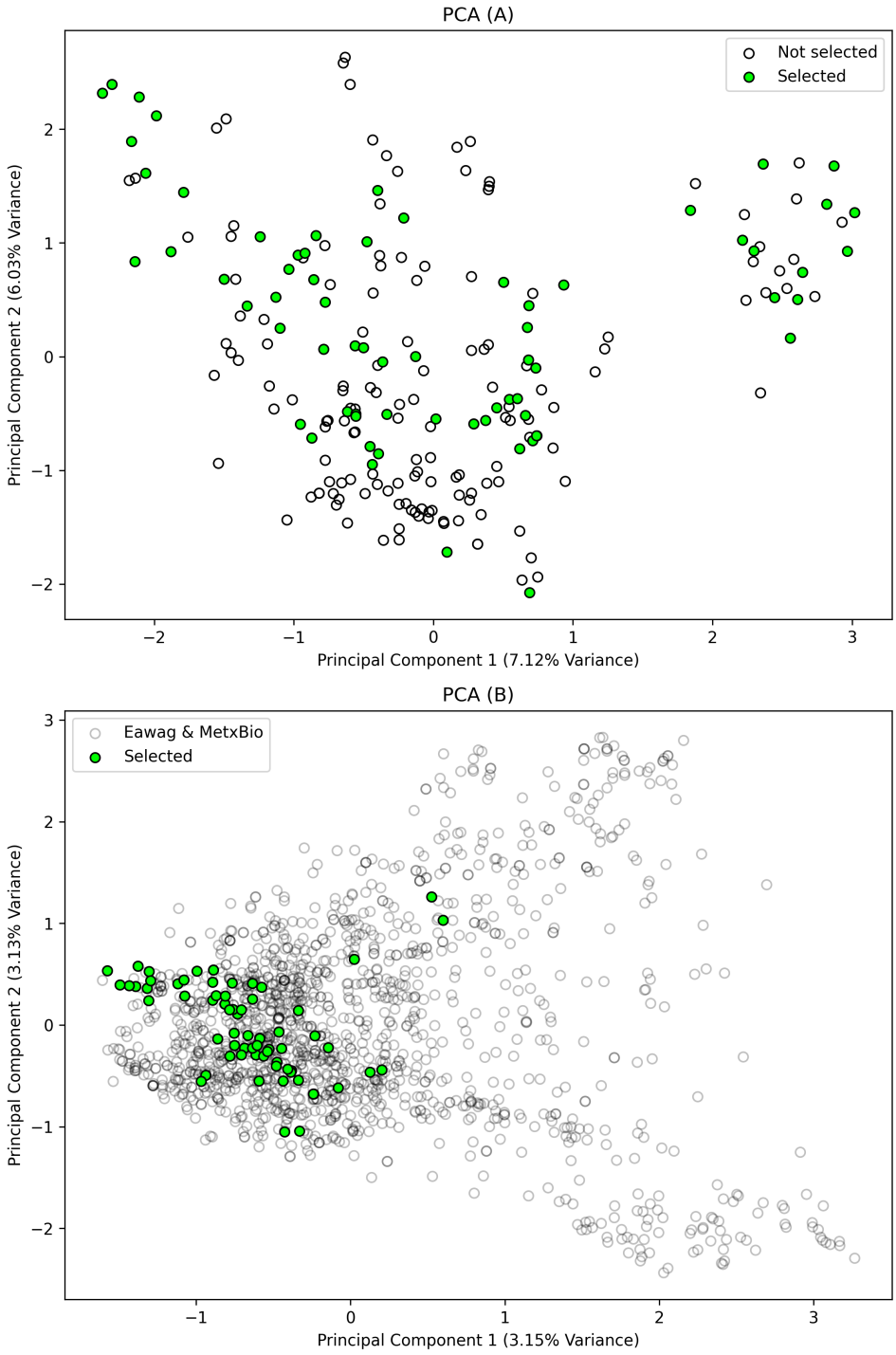

Figure 5.5: Principal component analysis (PCA) (A) shows the chemical similarity between the 67 selected parent and biotransformation products (including the anchor chemical) for the ionisation efficiency study (green) and the collated commercial database of 213 parent and biotransformation products (black circles) with PC 1 and PC 2 explaining 7.12% and 6.03% of the variance, respectively. PCA (B) shows the chemical similarity between the 67 selected parent and biotransformation products (including the anchor chemical) (green) and the Eawag parent-biotransformation pair and MetXBioDB databases (black circles) with PC 1 and PC 2 explaining 3.15% and 3.13% of the variance respectively.

5.4 Discussion

This research has resulted in the development of a random forest regression model for the prediction of $\log_{10}$RIE values for 67 parent and BTPs, as a route towards the semi-quantification of BTPs with the overarching aim of integrating BTP concentration predictions into TK models. The performance of the model was analysed by comparing the predicted $\log_{10}$RIE against the experimental $\log_{10}$RIE for the training and test chemical sets and calculating the RMSE of $\log_{10}$RIE and Pearsons R^2 values. The RMSE of $\log_{10}$RIE for the test set was 0.77 compared to 0.37 in the training set, with R^2 values of 0.64 and 0.92 for the test and training set, respectively. Comparison between the evaluation metrics of the training and test sets showed the model was able to predict the $\log_{10}$RIE of unseen data relatively well and could have generalisability to other datasets within the same $\log_{10}$RIE ranges. The model performed comparably to other $\log_{10}$RIE predictive models, such as the multi linear regression model developed by Oss et al. (2021) on 334 chemicals, which had a RMSE in the range of 0.7 - 0.8 log IE units compared to the RMSE of the test set in this study of 0.77 [202]. Additionally, the Bayesian ridge regression model related to molecular fingerprints developed by Mayhew et al. (2020) resulted in an RMSE of 0.362 and R^2 of 0.62 compared to the RMSE and R^2 test set values in this study of 0.77 and 0.64, respectively [173]. Moreover, the RMSE obtained from predictions of IE in different biological matrices varied between 0.36 and 1.31 and R^2 varied between 0.55 and 0.81 [157]. Furthermore, the RMSE of $\log_{10}$RIE for the test set had a multiplicative factor of 6. While not directly comparable these results are within the same order of magnitude as Liigand et al. (2020) findings that showed a 2.2 times prediction error in IE values [156]. It is important to highlight that many studies are restricted to specific groups of similar chemicals, which limits the application of the established models and will generally improve results for similar chemicals as the model is learning from a homogenous group of chemicals in terms of structure and function. Alternatively, in this study a wide range of parent and BTPs were included in the model regardless of chemical grouping or function, which makes modelling more challenging because of the diversity

of chemical structures for the model to learn from and results in dependency on the test and training split especially with a smaller group of chemicals.

To further evaluate the generalisability of the model a 10-fold cross validation of the training set was performed. The average RMSE of $\log_{10}$RIE was virtually the same as the RMSE test set but higher than the RMSE training set. It could be argued that the model overfits to some of the specific characteristics in the training set but on average it performs well and generalises to the different subsets of the data. The average RMSE had a central tendency that shows it represents the performance of the model well and highlights the stability of the model in different scenarios, however, at the extreme end the RMSE was 1.54. The minimum RMSE of $\log_{10}$RIE had a multiplicative factor of approximately 2. However, the maximum RMSE of $\log_{10}$RIE had a multiplicative factor of 35. This highlights the significant variation in performance between the different folds of the cross-validation. Further work could investigate the subsets of chemicals with the worst RMSE to understand the potential characteristics that result in poor prediction performance. Finally, the robustness of the model was tested by performing a y-scramble on the test set to evaluate whether the model performance results were a result of randomness. Through 1000 random iterations the RMSE and Pearson R^2 values were never better than the RMSE and Pearsons R^2 test set values, which suggests there is a relationship between the PaDEL descriptors and the $\log_{10}$RIE that the model is able to capture from the training data.

Even though the model performs relatively well on unseen data and the results of the model outperformed randomness further optimisation of the model performance could be achieved by increasing the number of the parent and BTPs in the training set. It has been argued that prediction models should be continually updated with new data to improve the coverage of chemical structures [143]. Sixty-seven total parents and BTPs including the anchor is relatively small in comparison to some predictive models that have been shown to perform well. For example, the random forest regression model developed by

Liigand et al. (2020) had 353 unique chemicals in positive ion mode and 109 in negative ion mode [156], while Oss et al. (2021) developed a model on 334 chemical IE values [202]. It would be interesting to investigate the performance of these other predictive models developed on a wider set of chemicals on the 67 parent and BTPs in this study. Additionally, in this study the model chosen was the random forest regression. However, it is plausible that other models could provide similar or better predictions. For example, other machine learning algorithms have shown promise in predicting IEs with examples including artificial neural networks [204] and Bayesian ridge regressions [173]. However, while other methods can provide good predictions they may be less interpretable than the random forest regression, which has clear feature importance measures. Alternatives, such as, artificial neural networks have complex structures that are considered black-box models, due to the inability to interpret the internal logic of the network [205].

Three of the most influential PaDEL descriptors represented by the highest SHAP values and qualitatively had the best colour and value separation were auto-correlated descriptors ("ATSC8m", "AATSC3s", and "ATSC8p") proposed by Moreau & Broto [180]. Auto-correlation descriptors are able to encode the chemical structure in addition to numerical properties atoms properties in a 2D conformation [118, 89]. Each autocorrelation descriptor highlighted in this study is weighted by a different physiochemical property and accounts for different atoms at various lengths of bonds away from each other. "ATSC8m" is the total mass of the terminal atoms that are 8 bonds away in the molecular structure [274]. "AATSC3s" is the intrinsic state of the atoms 3 bonds away in the molecular structure, which is related to the electrotopological state [109] "ATSC8p" is the polarisability of the atoms 8 bonds away in the molecular structure. The other important descriptors included "C1SP2" that relates to the number of double bonded carbons bound with another carbon [127], and the final descriptor was "MaxHother", which describes the maximum hydrogen electrotopological state for hydrogen atoms present on aromatic carbons, double bonded substituted carbon, and double bonded unsubstituted carbon [226]. Basicity descriptors, the size of the molecule, and hydrophobicity / charge (de) localisation of the

chemicals were highlighted as the most influential parameters in a multiple linear regression of a IE dataset of 334 chemicals spanning six orders of magnitude [202]. Therefore, it is unsurprising that the most influential descriptors in this study are also related to mass, polarisability, and descriptors related to electrotopological state. Oss et al. (2021) showed through PCA analysis of compounds and chemical properties that parameters related to size, such as mass, area, and molecular volume were all clustered together [202]. This may explain the reason other key parameters were not identified as the most influential in this study as there are many that have co-linearity and were therefore removed during processing. Further work could evaluate the relationships between these specific 2D descriptors that are hard to interpret and other generally explainable descriptors through correlation analysis.

The PCA of the selected parent and BTPs compared to those available commercially and compared to the Eawag and MetXBioDB databases highlighted that the chemical space of the commercially available parent and BTPs were captured well in this study. However, there are clusters of chemicals in the Eawag and MetXBioDB databases not captured in the selected chemicals. This suggests an even wider structural coverage of parent and BTPs could potentially improve the domain of applicability of the model. As previously stated, the availability of standards commercially available is limited but with more time and resources the number of parents and BTPs in the model could be increased. A unique use of this PCA to map chemical space would be in future studies to compare potential parent and BTP chemical structures before experimentation to make sure that similar structures were not already included in the model. This would enable the focus of resources on chemical structures not already accounted for within the model. Another use of the PCA would be to identify parents and BTPs that are representative of the data set through chemical similarity and use them for the test set instead of a random training test split, which has greater statistical fluctuations in smaller datasets [6].

A key ambition of this work is to enable the semi-quantification of BTPs that can be

utilised in TK models. Therefore, a reliable RIE predictive model that captures the relationship between the RIE and chemical descriptors is the first major step in this study. To achieve the long term ambition, the next phase of this research should focus on using the RIE predictive model to calculate concentrations of BTPs without reference standards using a comparable methodology presented by [156, 143, 203]. Firstly, RIEs are converted back to absolute IE values. Then a set of instrument specific calibration chemicals (internal standards) with known concentrations measured using the MS to obtain intensity values. Instrument specific RF values are calculated by dividing the measured intensity by the known concentration of each calibration chemical. The instrument specific RF account for variations in the instrumentation setup and result in a more accurate prediction of the BTPs of interest, which would have measured intensities using the same MS configuration. These instrument specific RFs are correlated with predicted IE values from the random forest model for the calibration chemicals using a linear regression, which enables the predictions of RF values. Finally, the predicted concentration of a BTP without a standard can be calculated assuming the structure is known. Firstly, the BTP IE value can be predicted from the random forest regression using calculated chemical descriptors from its structure. This is followed by inputting the IE prediction into the predicted RF linear regression and dividing the measured intensity of the BTP by the predicted RF to predict the BTP concentration. This method can then be validated by comparing the predicted concentrations against the known concentrations for a set of BTPs and analysing the prediction accuracy as described by [156]. If it possible to predict BTP concentrations within a reasonable margin of error then the need for standards for every BTP, which is not feasible, would be removed and these semi-quantitative values could be implemented in TK models. This could either be as specific TK compartments or as a calculated rate of change between the parent and BTP in the form of a biotransformation rate, which are key for understanding the true accumulation potential of chemicals [146].

To make the work more relevant to *D. magna* the impact of matrix effects also need to be understood. Most research in this area has been carried out using solvent mixtures, which

makes application to complex biological mixtures difficult [157]. Matrix effects can cause a change in the response of the chemical and effect its measurement accuracy in the MS [186]. It has been shown that IE of warfarin can differ by an order of magnitude between urine and solvent matrices [157]. Therefore, further work should conduct IE experiments with *D. magna* exposed to a select number of parents and BTPs to establish the matrix effects. Consequently, this could enable a *D. magna* specific IE predictive model, which would be an important step in achieving the integration of biotransformation predictions into *D. magna* TK models.

5.5 Conclusions

To conclude, this work demonstrates the capabilities of the random forest regression model to predict the $\log_{10}$RIE of parents and BTPs from PaDEL descriptors. In comparison to other predictive models, the performance quality was relatively similar. Additionally, the cross-validation of the training set highlighted the potential generalisability of the model to unseen data while the y-scramble illustrated that there was a significant relationship between the descriptors and the $\log_{10}$RIE values captured by the model. The data from this study is an essential contribution to the IE research area where BTPs have not generally been evaluated and can be used and accessed by the IE and TK modelling community. Further work could increase the structural diversity of the parent and BTP experimental RIE values unaccounted for within the new model reported here, which would improve the domain of applicability and potentially improve prediction performance. The work undertaken is a first step towards predicting the concentrations (termed semi-quantifying) of BTPs. Consequently, the early progress described here builds confidence in this approach towards supporting improved ERA and TK modelling, with the ability to predict concentrations of BTPs without internal standards, which could improve the understanding of the accumulation potential of BTPs.

CHAPTER 6 : CONCLUSIONS & FUTURE WORK

The overarching aim of the research presented in this thesis is to advance theoretical and probabilistic methods for toxicokinetic predictions in *Daphnia magna* for use in ERA. This aim is motivated by the significant increase in the variety and volume of chemicals used globally in the last few decades. This growth has resulted in increased pressure to generate toxicity data for these chemicals, to perform safety assessments, and to evaluate potential impacts on the environment. The need to generate toxicity data has motivated a necessary shift towards NAMs, including *in silico* and *in vitro* methods to replace traditional toxicity testing and help inform chemical safety decisions without the need for animal testing, which is low-throughput, expensive, and ultimately raises ethical concerns. The development of *in silico* TK models for environmentally relevant species, such as *D. magna*, are crucial for a NAMs-based safety strategy.

In silico methods are necessarily model and data driven. In Chapter 1, a comprehensive analysis of literature highlighted limited availability of quantitative *D. magna* TK data and a modelling approach focused on understanding chemical specific TK processes rather than a general modelling framework that applies to a large chemical space (as required for a modern NAMs-based ERA). Primarily, approaches for modelling TK data are deterministic point-wise predictions that rely on nonlinear regressions as opposed to probabilistic descriptions of predictions and (hierarchical) model parameters. In Chapter 2, a novel Bayesian framework is developed from limited available historical *D. magna*

TK data and evaluated as a NAM for ERA.

The literature review in Chapter 1 identified several key research questions relevant for TK modelling in *D. magna*. Ionisation has been shown to have an effect on TK predictions in fish, therefore, in Chapter 3 the effect of ionisation is investigated within a Bayesian framework by using D_{ow} over K_{ow}. Additionally, protein binding is a key parameter in human PBTK and is known to impact the pharmacological relevant concentration. In Chapter 4, a theoretical protein surface-binding model for *D. magna* that accounts for the protein binding as a function of the protein fraction and external concentration is developed through decomposition analysis. Finally, the biotransformation of chemicals has been highlighted as a key process in TK as it can lead to overestimations of the parent chemical and/or result in more toxic BTPs. *D. magna* BTP TK data is virtually absent from the literature and commercially available standards that are needed for their absolute quantification are limited. Chapter 5 defines a semi-quantification method using relative ionisation efficiencies as a first step to predicting BTP concentrations that can be implemented into semi-quantitative TK modelling of BTPs.

6.1 Bayesian framework conclusions

In Chapter 2 a proof-of-concept Bayesian analysis was developed to estimate the steady-state internal concentrations (and their ratios) on TK time-course data for *D. magna* from the *AquaTK* dataset. The Bayesian framework consists of two components, an inference of the steady-state internal concentration and an inference with a predictive component. The inference of the steady-state internal concentration assumes an exponential decay diffusion profile while the predictive model uses information about the experiment's external water concentration and the chemical's lipid partitioning. An important feature of a Bayesian inference and prediction framework is the common-sense probabilistic interpretation of the results. The inferred concentration ratio is a distribution over plausible concentration ratio values and so the prediction is also a distribution over all plausible values of the

concentration ratio.

In the proposed Bayesian framework in Chapter 2, the predictions can be intuitively benchmarked against the inferences by comparing the 95th percentiles of the predicted and the inferred steady-state concentration ratios. On average the 95th percentile of the predicted steady-state concentration ratios was 8-fold higher than the concentration ratios from the experimental data for 96% of the data. Therefore, using the 95th percentile in any of the data availability scenarios would be a conservative estimate and essentially avoids, in 95% of cases, underestimating the true concentration ratio values. Using the higher percentile provides a risk assessor with a calibrated uncertainty of the estimate, which is more meaningful than arbitrary safety factors usually applied to TK data.

In the atrazine case study in Chapter 2, atrazine data was presented at three different levels of precision given three different data availability scenarios (low to high). The multi-level modelling approach shows how Bayesian inference adapts its ability to improve (uncertainty reductions) in predictions of steady-state concentration ratios for increasing amounts of data availability. In a world with limited data it has been demonstrated that *in vivo* data can be readily substituted by mechanistic based information with substantial gains in performance, robustness, and transparency, resulting in an extremely valuable modelling approach that accurately reflects uncertainty given the lack or abundance of data.

A unique aspect of the analysis performed in Chapter 2 is the inclusion of only the uptake-phase time-course profile. Previous studies required both uptake and elimination time-course data to be present. However, relaxing this requirement and focusing on estimating steady-state concentration ratios, meant that the number of suitable studies was increased significantly. Moreover, the analysis is able to estimate steady-state concentrations even for singular data points and combine that information between studies where applicable, which is one of the benefits of using an integrated framework. This feature means that ERAs applying a Bayesian framework can estimate predictions and quantify uncertainties

where experimental data is limited.

Many of the chemicals in the *AquaTK* dataset are ionisable at environmentally relevant pHs, which if accounted for might affect prediction error. This effect is studied in Chapter 3 for neutral and ionisable chemicals within the Bayesian framework. The ionisation effect is first studied in a deterministic setting by comparing NLOM steady-state concentration ratios predictions using first K_{ow} and then D_{ow}. Mean fold-error is reduced from 11 to 9 when using D_{ow} over K_{ow} suggesting that, in a deterministic setting, accounting for ionisation improves predictions for the NLOM model. Although D_{ow} improved overall prediction error, predictions for ionisable chemicals using D_{ow} consistently underpredicted; underpredictions in ERAs are undesirable in general, suggesting that applying the NLOM model with D_{ow} for *D. magna* predictions is not beneficial in an ERA.

In Chapter 3 the evaluation of the effect of ionisation in the Bayesian predictive model showed that the magnitude of prediction error gets larger when ionisation is accounted for using D_{ow}. Overall, the prediction errors of the Bayesian prediction model were smaller than the NLOM model with geometric means of fold error of 3.4 and 4.3 respectively. A key feature of the Bayesian model is that the steady-state concentration ratios regress towards the mean, which bounds the prediction error variance by the steady-state concentration ratio variance in the *AquaTK* dataset.

The cross-validation of each study ID to test the sensitivity of the Bayesian parameters to the input data in Chapter 3 highlighted that the expected value of the standard deviation parameter was similar across all cross-validation folds apart from when source ID 14 removed. Consequently, the largest prediction errors were primarily neutral chemicals, which suggests that factors other than ionisation need to be considered to improve model performance.

Chapters 2 and 3 provide concrete examples of applying Bayesian methods in different settings relevant for ERAs: steady-state concentration ratio inference at varying levels of data availability, probabilistic predictions that are consistent and intuitive to benchmark

against inferences, hypothesis and model testing and evaluation, complex data structure modelling, and an example of an ERA that combine all of the components together.

6.2 Bayesian framework future work

Chapters 2 and 3 provide a proof-of-concept Bayesian framework for the quantification of uncertainty in *D. magna* TK predictions and the ability to compare parameters to investigate the effects of ionisation on prediction performance. As a proof-of-concept, the Bayesian framework used the experiment's external water concentration and the chemicals lipid partition coefficient for predictions because they were two experimental factors that were expected to significantly affect the internal concentration. The benefit of the Bayesian approach is the ability to include any number of parameters to make more accurate inferences and predictions. Future work should aim to expand the Bayesian predictive model with other key parameters, such as temperature, wet weight, or size.

In Chapter 3, to improve the NLOM model predictions rather than treat each partition coefficient independently a geometric mean of K_{ow} and D_{ow} predictions could be used to achieve a closer prediction to the target steady-state concentration ratio. Furthermore, different models, such as the NLOM model, could easily be implemented into the Bayesian framework and the performance of the various models predictive performance evaluated in a consistent and open framework.

The Bayesian framework in Chapters 2 and 3 could be advanced further by potentially including important structural features of the chemical including the number of hydrogen acceptors and donors or number of halogenated atoms. Collectively, an important next step would be to test the Bayesian framework on a wider set of chemicals and environmentally relevant species to establish the domain of applicability. Even though ECOTOX database is generally only point estimates it could be useful for a more extensive analysis to test whether the 95^{th} percentile could serve as a conservative estimate across all ERA relevant species and chemicals.

6.3 Theoretical protein surface-binding model conclusions

In Chapter 4 a theoretical protein surface-binding model for predicting TK predictions in *D. magna* was developed to integrate external concentration as a parameter. As a result of recent work in *G. pulex* by Raths et al. (2023), it has been suggested that the external concentration measured effect is related to protein surface-binding [215]. The PSB model provides an expression for the concentration ratio upper bound expressed as a function of the external concentration and the protein fraction of the organism.

A theoretical upper-bound was evaluated against the inferred *D. magna* steady-state concentration ratios from Chapter 3 to show that the theoretical model is valid for a range of external concentration scenarios. Additionally, the PSB model was used as a predictive model and benchmarked against the NLOM model, which resulted in 70.21% of time-courses within 10-fold error while the NLOM model predicted 59.57% time-courses within 10-fold error. This will have positive implications for ERA as the PSB model can estimate TK predictions across any external concentration scenario with minimum data requirements and without the need for experimental data. A risk assessor would only need the lipid and protein fractions, which are readily available in this work and in the literature, an *in silico* partition coefficient value, and the chosen external concentration scenario to predict an upper bound of the steady-state concentration ratio. The true value of the *D. magna* steady-state concentration will in most cases be less than the predicted value, but this allows a conservative estimate for use in ERA.

6.4 Theoretical protein surface-binding future work

As the theoretical protein surface-binding model was shown to successfully capture TK of *D. magna* and *G. pulex*, the next logical phase for further work would be to evaluate the model across multiple trophic levels and an extended array of chemicals. This phase was outside the scope of the PhD research, however, an initial investigation of the application of the theoretical upper bound using similar methods as in Chapter 4 was conducted on the

Arnot & Gobas. (2006) database that is also contained in the ECOTOX database [17]. A dataset was curated of 56 species covering both aquatic vertebrates and invertebrates apart from algae and only included the highest quality rated data (criteria score = 1). The total number of high quality steady-state concentration ratio values equalled 2,315. CompTox Chemical Dashboard was used to predict $\log_{10} D_{ow}7.4$ values for all the chemicals [253]. As there were a range of species the lowest protein fraction was set to 0.01 and the highest lipid fraction was set to 0.1. The external concentration scenarios were chosen as 0.026, 1, 5.4, 50, 300, 1000 $\mu g\ kg^{-1}$ (corresponding to the 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th}, and 95^{th} percentiles).

Figure 6.1 shows the steady-state concentrations ratios plotted against the $D_{ow}7.4$ values from CompTox with the PSB model theoretical upper bound represented by the red dashed line. This initial investigation suggests that the PSB model is applicable to a wider range of environmentally relevant species, which suggests this model could have widespread positive impacts on ERA with predictions without the need for animal testing. Future work could focus on quantifying the performance for different species and identifying any outliers that break the theoretical upper bound.

From Chapter 4 for $1 < P < (\frac{1}{\rho} - 1)$ the PSB model defines a maximum theoretical steady-state concentration given the protein and lipid fraction. The use of a PSB model as a predictive model was shown to perform more accurately than the NLOM model in predicting *D. magna* steady-state concentration ratios. However, there was still large variance in prediction performance across the chemical space. There are many processes that will affect the accumulation of the chemical resulting in a concentration ratio less than the upper bound, e.g. biotransformation, inhomogeneous chemical distribution, and imperfect diffusion, due to respiration. Therefore, the loss of chemical mass from the maximum upper bound can be represented by a random variable,

$$\text{Loss} := \log\left(\frac{\text{PSB prediction}}{\text{Measurement}}\right) > 0, \quad \text{Measurement} \leq \text{PSB prediction} \tag{6.1}$$

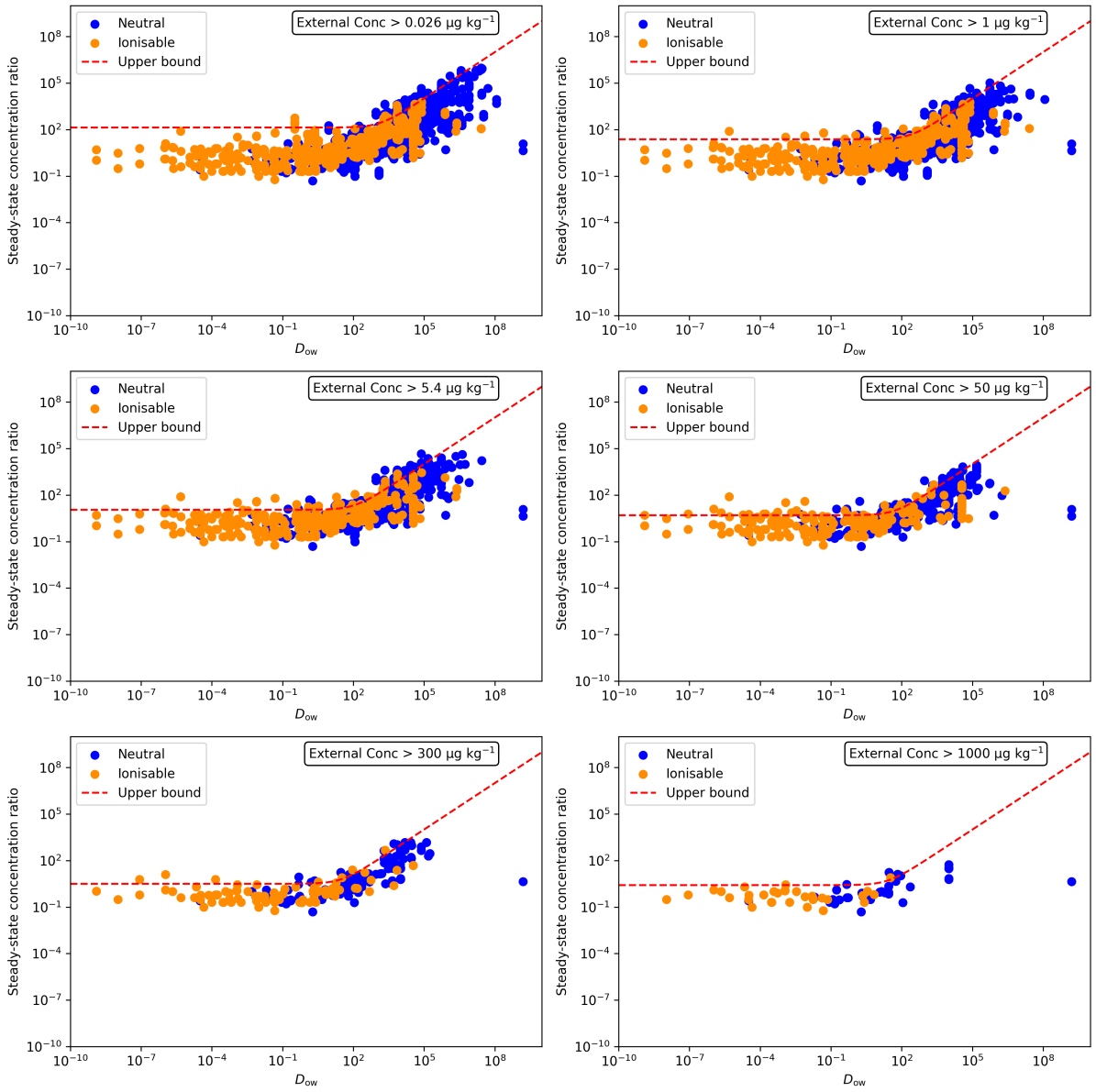

Figure 6.1: 2315 steady-state concentration ratios across 56 species of aquatic vertebrates and invertebrates from [17] and ECOTOX database plotted against the D_{ow}7.4 of the chemical with the theoretical upper of the protein surface-binding (PSB) model (red dashed line) plotted for each external concentration scenario (0.026, 1, 5.4, 50, 300, 1000 $\mu g\ kg^{-1}$). Steady-state concentration ratio data was only included if the external concentration met the threshold of the scenario. Ionisable (orange) and neutral (blue) chemicals were highlighted.

where the measurement is the concentration ratio and the PSB prediction is the predicted steady-state concentration ratio based on the maximum upper bound. The measured concentration ratios were taken from the ECOTOX database again but a range of physiochemical descriptors were predicted from ADMET Predictor^{TM} software, which

resulted in marginally more concentration ratios being available. Overall, there was 2,444 steady-state concentration ratios defined as high quality to use for this analysis. As the external concentrations are the same the loss is only determined by comparison of the internal concentrations. Figure 6.2 plots the steady-state concentration ratio prediction for the PSB and NLOM models against the measured steady-state concentration ratios from the ECOTOX database. It is clear that most of the steady-state concentration ratios are bound by the PSB theoretical upper bound. Comparatively, the NLOM model predictions are not bound and tend to underpredict compared to the measured data from the ECOTOX database.

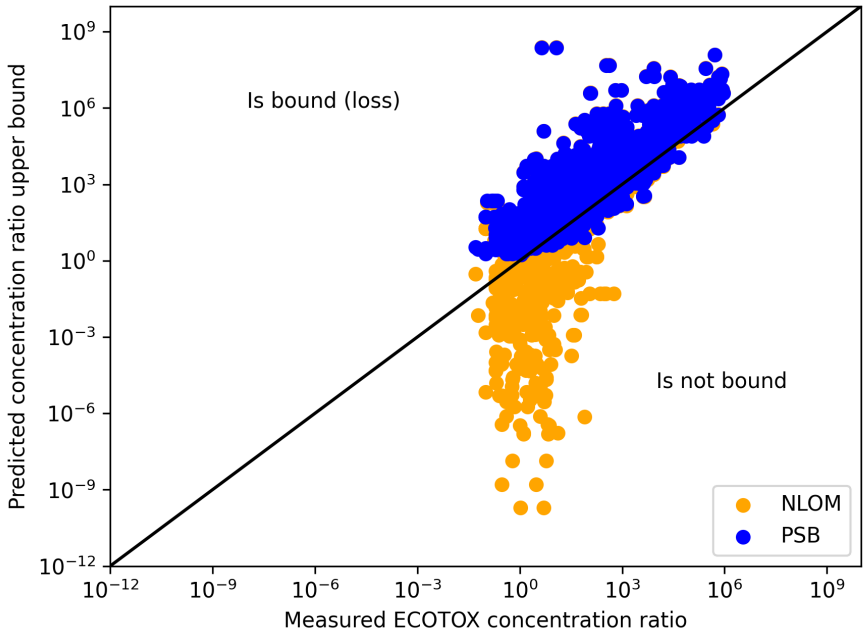

Figure 6.2: Predictions from the protein surface-binding (PSB) model (blue) and the non-lipid organic matter (NLOM) model (orange) plotted against the measured steady-state concentration ratio from the ECOTOX database for 2444 high quality steady-state concentration ratios. The black bisection highlights those predicted steady-state concentration ratios captured by the theoretical upper bound of the PSB model with a positive loss and those not bound.

A simple predictive model of the "true" internal organism concentration can be developed using the loss where the expected true internal organism concentration can be estimated using the mean of the loss as a correction factor. Figure 6.3 shows the distribution of

the internal concentration loss across the steady-state concentration ratios with the mean value represented by a red dashed line.

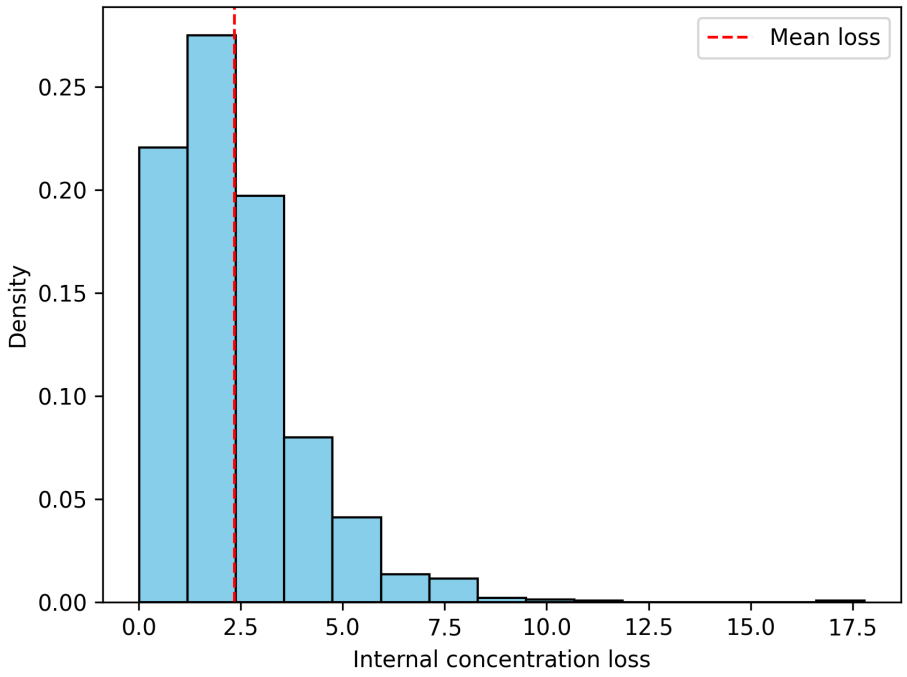

Figure 6.3: Distribution of the internal concentration loss across the 2,344 high quality steady-state concentration ratios with the mean loss represented with a red dashed line.

A more advanced predictive model can be constructed by looking for correlations between the loss and chemical ADMET properties. Following similar methods to Chapter 5 a random forest regression model was developed with the aim of predicting the internal concentration loss. As a simple proof-of-concept the steady-state concentration ratios from the ECOTOX database were used as the training set and the *AquaTK* dataset was used an independent test set. The loss was calculated using Equation 6.1 and only positive loss was used to train the model and predict positive loss in *D. magna*. If the loss was negative it would be mean it was breaking the theoretical upper bound and therefore would not be applicable for regression prediction. The default 50 ADMET predicted chemical properties were used as the X variable and the internal concentration loss as the y variable. As a result of chemicals needing positive internal concentration loss and ADMET prediction properties the ECOTOX and *AquaTK* datasets were reduced to 2,237

and 14, respectively. The predicted internal concentration loss was plotted against the measured loss for the training set (ECOTOX) and test set ($AquaTK$) (Figure 6.4). There was a strong prediction performance on the training set (ECOTOX) with a R^2 value of 0.96. Moreover, there was a good prediction performance on the test set ($AquaTK$) with a R^2 value of 0.85. Overall, the predictive model was able to predict the *D. magna* internal concentration loss within 3-fold of the measured loss, which is a significant improvement on the safety factors usually applied to TK data.

The number of times a descriptor was chosen as the root node was used to evaluate the most important descriptors. The root node describes the first descriptor used to split the data. If a descriptor is consistently chosen as the root node across different trees in the random forest it highlights its importance. This resulted in the descriptor with the highest (negative) correlation being the molecular diffusion coefficient. This suggested that the lower the diffusion coefficient the greater the internal concentration loss. It could be argued that chemicals with a low diffusion coefficient will have limited absorption and struggle to cross the membrane barriers of organisms. Consequently, leading to reduced bioavailability of the chemical and less accurate predictions of the internal concentration. Further investigation could identify a range of neutral chemicals (limit influential co-factors) that have a range of diffusion coefficients and try to evaluate the relationship with the internal concentration loss. The loss could be a function of the diffusion coefficient and be implemented as a correction factor into the PSB model to potentially improve prediction performance.

Another avenue of future work is an estimation of a tighter upper bound by representing the protein affinity P as a pharmacokinetic parameter or physiochemical property. In Chapter 4 this is initially discussed and in Appendix C.3.1 the mathematical derivation of this idea is presented. One key parameter that could be related to the protein affinity is the fup. A key assumption of the protein surface-binding model is that all proteins interact with the chemical in a similar way. Therefore, fup was chosen as it is a key

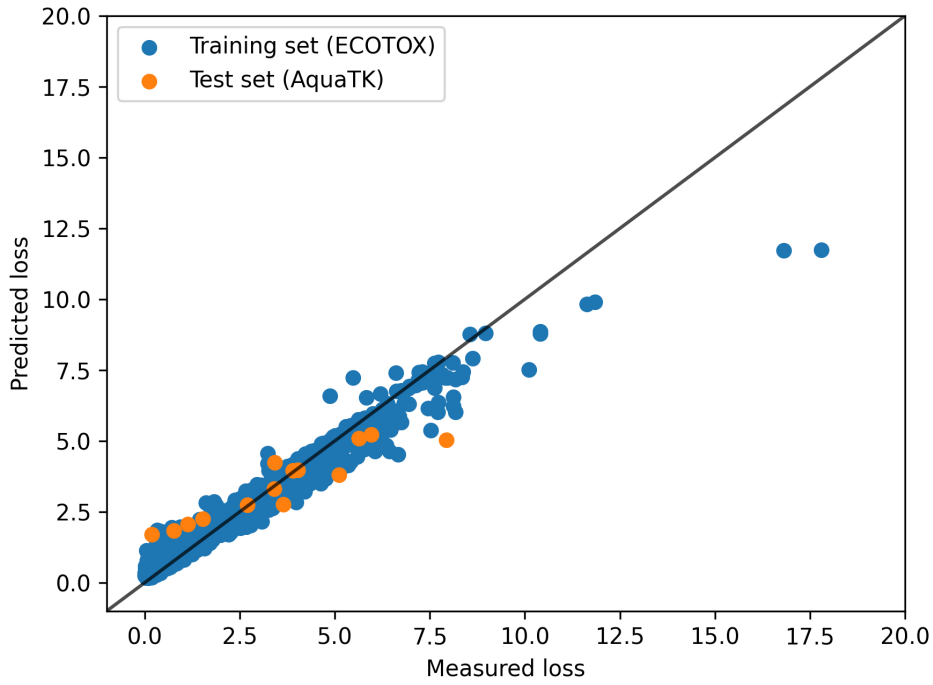

Figure 6.4: Predicted internal concentration loss from the random forest regression model against the measured internal concentration loss for the training set (2,237 positive internal concentration loss values from the ECOTOX database) and the test set (14 positive internal concentration loss values from the *AquaTK* dataset). The R^2 of the training (ECOTOX) and test set (*AquaTK*) were 0.96 and 0.85, respectively.

parameter in human PBTK, which is supported by extensively evaluated empirical data rather than other specific proteins that have more limited data. An initial investigation was undertaken to relate human and rat fup values with P values and to test whether they break the theoretical protein surface-binding model. Fup values were predicted using chemical canonical SMILES in ADMET Predictor^{TM} software. Table 6.1 shows the method to test whether the protein affinity breaks the density (i.e. $P > ((1/\rho) - 1)$) when human and rat fup values are used instead. The method for estimating fup values is presented in detail in Appendix C.3.2. An example would be prochloraz for rats that has a $fup = 6.23\%$ giving an estimate for $P \approx 15$. This is then compared to the protein surface-binding density ($\rho = \rho_p^{(S)}$) in normalised units, which results in an external concentration, $y_w = 640.322$ μg / L = 6.4e-6 kg / L. The protein surface-binding density is then estimated as $\left(\frac{1}{0.034}\right)^{\frac{2}{3}} \sqrt{6.4 \times 10^{-6}} \approx 0.024$, and the bound is set at $\frac{1}{0.024} - 1 \approx 40 > 15$. This would

mean that it would not break the theoretical model, which is represented by a "FALSE" in Table 6.1. Overall, the predictions for P would not break the model for human and rat fup values. Considering the substantial differences between the biochemical composition of *D. magna* and humans and rats and the fact the P predictions are within an order of magnitude suggest there is some relationship between the protein affinity and predicted fup values. However, from the analysis of the theoretical upper bound in Chapter 4 it is clear that some chemicals break the model. Therefore, to better replicate the *D. magna* dataset specifically further work could work on deciphering a scaling metric or correction factor between the fup values of humans and rats and *D. magna* or re-analyse the dataset with *D. magna* specific fup data using *in vitro* assays. Alternatively, research could decipher other surrogates of protein affinity. For example, protein affinity has been shown to depend on the lipophilicity of the chemical, which suggests some function of K_{ow} could represent P and be tested in a similar framework.

Table 6.1: Relationship between the predicted fraction unbound in the plasma ($fup\%$) and theoretical protein surface-binding model. Information includes Chemical name, Source ID, external concentration in μg/L & kg/L, the protein surface-binding density ($\rho = \rho_p^{(S)}$), the upper bound for $((1/\rho) - 1)$, the predicted human ($fup_{hum}\%$), the predicted rat fup ($fup_{rat}\%$), P values for humans (P_{hum}) and rats (P_{rat}) based on the predicted fup values, human and rat relative P values ($P_{\text{rat}}/P_{\text{human}}$), TRUE/FALSE whether the predicted P values for humans and rats broke the assumption that $P < ((1/\rho) - 1)$, and the predicted P values for humans and rats normalised by the upper bound $((1/\rho) - 1)$.

Chemical name	Source ID	Ext. Conc. μg/L	Ext. Conc. kg/L	ρ	$((1/\rho) - 1)$	$fup_{hum}\%$	$fup_{rat}\%$	P_{hum}	P_{rat}	P_{rat}/P_{hum}	$P_{hum} > ((1/\rho) - 1)$	$P_{rat} > ((1/\rho) - 1)$	$P_{hum}/((1/\rho) - 1)$	$P_{rat}/((1/\rho) - 1)$
2,4-Dichlorophenoxyacetic Acid	d12	1000	0.00001	0.030131035	32.18837164	7	4.64	13.28571429	20.55172414	1.546903967	FALSE	FALSE	0.412748878	0.638482877
Atrazine	d12	100	0.000001	0.00952827	103.9508462	18.9	10.6	4.291005291	8.433962264	1.965497988	FALSE	FALSE	0.041279176	0.081134138
Carbamazepine	d08	100	0.000001	0.00952827	103.9508462	9.07	13.22	10.02535832	6.56429652	0.654769267	FALSE	FALSE	0.096443258	0.063148082
Carbamazepine	d08	5	0.00000005	0.002130586	468.3544529	9.07	13.22	10.02535832	6.56429652	0.654769267	FALSE	FALSE	0.021405494	0.014015659
Chlorpyrifos	d05	0.08	8E-10	0.0002695	3709.572753	5.93	5.34	15.86340641	17.72659176	1.117451782	FALSE	FALSE	0.004276343	0.004778607
DDT	d12	10	0.0000001	0.003013104	330.8837164	2.54	2.02	38.37007874	48.5049505	1.264134766	FALSE	FALSE	0.115962427	0.146592135
Diazinon	d02	1.5821	1.5821E-08	0.001198481	833.3897866	10.61	17.87	8.425070688	4.595970901	0.545511257	FALSE	FALSE	0.0101094	0.005514791
Diazinon	d02	1.4604	1.4604E-08	0.001151463	867.4604245	10.61	17.87	8.425070688	4.595970901	0.545511257	FALSE	FALSE	0.00971234	0.005298191
Diclofenac	d09	100	0.000001	0.00952827	103.9508462	3.73	2.74	25.80965147	35.49635036	1.375313045	FALSE	FALSE	0.248287074	0.341472452
Diclofenac	d09	5	0.00000005	0.002130586	468.3544529	3.73	2.74	25.80965147	35.49635036	1.375313045	FALSE	FALSE	0.055107091	0.075789501
Fluoxetine	d10	5	0.00000005	0.002130586	468.3544529	15.89	9	5.293266205	10.11111111	1.910183754	FALSE	FALSE	0.011301838	0.021588588
Fluoxetine	d10	0.5	0.000000005	0.00067375	1483.229101	15.89	9	5.293266205	10.11111111	1.910183754	FALSE	FALSE	0.003568745	0.006816958
Imidacloprid	d04	240	0.0000024	0.014761132	66.74547994	58.39	36.72	0.712622024	1.723311547	2.418268715	FALSE	FALSE	0.010676708	0.02581915
Imidacloprid	d04	190	0.0000019	0.013133814	75.13934637	58.39	36.72	0.712622024	1.723311547	2.418268715	FALSE	FALSE	0.009484006	0.022934875
Irgarol	d03	200	0.000002	0.013475009	73.21145506	10.73	7.01	8.319664492	13.26533524	1.594455551	FALSE	FALSE	0.113638835	0.181192072
Kepone	d15	0.175	1.75E-09	0.000398596	2507.80508	8.25	7.28	11.12450867	12.73742967	1.144988067	FALSE	FALSE	0.004435954	0.005079115
Mirex	d15	0.118	1.18E-09	0.000327307	3054.237255	4.97	3.60	19.12630274	26.79815131	1.401115086	FALSE	FALSE	0.006262219	0.00877409
Pentachlorophenol	d16	20	0.0000002	0.004261172	233.6772265	4.40	1.94	21.73167933	50.52885268	2.325124162	FALSE	FALSE	0.092998704	0.216233535
PFDA	d06	5	0.00000005	0.002130586	468.3544529	3.45	4.12	27.98550725	23.27184466	0.831567727	FALSE	FALSE	0.059752837	0.049688531
PFDA	d07	5	0.00000005	0.002130586	468.3544529	3.45	4.12	27.98550725	23.27184466	0.831567727	FALSE	FALSE	0.059752837	0.049688531
PFDoA	d06	5	0.00000005	0.002130586	468.3544529	3.08	3.44	31.46753247	28.06976744	0.89202315	FALSE	FALSE	0.067187431	0.059932744
PFDoA	d07	5	0.00000005	0.002130586	468.3544529	3.08	3.44	31.46753247	28.06976744	0.89202315	FALSE	FALSE	0.067187431	0.059932744
PFNA	d06	5	0.00000005	0.002130586	468.3544529	3.71	4.63	25.9541779	20.59827214	0.793639938	FALSE	FALSE	0.055415674	0.043980092
PFNA	d07	5	0.00000005	0.002130586	468.3544529	3.71	4.63	25.9541779	20.59827214	0.793639938	FALSE	FALSE	0.055415674	0.043980092
PFOA	d06	5	0.00000005	0.002130586	468.3544529	4.06	5.34	23.63054187	17.72659176	0.750155957	FALSE	FALSE	0.050454398	0.037848667
PFOA	d07	5	0.00000005	0.002130586	468.3544529	4.06	5.34	23.63054187	17.72659176	0.750155957	FALSE	FALSE	0.050454398	0.037848667
PFOS	d06	5	0.00000005	0.002130586	468.3544529	3.83	4.1	25.10966057	23.3902439	0.931523699	FALSE	FALSE	0.053612516	0.049941329
PFOS	d07	5	0.00000005	0.002130586	468.3544529	3.83	4.1	25.10966057	23.3902439	0.931523699	FALSE	FALSE	0.053612516	0.049941329
PFUnA	d06	5	0.00000005	0.002130586	468.3544529	3.24	3.74	29.86419753	25.73796791	0.861833568	FALSE	FALSE	0.063764094	0.054954037
PFUnA	d07	5	0.00000005	0.002130586	468.3544529	3.24	3.74	29.86419753	25.73796791	0.861833568	FALSE	FALSE	0.063764094	0.054954037
Prochloraz	d13	640.322	6.40322E-06	0.024110891	40.47503231	4.63	6.23	20.59827214	15.05136437	0.730710045	FALSE	FALSE	0.50891305	0.371867878
Propiconazole	d13	479.108	4.79108E-06	0.020855988	46.94786053	8.19	9.38	11.21001221	9.66098081	0.861817153	FALSE	FALSE	0.238775784	0.205781067
Propranolol	d01	100	0.000001	0.00952827	103.9508462	34.83	35.38	1.871088142	1.826455625	0.976146224	FALSE	FALSE	0.017999739	0.017570378
Propranolol	d01	5	0.00000005	0.002130586	468.3544529	34.83	35.38	1.871088142	1.826455625	0.976146224	FALSE	FALSE	0.003995026	0.003899729
Roxithromycin	d01	100	0.000001	0.00952827	103.9508462	37.62	75.73	1.658160553	0.320480655	0.193274803	FALSE	FALSE	0.015951391	0.003083002
Roxithromycin	d01	5	0.00000005	0.002130586	468.3544529	37.62	75.73	1.658160553	0.320480655	0.193274803	FALSE	FALSE	0.003540397	0.000684269
TBOEP	d14	100	0.000001	0.00952827	103.9508462	24.44	27.63	3.092326636	2.619081324	0.846961409	FALSE	FALSE	0.02974797	0.025195382
TBOEP	d14	20	0.0000002	0.004261172	233.6772265	24.44	27.63	3.092326636	2.619081324	0.846961409	FALSE	FALSE	0.013233325	0.011208115
TCEP	d14	100	0.000001	0.00952827	103.9508462	50.75	39.50	0.970598502	1.531474855	1.577866494	FALSE	FALSE	0.009337091	0.014732683
TCEP	d14	20	0.0000002	0.004261172	233.6772265	50.75	39.50	0.970598502	1.531474855	1.577866494	FALSE	FALSE	0.004153586	0.006553804
TDCPP	d14	100	0.000001	0.00952827	103.9508462	17.65	15.54	4.664776466	5.433986223	1.164897453	FALSE	FALSE	0.044874829	0.052274574
TDCPP	d14	20	0.0000002	0.004261172	233.6772265	17.65	15.54	4.664776466	5.433986223	1.164897453	FALSE	FALSE	0.019962478	0.02325424
Terbutryn	d03	200	0.000002	0.013475009	73.21145506	12.58	8.43	6.949125596	10.8623962	1.563131369	FALSE	FALSE	0.094918556	0.148370172
Tetracycline	d11	1000	0.00001	0.030131035	32.18837164	31.66	22.22	2.158559697	3.500450045	1.621660059	FALSE	FALSE	0.067060233	0.108748901
Tetracycline	d11	100	0.000001	0.00952827	103.9508462	31.66	22.22	2.158559697	3.500450045	1.621660059	FALSE	FALSE	0.020765196	0.033674089
TPHP	d14	100	0.000001	0.00952827	103.9508462	3.38	3.74	28.59908135	25.7510121	0.90041396	FALSE	FALSE	0.275121198	0.247722967
TPHP	d14	20	0.0000002	0.004261172	233.6772265	3.38	3.74	28.59908135	25.7510121	0.90041396	FALSE	FALSE	0.122387114	0.110199066

A final direction for further research could integrate the PSB model into the Bayesian statistical framework and evaluate its model performance and application to ERA. As the previous deterministic models used in the Bayesian framework in Chapters 2 and 3 were essentially placeholder models it would be more appropriate to input the mechanistically derived PSB model. This would hopefully result in improved model predictions compared with the inferred steady-state concentration ratios and reduced uncertainty, which could lead to better estimates for ERA. An interesting comparison would be to re-run the methods in Chapter 2 and 3 to investigate the effect of ionisation on the PSB model in the Bayesian framework. In a deterministic setting the NLOM & PSB model showed improved predictions when using D_{ow} over K_{ow}, whereas the Bayesian model predictions performed better when using K_{ow}. Therefore, it would be interesting to see if a different model choice improved the predictions with D_{ow} within the Bayesian framework.

6.5 Relative ionisation efficiency conclusions

In Chapter 5 a random forest regression model was developed to predict the $\log_{10}$RIE values of parents and BTPs as the first step in a workflow to predicting concentrations of BTPs for use in TK models. Overall, the predictive capabilities of the random forest regression model were demonstrated with a positive correlation between the predicted $\log_{10}$RIE and experimental $\log_{10}$RIE values for the independent test set of parent and BTPs. Further evaluation of the models prediction performance was performed with a 10-fold cross validation of the training set. This analysis highlighted that the average RMSE across each cross validation was equivalent to the RMSE on the training set, which demonstrates the consistency of the model predictions on unseen data. The robustness of the predictive model was tested with a y-scramble. It confirmed that the predictive performance could not be achieved by chance through 1000 random iterations and therefore captures a significant relationship between the calculated chemical descriptors and the $\log_{10}$RIE.

The calculated SHAP values revealed the most influential chemical descriptors on the $\log_{10}$RIE across the training and test set of chemicals. It was clear that the model made the same descriptor choices across the training and test set. The most influential parameters were related to mass, polarisability, and descriptors related to electrotopological state, which is consistent with findings of important descriptors from other studies who have analysed larger and more diverse IE datasets.

An important aspect of the research in Chapter 5 is the generation of RIE data for a group of chemicals that have not typically been studied. The data from Chapter 5 is now available and can contribute to the IE and TK modelling community.

6.6 Relative ionisation efficiency future work

In Chapter 5 the random forest regression model was shown to provide reliable predictions of the $\log_{10}$RIE of the parents and BTPs. Therefore, the next phase in this research should focus on the models implementation into a workflow for predicting concentrations of BTPs without standards that can then be applied to TK modelling. To achieve this goal requires several steps. Firstly, intensities from a set of specific calibration chemicals (internal standards) with known concentrations are measured using the MS. Instrument specific RF are calculated by dividing the measured intensity and the known concentration. These are important to account for any variation in instrumentation setup and will more accurately reflect the concentration of the BTP that will be measured on the same MS configuration. A linear regression between the predicted IE values from the random forest regression model and the instrument specific RF values for each calibration chemical enables the predictions of the RF. Consequently, as long as the structure of the BTP of interest is known the concentration can be determined from its measured intensity in the MS. A predicted IE from the random forest regression is calculated using the chemical descriptors of the BTP. This predicted IE is then used to predict the RF from the linear regression. Finally, the predicted concentration of the BTP is the ratio

between the measured intensity in the MS and the predicted RF value. This method can then be validated by comparing a set of BTPs with known concentrations against their predicted concentrations and measure the prediction accuracy. If it is possible to reliably predict BTP concentrations without the reliance on reference standards within an acceptable margin of error then the semi-quantitative values can be implemented in TK models. These quantitative values can be converted to biotransformation rates for use in TK modelling. For example, imagine there is a *D. magna* TK experiment where the *D. magna* are exposed to a chemical analysed at different time-points over a 48 hour period. The concentration of a parent chemical with a standard could be calculated at each time point using a calibration curve alongside the prediction of a BTP concentration using the intensity measured assuming the structure is known. The structure is important not only for the predictive element of the workflow but for determining the presence of the BTP and its intensity in the MS. As the intensity changes with concentration it is possible to create an absolute quantification time-course of the parent and semi-quantification time-course of the BTP over the 48 hours. The difference between the concentration of the parent and predicted concentration of the BTP can be calculated at each time point and divided by the time interval to obtain a biotransformation rate. This approach will enable biotransformation to be accounted for in TK models, leading to an improved understanding of the accumulation potential of BTPs and enhancement of ERA.

To make the research in Chapter 5 more relevant to *D. magna* future work could investigate the matrix effects on the IE predictions and concentration predictions. The experimental IE data in Chapter 5 was determined in solvents, however, it has been illustrated that IE predictions can change significantly dependent on the matrix. Therefore, to determine the matrix effects future work could re-run the experimental workflow in Chapter 5 but by exposing *D. magna* to each chemical. This would result in a *D. magna* specific IE predictive model that would be more relevant for the prediction of concentrations for use in *D. magna* TK models.

Additionally, the prediction performance and domain of applicability of the random forest regression in Chapter 5 could be improved further by generating more experimental IE data for parent and BTPs not included in the study. The more diverse the chemical dataset the greater the domain of applicability and the more chemicals for the model to learn from and identify stronger relationships between the structure and IE values. Consequently, resulting in more robust IE predictions, which could lead to more accurate concentration predictions.

6.7 Final remarks

An overarching theme of this thesis is the use of *in silico* NAMs and an "open data" approach to generating and sharing important data for ERA to reduce the need for animal testing. The main data source provided is the *AquaTK* R package used in Chapters 2, 3 and 4, which contains *D. magna* TK time-course data with experimental parameters, such as experimental and predicted $\log_{10} K_{ow}$ values, pH, wet weight, and temperature where applicable. Other key data provided in this thesis include the ACDLabs predictions of $\log_{10} K_{ow}$ and $\log_{10} D_{ow}$ in Chapter 3, collated *D. magna* biochemical composition data used in Chapters 3 and 4, the RIE experimental dataset and predicted PaDEL descriptors for the chemicals in Chapter 5, the predicted fup data from ADMET Predictor™ software and the subsets of the US EPA ECOTOX database including 2315 and 2444 steady-state concentration ratios defined as high quality with $\log_{10} D_{ow}$ predictions from CompTox Chemical Dashboard and descriptors from ADMET Predictor™ software where applicable. Each one of these datasets has multiple uses in terms of modelling in ERA and together make a compelling case highlighting alternative approaches available to risk assessors. Importantly, each dataset alongside the relevant code will be made available in separate Github repositories for public access (https://github.com/J-Collins1294/thesis-data & https://github.com/J-Collins1294/thesis-code).

Ultimately, the research contained in this thesis provides a proof-of-concept for the next

generation of ERA. It lays the foundations of a modern ERA that does not require animal testing through the increased use of theoretical and probabilistic methods. Future work should endeavour to integrate these methods into an all-purpose predictive tool in the form of an R package that risk assessors can use for ERA.

Throughout this thesis, data-driven computational methods were explored to relate internal and external concentrations of chemicals in a key aquatic invertebrate, *D. magna*, using relevant biological and physiochemical properties that govern these processes. This work highlighted the application of these methods to ERA and demonstrated their significance in eliminating the need for further animal testing, which is of considerable importance to society and regulatory bodies.

APPENDIX A

A.1 AquaTK chemical overview

Table A.1: 30 unique chemicals from 17 studies digitised and collated from journal publication repositories with their associated CAS registry number for identification and references.

Chemical	**CAS Number**	**Reference**
Atrazine	001912-24-9	[87]
2-4-Dichlorophenoxyacetic Acid	000094-75-7	[87]
Carbamazepine	000298-46-4	[194]
Chlorpyrifos	002921-88-2	[227]
Diazinon	000333-41-5	[139]
Dichlorodiphenyltrichloroethane (DDT)	000050-29-3	[87]
Diclofenac	015307-86-5	[193]
Fluoxetine	054910-89-3	[81]
Imidacloprid	138261-41-3	[155]
Irgarol	028159-98-0	[123]

Table A.1 – continued from previous page

Chemical	**CAS Number**	**Reference**
Kepone	000143-50-0	[230]
Mirex	002385-85-5	[230]
Pentachlorophenol	000087-86-5	[144]
Perfluorodecanoic Acid (PFDA)	000335-76-2	[68][271]
Perfluorododecanoic Acid (PFDoA)	000307-55-1	[68][271]
Perfluorononanoic Acid (PFNA)	000375-95-1	[68][271]
Perfluorooctanesulfonic Acid (PFOS)	001763-23-1	[68][271]
Perfluorooctanoic Acid (PFOA)	000335-67-1	[68][271]
Perfluoroundecanoic Acid (PFUnA)	002058-94-8	[68][271]
Prochloraz	067747-09-5	[70]
Propiconazole	060207-90-1	[70]
Propranolol	000525-66-6	[80]
Roxithromycin	080214-83-1	[80]
Terbutryn	000886-50-0	[123]
Tetracycline	000060-54-8	[131]
Tributyltin chloride	001461-22-9	[92]
Triphenyl Phosphate (TPHP)	000115-86-6	[161]

Table A.1 – continued from previous page

Chemical	**CAS Number**	**Reference**
Tris(1,3-Dichloro-2-Propyl) Phosphate (TDCPP)	013674-87-8	[161]
Tris(2-Butoxyethyl) Phosphate (TBOEP)	000078-51-3	[161]
Tris(2-Chloroethyl) Phosphate (TCEP)	000115-96-8	[161]

A.2 Posterior internal concentration time-course estimates

Samples from the posterior distribution of the model, conditional on the *D. magna* time course data, were used to sample continuous time-courses. The distribution of sampled time courses was summarised in terms of centred 95% and 50% credible intervals. These intervals are plotted against measured internal concentrations in Figure A.1.

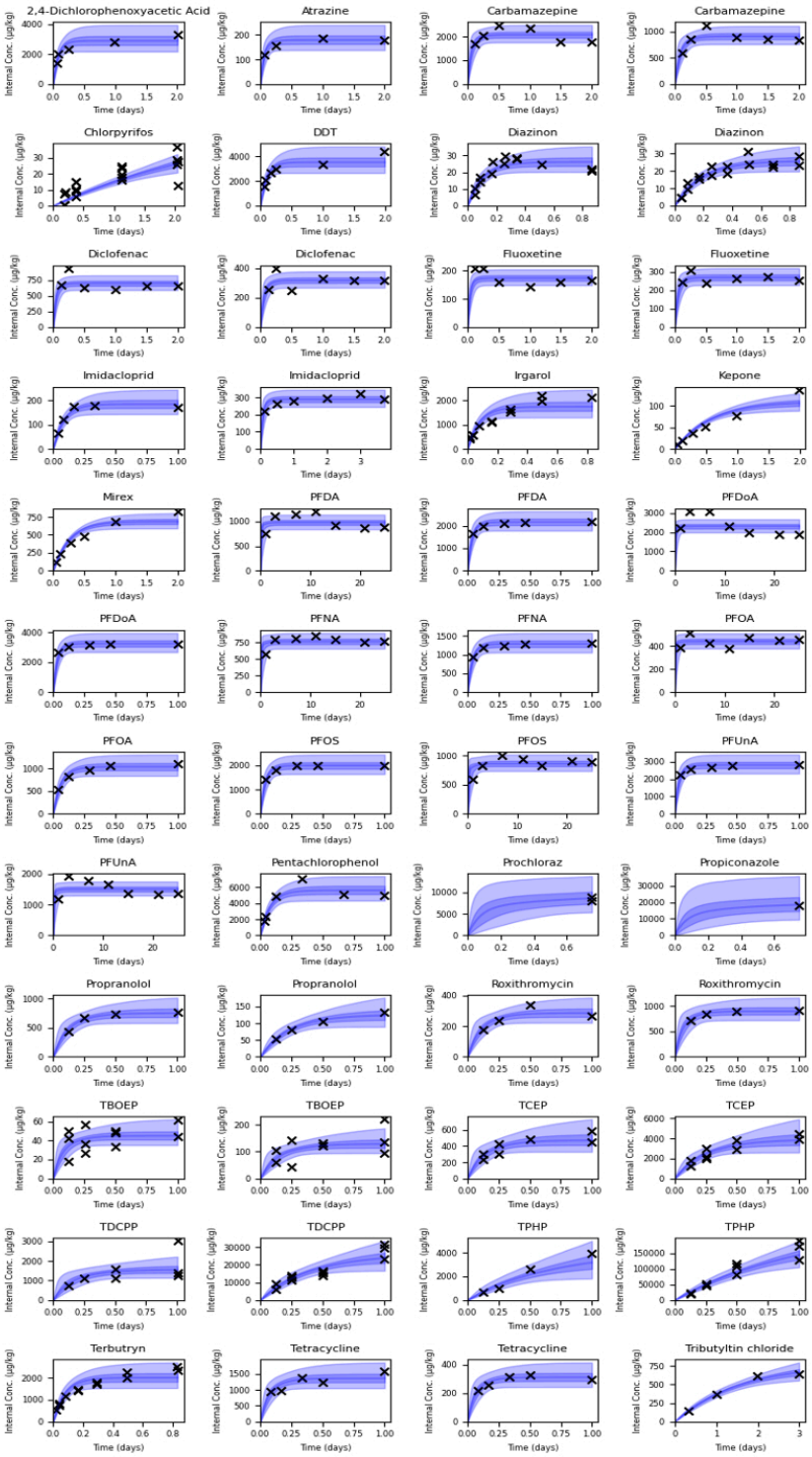

Figure A.1: Posterior predictive plots for the internal concentration time course for the 30 chemicals at each external concentration. Light blue regions cover a 95% centred interval for $c_i(t)$ and dark blue regions a 50% centred interval. Measured concentrations are shown as black crosses.

APPENDIX B

B.1 Time-course data ionisation classification

Table B.1: 47 chemical time-courses, Source ID for each time-course, pH either from the study or a mean value of 7.55 overall, octanol-water partition coefficient (K_{ow}) and dissociation constant (D_{ow}) predictions from ACDLabs, and an ionisable or neutral classification based on whether the $K_{ow} = D_{ow}$ (neutral) or $D_{ow} < K_{ow}$ (ionisable).

Chemical name	**Source ID**	**Ext. Conc.**	**pH**	**ACD** $\log_{10} K_{ow}$	**ACD** $\log_{10} D_{ow}$	**Ionisable/ Neutral**
2,4-Dichlorophenoxyacetic Acid	d12	1000	7.55	2.84	-0.84	Ionisable
Atrazine	d12	100	7.7	2.66	2.66	Neutral
Carbamazepine	d08	5	7	2.28	2.28	Neutral
Carbamazepine	d08	100	7	2.28	2.28	Neutral
Chlorpyrifos	d05	0.08	7.91	4.78	4.78	Neutral
DDT	d12	10	7.55	6.22	6.22	Neutral
Diazinon	d02	1.4604	7.8	3.8	3.8	Neutral
Diazinon	d02	1.5821	7.8	3.8	3.8	Neutral
Diclofenac	d09	5	7	4.48	1.71	Ionisable
Diclofenac	d09	100	7	4.48	1.71	Ionisable
Fluoxetine	d10	0.5	7	4.23	1.46	Ionisable
Fluoxetine	d10	5	7	4.23	1.46	Ionisable
Imidacloprid	d04	190	8.05	0.38	0.38	Neutral
Imidacloprid	d04	240	8.05	0.38	0.38	Neutral

Table B.1 – continued from previous page

Chemical name	**Source ID**	**Ext. Conc.**	**pH**	**ACD** $\log_{10} K_{ow}$	**ACD** $\log_{10} D_{ow}$	**Ionisable/ Neutral**
Irgarol	d03	200	7.9	3.21	3.21	Neutral
Kepone	d15	0.175	7.3	5.37	5.37	Neutral
Mirex	d15	0.118	7.3	6.91	6.91	Neutral
Pentachlorophenol	d16	20	5.5	5.04	4.16	Ionisable
PFDA	d06	5	7.8	7.39	3.63	Ionisable
PFDA	d07	5	7.8	7.39	3.63	Ionisable
PFDoA	d06	5	7.8	8.33	4.58	Ionisable
PFDoA	d07	5	7.8	8.33	4.58	Ionisable
PFNA	d06	5	7.8	6.6	2.85	Ionisable
PFNA	d07	5	7.8	6.6	2.85	Ionisable
PFOA	d06	5	7.8	5.58	1.83	Ionisable
PFOA	d07	5	7.8	5.58	1.83	Ionisable
PFOS	d06	5	7.8	4.17	0.68	Ionisable
PFOS	d07	5	7.8	4.17	0.68	Ionisable
PFUnA	d06	5	7.8	8	4.25	Ionisable
PFUnA	d07	5	7.8	8	4.25	Ionisable
Prochloraz	d13	640.322	7.9	4.08	4.08	Neutral
Propiconazole	d13	479.108	7.15	3.52	3.52	Neutral
Propranolol	d01	5	7	3.26	0.81	Ionisable
Propranolol	d01	100	7	3.26	0.81	Ionisable
Roxithromycin	d01	5	7	3.55	2.45	Ionisable
Roxithromycin	d01	100	7	3.55	2.45	Ionisable
TBOEP	d14	20	7.55	3.46	3.46	Neutral
TBOEP	d14	100	7.55	3.46	3.46	Neutral
TCEP	d14	20	7.55	1.42	1.42	Neutral

Table B.1 – continued from previous page

Chemical name	**Source ID**	**Ext. Conc.**	**pH**	**ACD** $\log_{10} K_{ow}$	**ACD** $\log_{10} D_{ow}$	**Ionisable/ Neutral**
TCEP	d14	100	7.55	1.42	1.42	Neutral
TDCPP	d14	20	7.55	3.26	3.26	Neutral
TDCPP	d14	100	7.55	3.26	3.26	Neutral
Terbutryn	d03	200	7.9	3.35	3.35	Neutral
Tetracycline	d11	100	6.8	-0.45	-3.25	Ionisable
Tetracycline	d11	1000	6.8	-0.45	-3.25	Ionisable
TPHP	d14	20	7.55	4.12	4.12	Neutral
TPHP	d14	100	7.55	4.12	4.12	Neutral

B.2 Bayesian reference model

Table B.2: Reference model used to estimate steady-state concentration ratios.

Parameter type	**Parameter symbol and definition**
Random variable	$c_i(t)$ - Concentration in time-course i at time t
Random variable	K_i - Steady-state concentration rate for time-course i
Random variable	K_{loc} - Average $\log_{10}$ steady-state concentration ratio
Random variable	K_{scale} - Standard deviation of $\log_{10}$ steady-state concentration ratios
Random variable	$k_{\text{out},i}$ - Clearance rate for time-course i
Random variable	t_{shape} - Shape parameter of distribution describing time to 95% steady state

Table B.2 – continued from previous page

Parameter type	**Parameter symbol and definition**
Random variable	t_{scale} - Scale parameter of distribution describing time to 95% steady state
Random variable	σ - Scale parameter of distribution of data with respect to estimated internal concentration
Data	$C_{\text{water},i}$ - External concentration for time-course i
Data	$y_{i,j}$ - jth measured internal concentration for time-course i
Data	$n_{i,j}$ - Number of replicates underpinning the average $y_{i,j}$
Prior	**Likelihood**
$\sigma \sim \text{HalfNormal}(0, 1)$	$t_{95,i} \sim \text{InverseGamma}\left(t_{\text{shape}}, t_{\text{scale}}\right)$
$K_{\text{loc}} \sim \text{Normal}(2, 1)$	$k_{\text{out},i} = \frac{3}{t_{95,i}}$
$K_{\text{scale}} \sim \text{HalfNormal}(0, 1)$	$\log_{10} K_i \sim \text{Normal}\left(K_{\text{loc}}, K_{\text{scale}}\right)$
$t_{\text{shape}} \sim \text{HalfNormal}(0, 1)$	$c_i(t) = C_{\text{water},i} K_i \left(1 - \text{e}^{-k_{\text{out},i} t}\right)$
$t_{\text{scale}} \sim \text{HalfNormal}(0, 10)$	$\log y_{i,j} \sim \text{Normal}\left(\log\left(c_i\left(t_{i,j}\right)\right), \frac{\sigma}{\sqrt{n_{i,j}}}\right)$

B.3 Bayesian predictive model

Table B.3: Specification of the Bayesian predictive model that uses chemical-specific partition coefficients to estimate steady-state concentration ratios.

Parameter type	**Parameter symbol and definition**
Random variable	$c_i(t)$ – Concentration in time-course i at time t
Random variable	K_i – Steady-state concentration rate for time-course i
Random variable	α – intercept for predicting steady-state concentration ratio using partition coefficient
Random variable	β – slope for predicting steady-state concentration ratio using partition coefficient
Random variable	K_i^{pred} – predicted steady-state concentration rate for time-course i
Random variable	K_{scale} – standard deviation of $\log_{10}$ steady-state concentration ratio prediction error
Random variable	$k_{\text{out},i}$ – Clearance rate for time-course i
Random variable	t_{shape} – Shape parameter of distribution describing time to 95% steady state
Random variable	t_{scale} – Scale parameter of distribution describing time to 95% steady state
Random variable	σ – Scale parameter of distribution of data with respect to estimated internal concentration
Data	$C_{\text{water},i}$ – External concentration for time-course i

Parameter type	**Parameter symbol and definition**
Data	$y_{i,j}$ – jth measured internal concentration for time-course i
Data	$n_{i,j}$ – number of replicates underpinning the average $y_{i,j}$
Data	$\log_{10}\mathrm{P}_i$ – base 10 logarithm of the partitioning coefficient for the chemical corresponding to time-course i
Prior	**Likelihood**
$\sigma \sim \text{HalfNormal}(0, 1)$	$t_{95,i} \sim \text{InverseGamma}\left(t_{\text{shape}}, t_{\text{scale}}\right)$
$\alpha \sim \text{Normal}(2, 1)$	$k_{\text{out},i} = \frac{3}{t_{95,i}}$
$\beta \sim \text{Normal}(0, 1)$	$\log_{10} K_i^{\text{pred}} = \alpha + \beta \log_{10}\mathrm{P}_i$
$K_{\text{scale}} \sim \text{HalfNormal}(0, 1)$	$\log_{10} K_i \sim \text{Normal}\left(\log_{10} K_i^{\text{pred}}, K_{\text{scale}}\right)$
$t_{\text{shape}} \sim \text{HalfNormal}(0, 1)$	$c_i(t) = C_{\text{water},i} K_i \left(1 - \mathrm{e}^{-k_{\text{out},i} t}\right)$
$t_{\text{scale}} \sim \text{HalfNormal}(0, 10)$	$\log\, y_{i,j} \sim \text{Normal}\left(\log\left(c_i\left(t_{i,j}\right)\right), \frac{\sigma}{\sqrt{n_{i,j}}}\right)$

APPENDIX C

C.1 Partitioning model

C.1.1 Assumptions

Steady-state is assumed for all quantities unless explicitly stated, i.e., the internal concentration is assumed to be the *steady-state* internal concentration.

C.1.2 Two-phase partitioning model

Let m be the *total chemical mass*, and let M be the *total organism mass*, then y is the *total internal concentration* defined as the ratio,

$$y := \frac{m}{M},\ M > 0$$

Let $y_\partial = \frac{m_\partial}{(M_\partial + m_\partial)}$ be the *partition internal concentration* in a two-phase partition such that,

$$M = M_\ell + M_\partial + m,$$

and

$$m = m_\ell + m_\partial.$$

In environmental exposure scenarios, the mass of the chemical in each partition is assumed to be much smaller than the mass of the partition it "occupies" and therefore the mass of the compartment, which implies that $y_\partial \approx \frac{m_\partial}{M_\partial}$ and $M \approx M_\ell + M_\partial$.

Show that,

$$\frac{y}{y_\partial} = \nu \cdot K + 1 - \nu,\ y_\partial > 0$$

where ν is defined as the *ℓ-partition mass fraction*,

$$\nu := \frac{M_\ell}{M},$$

and K is the *true organism-specific partitioning coefficient* defined as the ratio of the concentrations of each partition,

$$K := \frac{y_\ell}{y_\partial} = \frac{m_\ell/M_\ell}{m_\partial/M_\partial} = \frac{m_\ell}{m_\partial} \cdot \frac{M_\partial}{M_\ell}.$$

From the definitions of y, y_∂, m, and K,

$$\begin{aligned} z := \frac{y}{y_\partial} &= \nu \frac{m/M_\ell}{m_\partial/M_\partial} \\ &= \nu \frac{m}{m_\partial} \frac{M_\partial}{M_\ell} \\ &= \nu \left(\frac{m_\ell + m_\partial}{m_\partial} \right) \frac{M_\partial}{M_\ell} \\ &= \nu \left(\frac{m_\ell}{m_\partial} + 1 \right) \frac{M_\partial}{M_\ell} \\ &= \nu \frac{m_\ell}{m_\partial} \frac{M_\partial}{M_\ell} + \nu \frac{M_\partial}{M_\ell} \\ &= \nu K + 1 - \nu \end{aligned}$$

where it possible to use the fact that,

$$y = \frac{m}{M} = \nu \frac{m}{M_\ell},$$

in the first step, and the last step uses the definition of K and the fact that $M_\partial = (1-\nu)M$ and $M_\ell = \nu M$ giving,

$$\nu \frac{M_\partial}{M_\ell} = \nu \frac{1-\nu}{\nu} = 1 - \nu.$$

C.1.3 General partitioning model

The total internal concentration can be expressed as,

$$y = \frac{m_\partial}{M_\partial} \left(\nu \frac{m_\ell}{m_\partial} \frac{M_\partial}{M_\ell} + \frac{M_\partial}{M} \right) = \nu \frac{m_\ell}{M_\ell} + \frac{m_\partial}{M},$$

composed of two terms that contain only masses from either partition where M is assumed constant. Let $m_\ell = m_1 + m_2$ and $M_\ell = M_1 + M_2$, which implies that,

$$\nu = \frac{M_\ell}{M} = \frac{M_1}{M} + \frac{M_2}{M} = \nu_1 + \nu_2$$

Substituting into the left-hand term in the expression for total internal concentration,

$$
\begin{aligned}
\nu\frac{m_\ell}{M_\ell} &= \nu_1\frac{m_\ell}{M_\ell} + \nu_2\frac{m_\ell}{M_\ell} \\
&= \nu_1\frac{m_1+m_2}{M_1+M_2} + \nu_2\frac{m_1+m_2}{M_1+M_2} \\
&= \frac{M_1}{M}\frac{m_1+m_2}{M_1+M_2} + \frac{M_2}{M}\frac{m_1+m_2}{M_1+M_2} \\
&= \frac{m_1}{M} + \frac{m_2}{M} \\
&= \nu_1\frac{m_1}{M_1} + \nu_2\frac{m_2}{M_2}
\end{aligned}
$$

Thus, for any sub-partition of a partition, results in a linear combination of each of the sub-partitions. In general, then, for i sub-partitions

$$
m = \sum_i m_i + m_\partial
$$

and,

$$
M = \sum_i M_i + M_\partial + m,
$$

the concentration ratio can then be expressed as,

$$
\frac{y}{y_\partial} = \sum_i \phi_i K_i + 1 - \sum_i \phi_i,
$$

where ϕ_i is the mass fraction of the ith sub-partition, K_i is the partitioning coefficient of the ith sub-compartment, and the last term results from the fact that $M_\partial = \left(1 - \sum_i \phi_i\right) M$ must be over all partition's mass fractions.

C.2 Protein surface-binding model

C.2.1 Limitations of current models

TK studies for propranolol show that for the same study, for the same chemical, over the same time exposure, but at different external concentrations, significant differences in steady-state internal-external ratios are observed. This motivates the development of a model that captures an external concentration dependence, i.e.,

$$z = y/y_w = f\left(K, y_w; \nu, \ \ldots\right), \ y_w > 0.$$

Studies of protein binding in *Gammarus pulex* at different concentrations provide empirical evidence that suggest the dependence on external concentration is linked to protein binding. Protein binding is typically modelled as a surface-binding process. This motivates the development of a protein surface-binding model.

C.2.2 Assumptions

Under the assumption that chemical mass contributions can be linearly decomposed into lipid-water, and protein contributions,

$$y = \frac{m}{M} = \frac{m_\ell + m_w + m_p}{M} = \frac{m_\ell + m_w}{M} + \frac{m_p}{M} = y^{(I)} + y^{(II)}.$$

The key point here being that each of the two processes can be modelled *independently* and combine their contributions. Specifically, the previous section shows that the first term can be expressed as a function of the partitioning coefficient and the mass fraction,

$$\frac{y}{y_w} = \frac{y^{(I)}}{y_w} + \frac{y^{(II)}}{y_w} = \nu K + 1 - \nu + \frac{1}{y_w}\frac{m_p}{M}.$$

The aim of the remaining text is to obtain an estimate of the protein chemical mass term

m_p since both y_w and M are both known.

C.2.3 Densities

The term m_p/M can be interpreted as both a mass concentration and a density, since Daphnia are 93% water, and the density of water is close to 1, the *protein-chemical density* is defined as,

$$\rho_p := \frac{m_p}{V} \approx \frac{m_p}{M} = y^{(II)}.$$

A similar argument applies to the water concentration, defined here with respect to the *water-chemical density,*

$$\rho_w := \frac{m_w}{V} \approx \frac{m_w}{M} \approx \frac{m_w}{M_w} = y_w.$$

It is expected that the surface density of chemical on the protein to be closely related to the volume density of chemical in the water. The aim is to develop a model relating the two quantities.

Remark 1. For unit length-scales length, surface, and volume densities interchangeably can be used interchangeably since, $\widehat{\rho} = \frac{m}{L^3} = \frac{m}{L^2} = \frac{m}{L}$.

Most environmental measurements of external water concentrations have unit length-scale, i.e., they are measured in X per unit litre. For $L \neq 1$, it is important to make a distinction between the volume density $\rho^{(v)}$ and the surface density $\rho^{(s)} = m/L^2$.

Remark 2. It is implicitly assumed in the previous statement that the *internal water-chemical density* in the organism that "encloses" the protein is the same as the external

water-chemical density.

In the next section the relationship between the number of available molecules in an enclosing space, an embedded $D-1$ dimensional subspace, and the maximum possible surface density is investigated.

C.2.4 Toy surface-binding model

Let V be a unit volume representing the water partition of the organism that is assumed to enclose a unit surface S representing the surface of the protein partition. Let b be the number of binding sites in a unit length-scale so that $B_S = b^2$ and $B_V = b^3$. For example, set $b = 50$ so that $B_S = 50^2 = 2,500$ and $B_V = 50^3 = 125,000$.

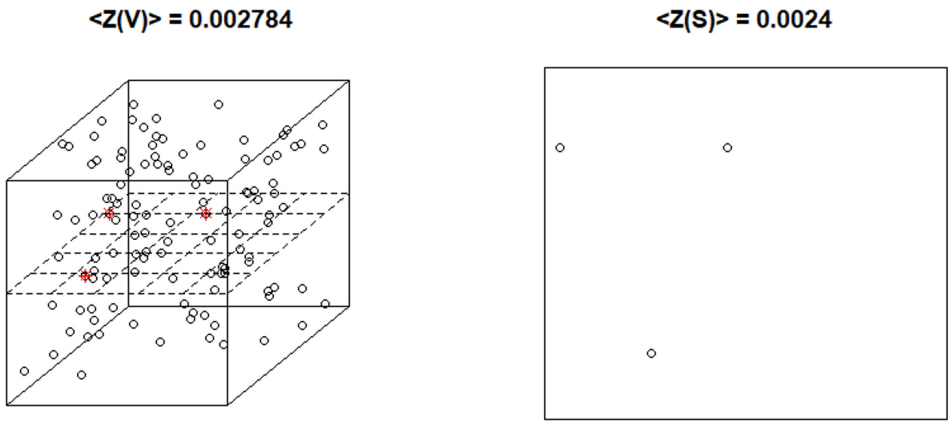

Figure C.1: Unit volume representing the water partition of the organism enclosing a protein unit surface. Both have approximately the same average density $\langle Z \rangle \approx 1/1000$.

Let $Z \sim Bern(p_\rho)$ be a random variable representing a molecule with unit mass that is randomly present with probability p_ρ at any given binding site i. Set $p_\rho = \rho$, so that the *expected density* of the volume is $E(Z) = \rho$. The randomness here corresponds to the random movement of the chemical molecules in the water partition of the organism.

Concretely, suppose the external water concentration is $y_w = 1,000$ μg/L. It is as-

sumed that the water partition in the organism has the same chemical-water density, so $\rho_w = y_w = 1,000\ \mu\text{g/L} = 10^{-6}\ \text{kg/L} \approx 10^{-6}\ \text{kg / kg}$.

Remark. It's important the density is dimensionless (kg/kg) to represent it as a probability – so the chemical mass must always be normalised with respect to the enclosing volume's mass. This avoids probabilities greater than 1.

Simulation steps. Sample from a $Bern(p_\rho)$ with $p_\rho = \rho_w$ at each binding site. Figure C.1 provides an example simulation for $\rho_w = 10^{-3}$. The average density across the volume (in fact any space) will be $\langle Z \rangle \approx \rho_w$. Any surface *embedded* in the volume will have the same density. This is simply because the density is defined at each binding site, independently of the geometry. An example of a volume simulation and surface embedded in the volume is presented in Figure C.1.

In other words, it is plausible to imagine each binding site independently of its geometry and talk about surface and volume densities interchangeably provided the space is suitably normalised. This provides the link between the external water concentration and the number of available chemical molecules at the surface of the protein.

C.2.5 Binding & unbinding probabilities

Given an enclosing volume density $\rho_w^{(V)}$ and an embedded surface with density $\rho_w^{(S)} = \rho_w^{(V)}$, the probability the molecule binds to the surface is defined as,

$$P(B|\,\{Z = 1\}) := p_b,$$

where B is the event the molecule binds. The probability the molecule binds *and then*

unbinds is defined as,

$$P(U|\{Z=1\}\cap B) := p_u,$$

where U is the event the molecule unbinds after it has first bound. The *protein affinity* is defined as,

$$P := \frac{p_b}{p_u}.$$

Let Y_0 be a random variable representing an available molecule with unit mass that has bound and not unbound to the surface of the protein, i.e., $Y_0 \sim Bern(p)$. The value of p can be calculated using Bayes law and the probability of intersection,

$$p = P\left(\{Z=1\}\cap B\cap U^c\right) = P(Z=1)P\left(B|Z=1\right)P\left(U^c|\{Z=1\}\cap B\right) = p_\rho p_b\left(1-p_u\right).$$

The expected density follows $E\left(Y_0\right) = p_\rho p_b\left(1-p_u\right)$ from the Bernoulli distribution. Figure C.2 is a simulation of a binding and unbinding process.

Remark. This is under the assumption that the time it takes a molecule to bind and unbind is essentially instantaneous. The timescales that protein binding and unbinding take place at are many orders of magnitude smaller than Fickian diffusion. Protein binding typically reaches steady-state after a few seconds compared with lipid-binding diffusion that reaches steady-state over potentially several days.

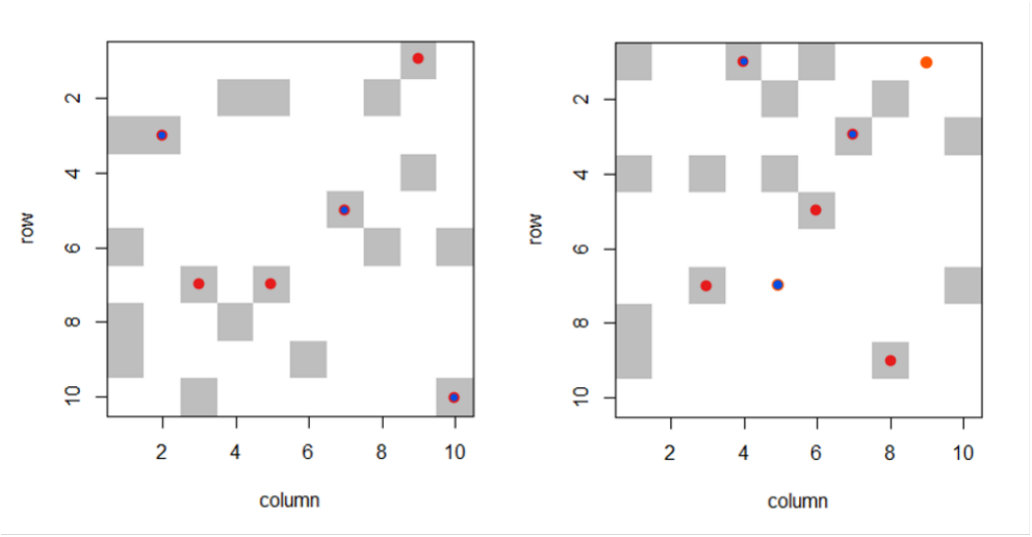

Figure C.2: Toy surface-binding model: $B_s\rho$ grey binding sites, $B_s p_\rho p_p$ red bound molecules, $B_s p_\rho p_b p_u$ blue unbound molecules, $B_s p_\rho p_b(1-p_u)$ orange accumulated molecules; left at initial time $t=0$, right: next step in time $t=1$.

C.2.6 Repeated binding & occupancy

Given an available binding site, the molecule at that site has a probability p_b of binding to that site. Moreover, once the molecule has bound, there is a probability p_u it unbinds. Figure 3 illustrates the toy model for an initial and following timestep.

Once a molecule has bound, that site is *occupied,* and a molecule cannot bind again to that site until the occupying molecule unbinds. Occupancy requires a notion of available space. Let B be the total number of available binding sites (this could be a surface or a volume) and let $\Delta B_n = B - BE(Y_n)$ be the expected number of remaining sites at the nth repeated binding process, since $BE\left(Y_n\right) = Bp$.

The expected number of molecules bound in the next experiment will depend on the number of sites that have been occupied in the previous experiment,

$$BE(Y_{n+1}) = \left(BE(Y_n) + \Delta B_n E\left(I_b\right)\right) E(1-I_u),$$

where I_b and I_u are indicator random variables representing the event of the molecules binding and unbinding with probabilities p_b and p_u as defined in the previous section.

In other words, the total expected number of occupied sites at time n will be the expected number of occupied sites from the previous time $BE(Y_n)$ plus the proportion of remaining unoccupied sites that bind $\Delta B_n E\left(I_b\right)$. The total is then multiplied by the proportion that do not unbind $E(1-I_u)$.

Substituting back $\Delta B_n = B - BE(Y_n)$ it is possible to factor out B,

$$E\left(Y_{n+1}\right) = \left(E\left(Y_n\right) + 1 - E\left(Y_n\right) E\left(I_b\right)\right) E\left(1 - I_u\right).$$

Expanding out the terms expressed as probabilities and taking $n \to \infty$, the steady-state surface density is given by,

$$\rho_{ss} \approx \frac{p_b \left(p_u - 1\right)^2 p_\rho}{p_b \left(p_u - 1\right)^2 p_\rho + p_u} = \frac{p_\rho q}{p_\rho q + p_u} = \frac{\rho q}{\rho q + p_u}.$$

Setting $\rho_{ss} = \rho$ and rearranging,

$$\rho = \frac{\rho q}{\rho q + p_u}$$

$$\rho q + p_u = q$$

$$\rho = \frac{q - p_u}{q} = \frac{p_b \left(p_u - 1\right)^2 + p_u}{p_b \left(p_u - 1\right)^2} \approx \frac{p_b - p_u}{p_b}$$

In other words, p_u needs to be at least a factor $\left(\frac{1}{\rho} - 1\right)$ larger than p_b for ρ_{ss} to be at

least as large as ρ. In general, $\rho \ll 1$, requiring the chemical protein affinity to be several orders of magnitude greater for the steady-state surface density to exceed the availability density ρ. Thus, the enclosing density, the number of molecules available to the surface, e.g., the water-chemical density enclosing the protein, can be considered an upper bound on the surface density.

Simulation. Create an array of B binding sites. Generate occupancy by sampling from a Bernoulli distribution with $Bern(p_\rho)$ at each binding site. For each binding sites sample another Bernoulli random variable with $Bern(p_b)$ and multiply the previous value. Then multiply again for $Bern(1 - p_u)$. In the next step, generate occupancy sampling for the sites that have 0. This tracks the occupancy from the previous step. Apply the binding step to the subset but apply the unbinding step to all binding sites. Repeat n times.

Figure C.3 simulates a repeated binding process and tracks the steady-state density. As the number of sites become occupied, the probability any single remaining site gets occupied becomes smaller and smaller, and the number of molecules that unbinds becomes larger and larger. An equilibrium is reached when $\frac{p_b(p_u-1)^2 p_\rho}{p_b(p_u-1)^2 p_\rho + p_u}$ sites are occupied. For the parameters in the simulation, $\rho_{ss} \approx 0.1611..$ which agrees with the simulation.

C.2.7 Estimating surface densities

In the previous section chemicals with a protein affinity $P < \frac{1}{\rho} - 1$ were shown that the enclosing density ρ is an upper bound on the surface density. The next step is to estimate ρ from the (internal) water concentration.

As a simple example, assume all the mass in the water in the organism is available to bind to the protein surface at any given instance of time, then,

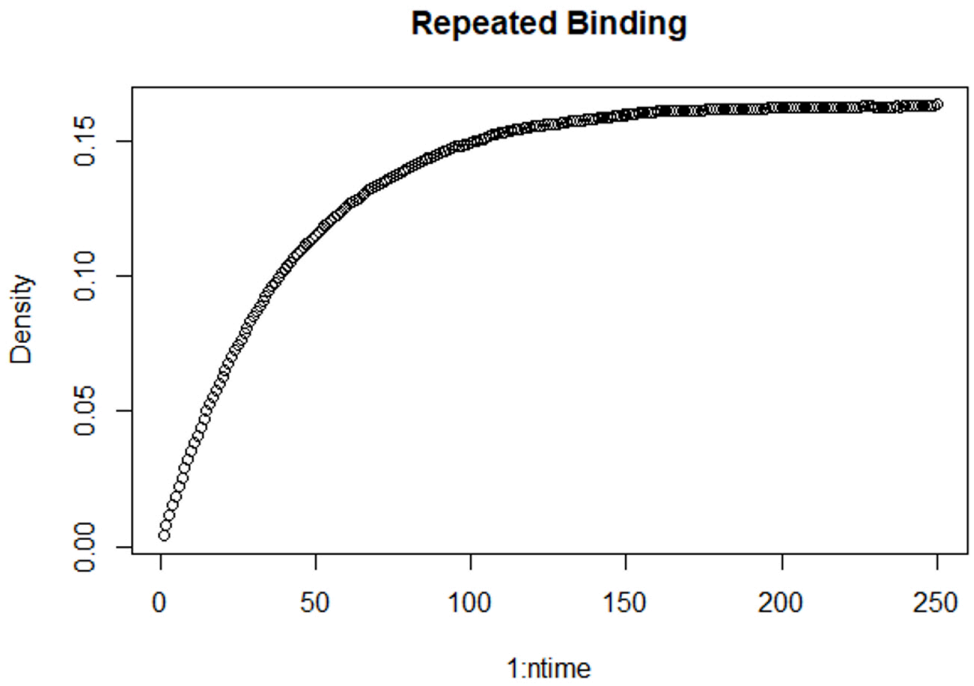

Figure C.3: Density of molecules on a surface of $B = 10^6$ binding sites with $\rho = 1/50$, $p_b = 1/5$, and $p_u = b_b/10$.

$$\rho_p^{(S)} \leq \rho_w^{(V)} = \frac{m_w}{V} \approx \frac{m_w}{M_w} = y_w.$$

Remark. It doesn't matter if all the mass binds to the protein and is replaced in the water over a long period of time. From the results in the previous section, the density on the protein surface cannot exceed the density of the enclosing volume provided $P < \frac{1}{\rho} - 1$. Note, surface and volume densities can discussed interchangeably at this point because the densities (so far) are defined with respect to unit length-scales as is usually the case, i.e., V has unit 1 (litre).

Suppose now that all the mass in the internal water occupies the protein volume, which it indeed must do before it binds. The density in the protein volume will be much higher than in the water volume because the protein volume is much smaller than the water volume. Figure C.4 illustrates the increase in density for a protein fraction $\frac{M_p}{M} \approx \frac{V_p}{V} = 1/2$.

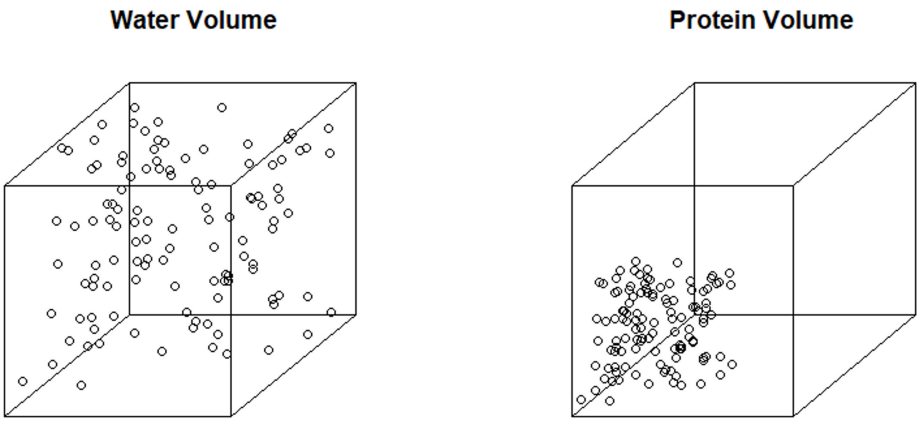

Figure C.4: Chemical mass in a water volume compacted into a protein volume.

In fact,

$$V_p = \varphi V \approx \varphi V_w,$$

where V_p is the *protein partition volume* and φ is the *protein mass fraction* where it is assumed that $\varphi = M_p/M \approx V_p/V$.

Thus, the density in the protein volume should instead be bound by,

$$\rho_p^{(S)} \leq \frac{m_w}{V_p} = \frac{m_w}{\varphi V} = \frac{1}{\varphi}\rho_w \approx \frac{1}{\varphi} y_w.$$

However, rescaling the volume by a factor of φ means the volume density is no longer defined over a unit length. The difference between a surface and a volume density must therefore be considered through a transform that preserves the density between a surface and a volume.

As a concrete example, consider estimating a scalar volume density $\rho = m/V = 10/100$ and keeping the mass fixed and converting the volume density into a surface density by transformation of the volume such that,

$$\rho^{(S)} = \left(\rho^{(V)}\right)^{\frac{2}{3}} = \left(\frac{m}{V}\right)^{\frac{2}{3}} = \left(\frac{m}{L^3}\right)^{\frac{2}{3}} = \left(\frac{m^2}{L^2}\right) = \left(\frac{10}{100}\right)^{\frac{2}{3}} = 0.21..$$

The resulting density is much higher than the original density creating an overestimation of the upper bound. For extremely small volume densities, which is the case in environmental settings, the overestimation could be several orders of magnitude.

This can be corrected in the upper bound by taking the square-root of the mass in the estimated water-chemical volume density,

$$\rho_p^{(S)} \leq \left(\frac{1}{\varphi}\right)^{\frac{2}{3}} \sqrt{y_w}.$$

The right-hand side can now be interpreted as an upper bound on the *protein-chemical surface density* $\rho_p^{(S)}$ with the scaling of φ adjusted accordingly.

C.3 Combined lipid and protein surface-binding model

The combined model follows from the inequality,

$$\frac{y}{y_w} = \frac{y^{(I)}}{y_w} + \frac{y^{(II)}}{y_w} = \nu K + 1 - \nu + \frac{1}{y_w}\frac{m_p}{M} \leq \nu K + 1 - \nu + \left(\frac{1}{\varphi}\right)^{\frac{2}{3}} \frac{1}{\sqrt{y_w}}$$

C.3.1 Tighter upper bound

It is observed that,

$$\rho \leq \rho_{ss}$$

when,

$$P < \left(\frac{1}{\rho} - 1\right).$$

Rearranging the expression above,

$$1 + P < \frac{1}{\rho},$$

and since $1 + P > \frac{1}{\rho} > 0$,

$$\frac{1}{1+P} > \rho$$

Since ρ as y_w can be interpreted,

$$y_w \equiv \rho < \frac{1}{1+P}$$

Substituting into the model,

$$\rho_p^{(S)} < \left(\frac{1}{\varphi}\right)^{\frac{2}{3}} \sqrt{\frac{1}{1+P}} \leq \left(\frac{1}{\varphi}\right)^{\frac{2}{3}} \sqrt{y_w}, \quad y_w > \frac{1}{1+P}$$

Given both P and y_w are known, a tighter upper bound can be achieved by knowing P, provided $\frac{1}{1+P}$ is less than y_w. Moreover, P can often be predicted from chemical properties, suggesting a reduced parameter version of the model might be possible. For example, if $P \approx f(K_{ow})$ the model can be reduced to a single variable model.

C.3.2 Fup% protein affinity estimates

A key condition of the model is the requirement that $P < ((1/\rho) - 1)$. P can be estimated with the fraction unbound in plasma (fup%) predicted from ADMET predictor software. Since P is the ratio of probability of binding to unbinding,

$$\frac{100 - fup\%}{fup\%} \approx \frac{p_b}{p_u} = P$$

APPENDIX D

D.1 Calibration curves and linear regression fits

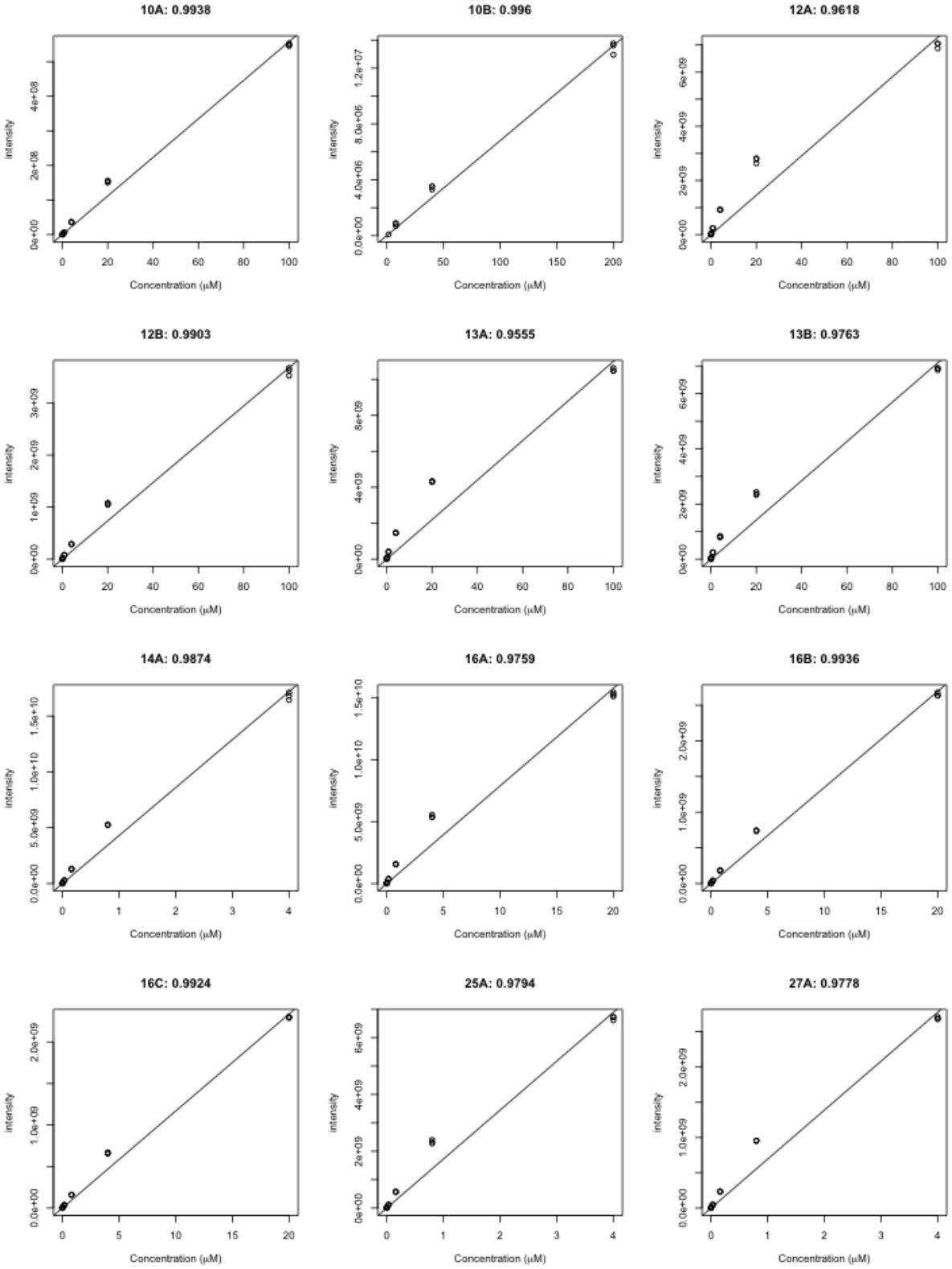

Figure D.1: Calibration curves and linear regression fits with R^2 values for chemicals 10A - 27A.

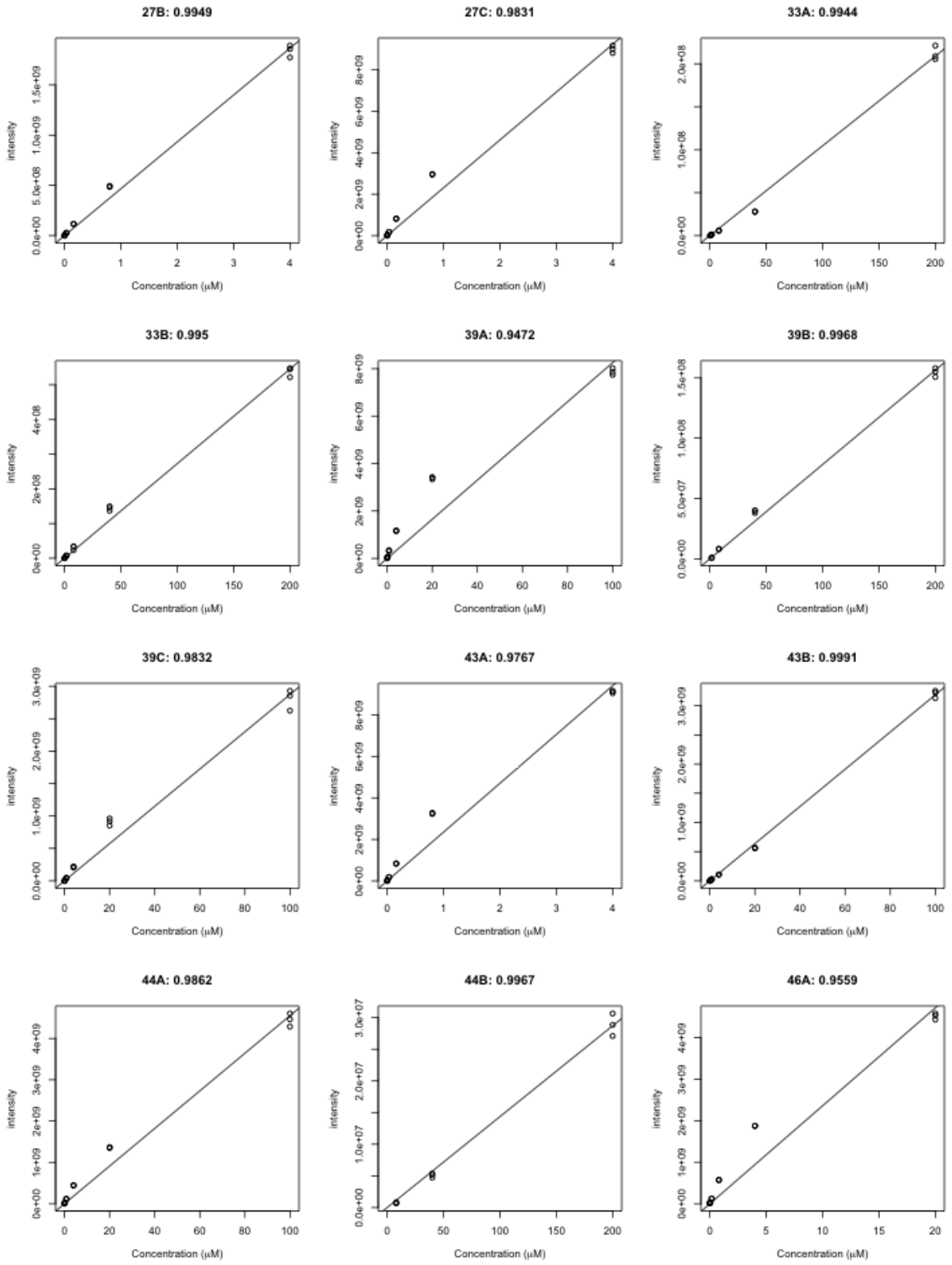

Figure D.2: Calibration curves and linear regression fits with R^2 values for chemicals 27B - 46A.

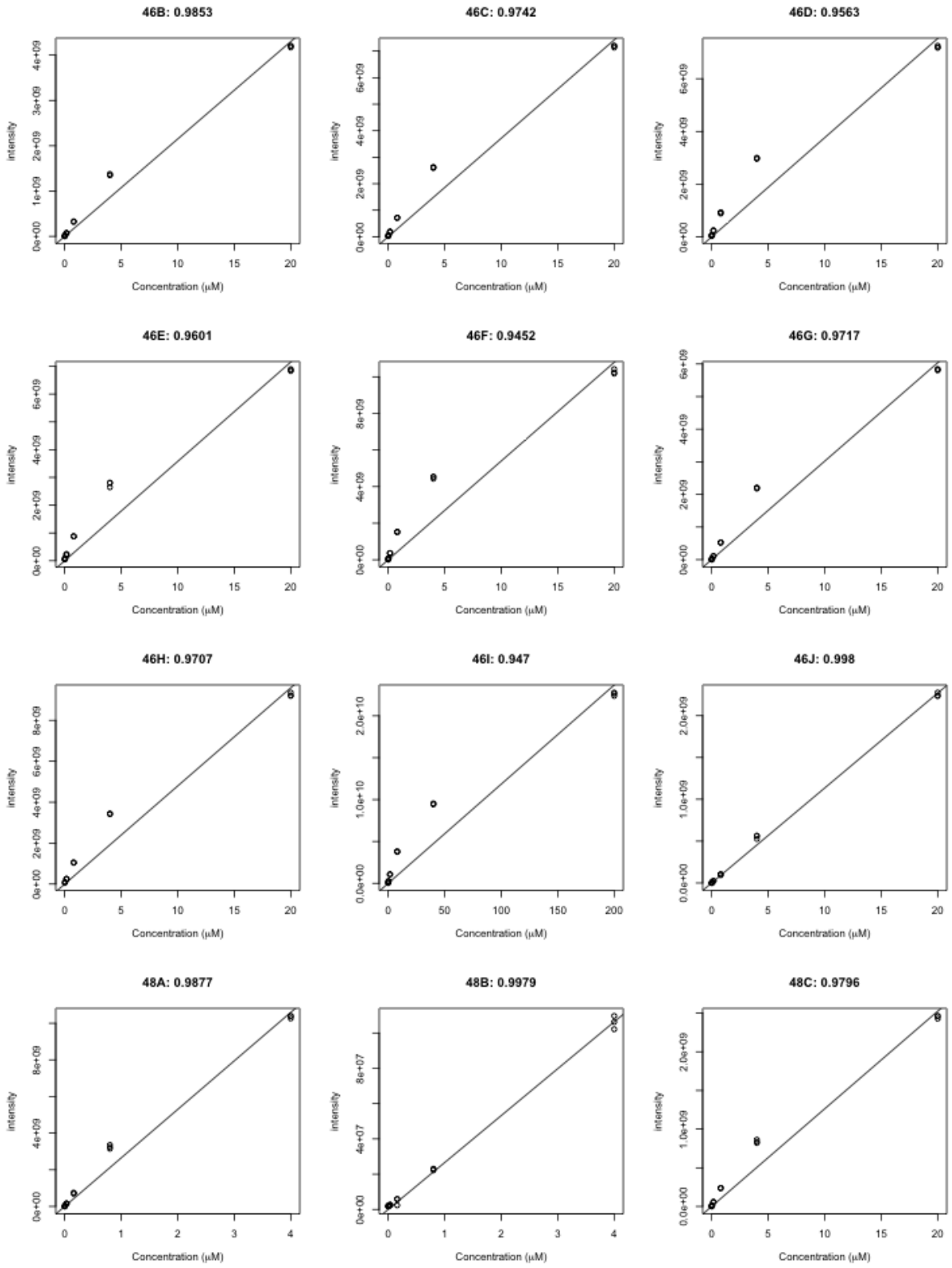

Figure D.3: Calibration curves and linear regression fits with R^2 values for chemicals 46B - 48C.

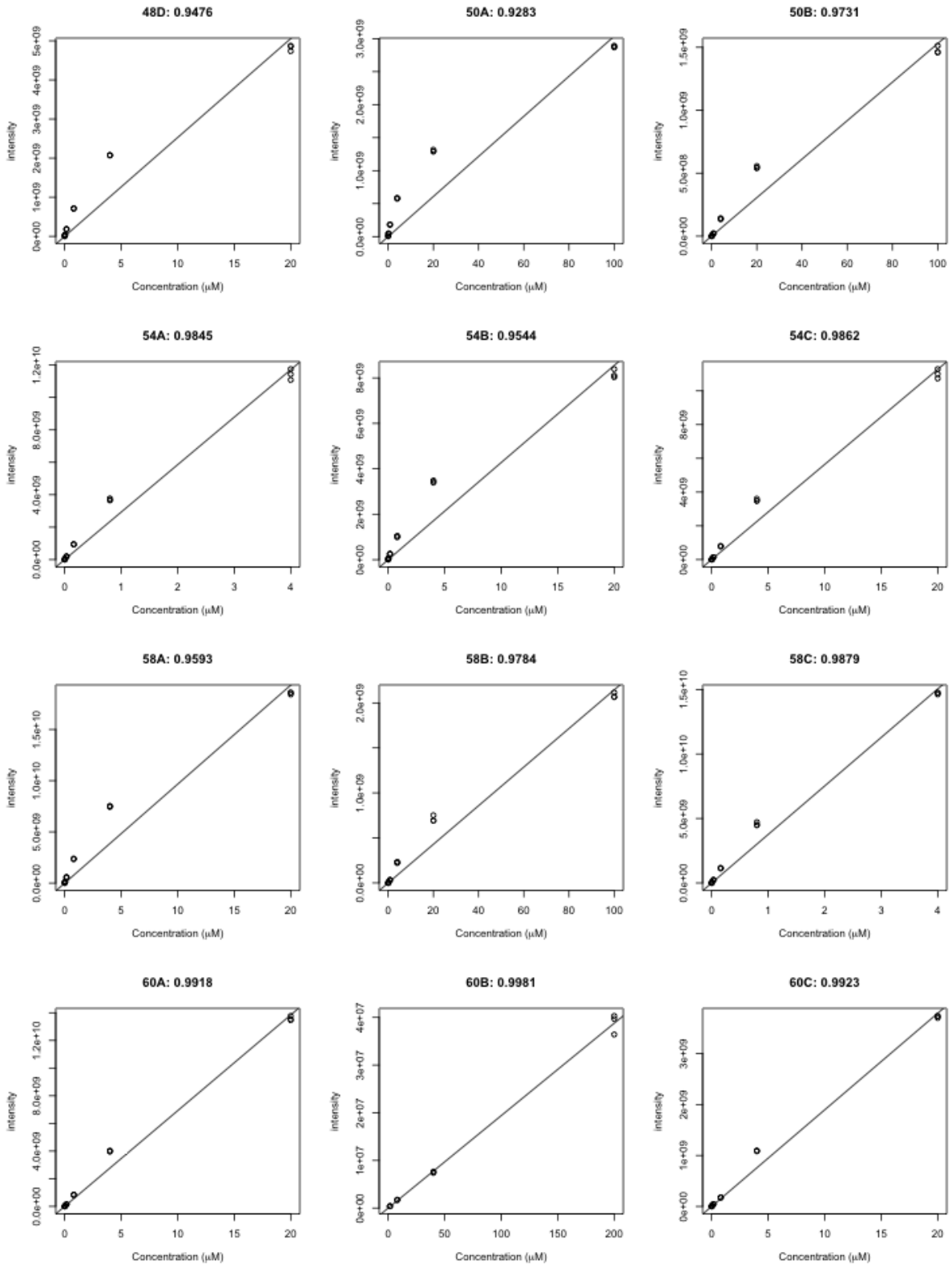

Figure D.4: Calibration curves and linear regression fits with R^2 values for chemicals 48D - 60C.

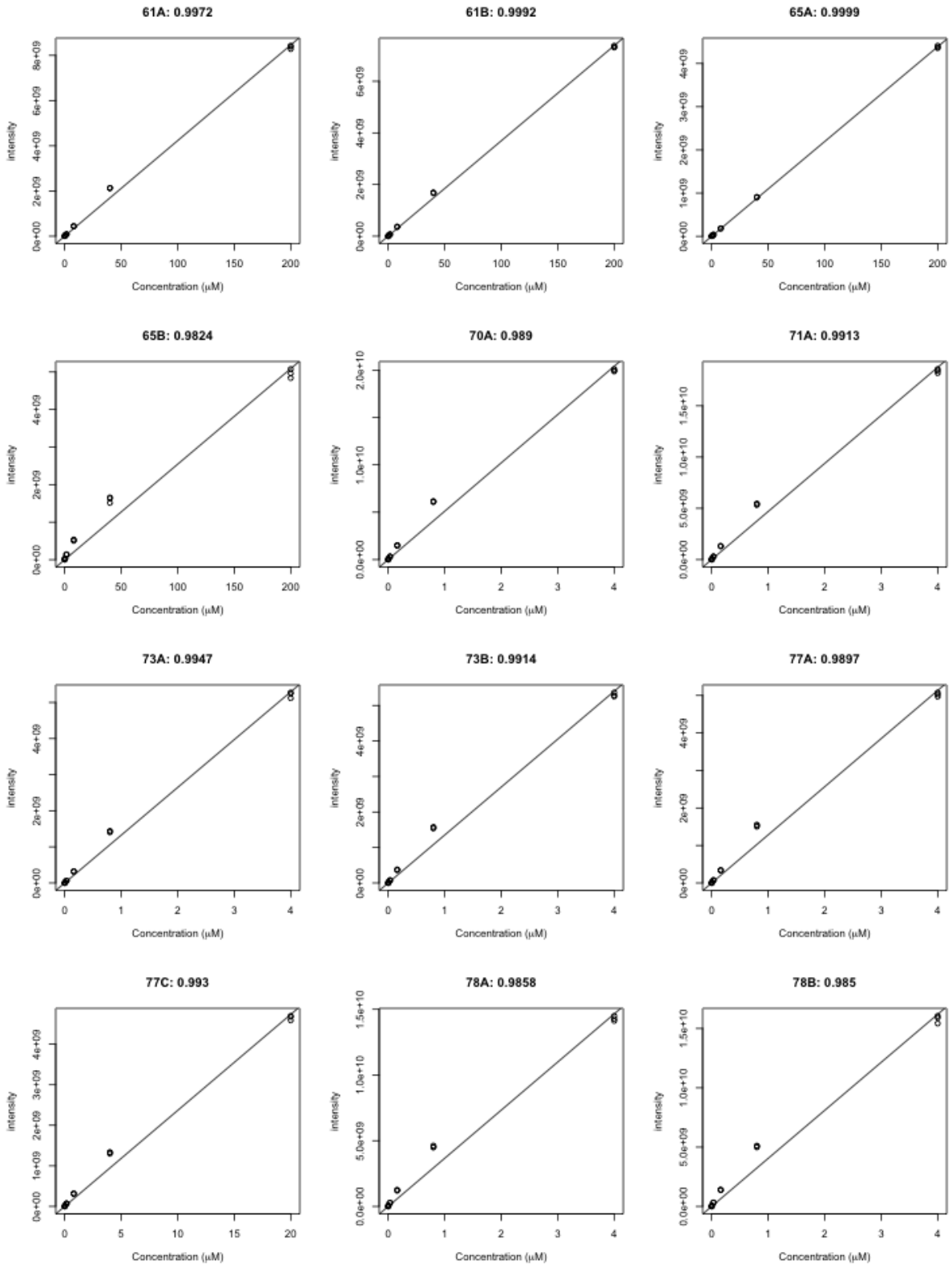

Figure D.5: Calibration curves and linear regression fits with R^2 values for chemicals 61A - 78B.

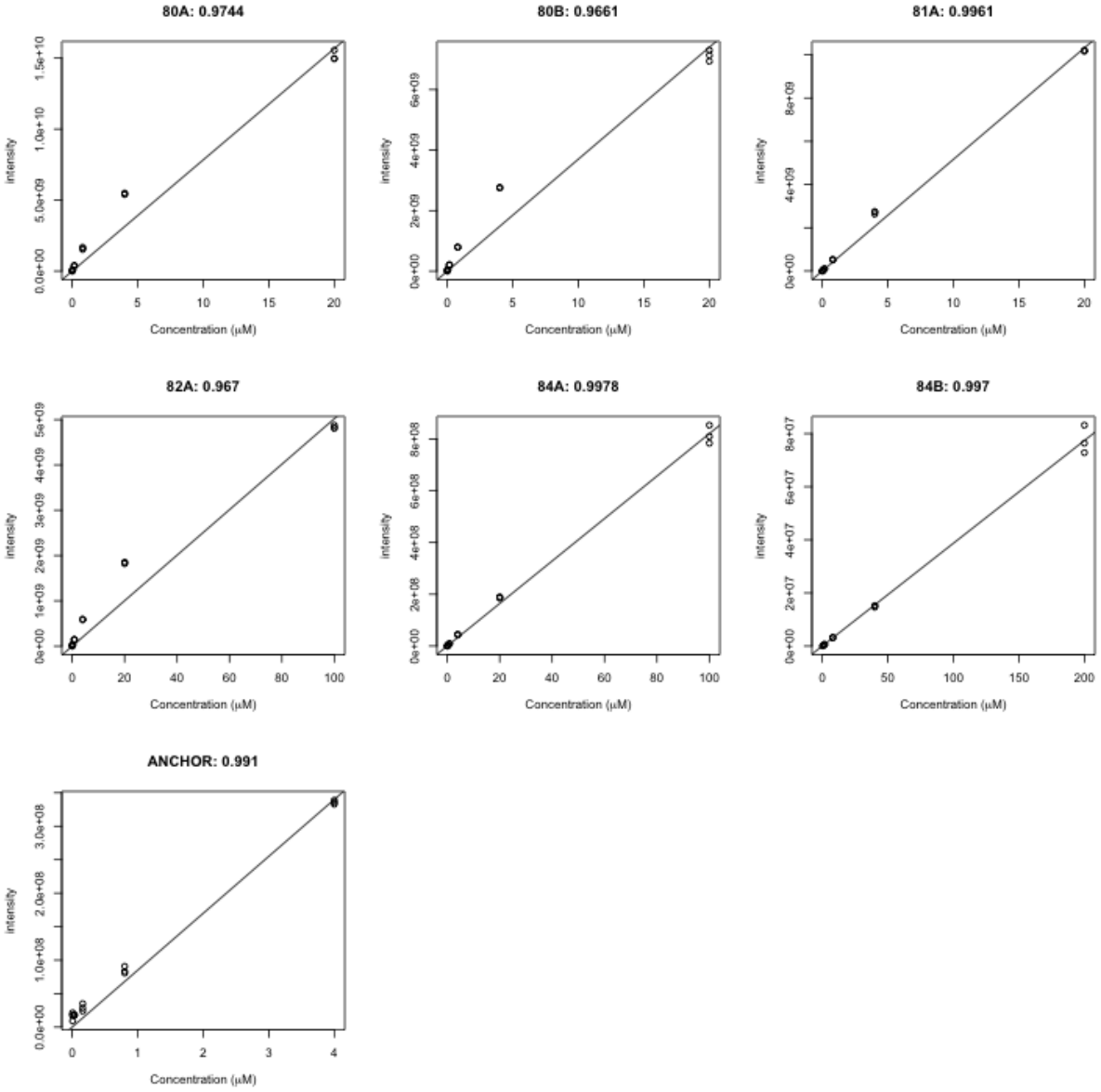

Figure D.6: Calibration curves and linear regression fits with R^2 values for chemicals 80A - Anchor.

D.2 Overview of relative ionisation efficiency data

Table D.1: Overview of chemical ID, chemical name, concentration group, gradient, R^2 value, relative ionisation efficiency (RIE) values, and $\log_{10}$RIE values for the 67 parent and biotransformation products including the anchor chemical.

ID	**Chemical name**	**Concentration group**	**Gradient**	R^2	**RIE**	$\log_{10}$**RIE**
10A	9-anthraldehyde	C	5.58×10^6	0.994	0.066	-1.183
10B	9-Anthracenecarboxylic acid	D	6.80×10^4	0.996	0.001	-3.097
12A	Acetochlor	C	2.33×10^8	0.996	2.743	0.438
12B	2-Chloro-N-(2-ethyl-6-methylphenyl)acetamide	C	3.68×10^7	0.990	0.433	-0.363
13A	Alachlor	C	2.22×10^8	0.981	2.615	0.418
13B	2-Chloro-N-(2,6-diethylphenyl)acetamide	C	7.11×10^7	0.976	0.837	-0.077
14A	Ametryn	A	4.30×10^9	0.987	50.547	1.704
16A	Atrazine	B	7.89×10^8	0.976	9.281	0.968
16B	Atrazine-desethyl	B	1.35×10^8	0.994	1.585	0.200
16C	Atrazine-desisopropyl	B	1.17×10^8	0.992	1.376	0.139
25A	Chlorpyrifos-oxon	A	1.72×10^9	0.979	20.277	1.307
27A	DEET	A	6.94×10^8	0.978	8.169	0.912
27B	N-ethyl-m-toluamide	A	4.66×10^8	0.995	5.490	0.740
27C	N,N-diethyl-m-hydroxymethylbenzamide	A	2.31×10^9	0.983	27.162	1.434
33A	Diflubenzuron	D	1.04×10^6	0.994	0.012	-1.910
33B	1-(4-Chlorophenyl)urea	D	2.73×10^6	0.995	0.032	-1.494
39A	Diuron	C	1.75×10^8	0.980	2.055	0.313
39B	Diuron-desdimethyl	D	7.82×10^5	0.997	0.009	-2.036
39C	Diuron-desmonomethyl	C	2.87×10^7	0.983	0.338	-0.471
43A	Fenthion-sulfoxide	A	2.35×10^9	0.977	27.669	1.442
43B	Fenthion-sulfone	C	3.19×10^7	0.999	0.375	-0.426
44A	Flufenacet	C	4.55×10^7	0.986	0.536	-0.271
44B	Flufenacet OXA	D	1.44×10^5	0.997	0.002	-2.772

Table D.1 – continued from previous page

ID	**Chemical name**	**Concentration group**	**Gradient**	R^2	**RIE**	$\log_{10}$**RIE**
46A	Imidazole	B	4.79×10^8	0.990	5.638	0.751
46B	1-methylimidazole	B	2.15×10^8	0.985	2.530	0.403
46C	1,2-dimethylimidazole	B	3.71×10^8	0.974	4.368	0.640
46D	2-methyl imidazole	B	7.62×10^8	0.988	8.966	0.953
46E	4-methyl imidazole	B	7.05×10^8	0.985	8.292	0.919
46F	2-ethyl-4-methyl imidazole	B	1.16×10^9	0.982	13.615	1.134
46G	1H-imidazole-1-propylamine	B	5.53×10^8	0.999	6.510	0.814
46H	2-ethylimidazole	B	1.32×10^9	0.996	15.576	1.192
46I	1-ethyl-1H-imidazole	D	4.82×10^8	0.993	5.671	0.754
46J	1-vinylimidazole	B	1.13×10^8	0.998	1.335	0.126
48A	Isoproturon	A	2.65×10^9	0.988	31.167	1.494
48B	4-Isopropylaniline	A	2.66×10^7	0.998	0.313	-0.505
48C	Isoproturon-didemethyl	B	1.26×10^8	0.980	1.486	0.172
48D	Isoproturon-monodemethyl	B	5.34×10^8	0.980	6.285	0.798
50A	Kresoxim-methyl	C	1.48×10^8	0.987	1.747	0.242
50B	Kresoxim-methyl acid	C	1.53×10^7	0.973	0.180	-0.745
54A	Metalaxyl	A	2.93×10^9	0.985	34.450	1.537
54B	Metalaxyl-hydroxymethyl	B	8.76×10^8	0.991	10.310	1.013
54C	Metalaxyl acid	B	5.64×10^8	0.986	6.633	0.822
58A	Metolachlor	B	1.92×10^9	0.987	22.558	1.353
58B	Metolachlor OXA	C	2.14×10^7	0.978	0.252	-0.598
58C	Metolachlor-Morpholinone	A	3.76×10^9	0.988	44.237	1.646
60A	Metsulfuron-methyl	B	6.92×10^8	0.992	8.147	0.911
60B	Methyl 2-(aminosulfonyl)benzoate	D	1.94×10^5	0.998	0.002	-2.642
60C	2-Amino-4-methoxy-6-methyl-1,3,5 triazine	B	1.89×10^8	0.992	2.230	0.348
61A	N,N-Dimethylaniline	D	4.23×10^7	0.997	0.498	-0.303
61B	N-Methylaniline	D	3.69×10^7	0.999	0.434	-0.362

Table D.1 – continued from previous page

ID	**Chemical name**	**Concentration group**	**Gradient**	R^2	**RIE**	$\log_{10}$**RIE**
65A	p-Toluidine	D	2.19×10^7	1.000	0.258	-0.588
65B	4'-Methylacetanilide	D	2.55×10^7	0.982	0.300	-0.524
70A	Prometon	A	5.10×10^9	0.989	60.045	1.778
71A	Prometryn	A	4.69×10^9	0.991	55.173	1.742
73A	Propazine	A	1.32×10^9	0.995	15.571	1.192
73B	Propazine-2-hydroxy	A	1.35×10^9	0.991	15.894	1.201
77A	Terbuthylazine	A	1.28×10^9	0.990	15.094	1.179
77C	Terbutylazine-desethyl	B	2.37×10^8	0.993	2.785	0.445
78A	Terbutryn	A	3.66×10^9	0.986	43.096	1.634
78B	Terbumeton	A	4.05×10^9	0.985	47.638	1.678
80A	Thiacloprid	B	7.83×10^8	0.974	9.217	0.965
80B	Thiacloprid-amide	B	7.02×10^8	0.992	8.263	0.917
81A	Thifensulfron methyl	B	5.16×10^8	0.996	6.073	0.783
82A	Triadimenol	C	1.49×10^8	0.998	1.751	0.243
84A	Trinexapac-ethyl	C	8.21×10^6	0.998	0.097	-1.015
84B	Trinexapac acid	D	3.87×10^5	0.997	0.005	-2.342
ANCHOR	Tetraethylammonium chloride	A	8.50×10^7	0.991	1.000	0.000

BIBLIOGRAPHY

[1] Aalizadeh, R., Nika, M.-C., and Thomaidis, N. S. (2019). Development and application of retention time prediction models in the suspect and non-target screening of emerging contaminants. *Journal of Hazardous materials*, 363:277–285.

[2] Aalizadeh, R., Nikolopoulou, V., Alygizakis, N., Slobodnik, J., and Thomaidis, N. S. (2022). A novel workflow for semi-quantification of emerging contaminants in environmental samples analyzed by LC-HRMS. *Analytical and Bioanalytical Chemistry*, 414(25):7435–7450.

[3] Achar, J., Cronin, M. T., Firman, J. W., and Öberg, G. (2024). A problem formulation framework for the application of in silico toxicology methods in chemical risk assessment. *Archives of Toxicology*, pages 1–14.

[4] Acharya, H., Vembanur, S., Jamadagni, S. N., and Garde, S. (2010). Mapping hydrophobicity at the nanoscale: Applications to heterogeneous surfaces and proteins. *Faraday discussions*, 146:353–365.

[5] Ahlers, J., Diderich, R., Klaschka, U., Marschner, A., and Schwarz-Schulz, B. (1994). Environmental risk assessment of existing chemicals. *Environmental science and pollution research*, 1:117–123.

[6] Alexander, D. L., Tropsha, A., and Winkler, D. A. (2015). Beware of R 2: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. *Journal of chemical information and modeling*, 55(7):1316–1322.

[7] Alin, A. (2010). Multicollinearity. *Wiley interdisciplinary reviews: computational statistics*, 2(3):370–374.

[8] Alonso, A., Marsal, S., and Julià, A. (2015). Analytical methods in untargeted metabolomics: state of the art in 2015. *Frontiers in bioengineering and biotechnology*, 3:23.

[9] Altshuler, I., Demiri, B., Xu, S., Constantin, A., Yan, N. D., and Cristescu, M. E. (2011). An integrated multi-disciplinary approach for studying multiple stressors in freshwater ecosystems: Daphnia as a model organism. *Integrative and comparative biology*, 51(4):623–633.

[10] Amézqueta, S., Subirats, X., Fuguet, E., Rosés, M., and Ràfols, C. (2020). Octanol-water partition constant. *Liquid-phase extraction*, pages 183–208.

[11] Anderson, D. H. (2013). *Compartmental modeling and tracer kinetics*, volume 50. Springer Science & Business Media.

[12] Ankley, G. T., Cureton, P., Hoke, R. A., Houde, M., Kumar, A., Kurias, J., Lanno, R., McCarthy, C., Newsted, J., Salice, C. J., et al. (2021). Assessing the ecological risks of per-and polyfluoroalkyl substances: Current state-of-the science and a proposed path forward. *Environmental toxicology and chemistry*, 40(3):564–605.

[13] Armitage, J. M., Arnot, J. A., Wania, F., and Mackay, D. (2013). Development and evaluation of a mechanistic bioconcentration model for ionogenic organic chemicals in fish. *Environmental toxicology and chemistry*, 32(1):115–128.

[14] Armitage, J. M., Erickson, R. J., Luckenbach, T., Ng, C. A., Prosser, R. S., Arnot, J. A., Schirmer, K., and Nichols, J. W. (2017). Assessing the bioaccumulation potential of ionizable organic compounds: Current knowledge and research priorities. *Environmental toxicology and chemistry*, 36(4):882–897.

[15] Arnot, J. A. and Gobas, F. A. (2003). A generic QSAR for assessing the bioaccumulation potential of organic chemicals in aquatic food webs. *QSAR & Combinatorial Science*, 22(3):337–345.

[16] Arnot, J. A. and Gobas, F. A. (2004). A food web bioaccumulation model for organic chemicals in aquatic ecosystems. *Environmental Toxicology and Chemistry: An International Journal*, 23(10):2343–2355.

[17] Arnot, J. A. and Gobas, F. A. (2006). A review of bioconcentration factor (BCF) and bioaccumulation factor (BAF) assessments for organic chemicals in aquatic organisms. *Environmental Reviews*, 14(4):257–297.

[18] Arnot, J. A., Mackay, D., and Bonnell, M. (2008). Estimating metabolic biotransformation rates in fish from laboratory data. *Environmental Toxicology and Chemistry: An International Journal*, 27(2):341–351.

[19] Aronoff, G. R., Bergstrom, R. F., Pottratz, S. T., Sloan, R. S., Wolen, R. L., and Lemberger, L. (1984). Fluoxetine kinetics and protein binding in normal and impaired renal function. *Clinical Pharmacology & Therapeutics*, 36(1):138–144.

[20] Arp, H. P. H., Brown, T., Berger, U., and Hale, S. (2017). Ranking REACH registered neutral, ionizable and ionic organic chemicals based on their aquatic persistency and mobility. *Environmental Science: Processes & Impacts*, 19(7):939–955.

[21] Ashauer, R. and Escher, B. I. (2010). Advantages of toxicokinetic and toxicodynamic modelling in aquatic ecotoxicology and risk assessment. *Journal of Environmental Monitoring*, 12(11):2056–2061.

[22] Astuto, M. C., Di Nicola, M. R., Tarazona, J. V., Rortais, A., Devos, Y., Liem, A. D., Kass, G. E., Bastaki, M., Schoonjans, R., Maggiore, A., et al. (2022). In silico methods for environmental risk assessment: Principles, tiered approaches, applications, and future perspectives. In *In Silico methods for predicting drug toxicity*, pages 589–636. Springer.

[23] Baker, M. T. and Van Dyke, R. A. (1984). Metabolism-dependent binding of the chlorinated insecticide DDT and its metabolite, DDD, to microsomal protein and lipids. *Biochemical pharmacology*, 33(2):255–260.

[24] Barron, M. G., Schultz, I. R., and Hayton, W. L. (1989). Presystemic branchial metabolism limits di-2-ethylhexyl phthalate accumulation in fish. *Toxicology and applied pharmacology*, 98(1):49–57.

[25] Bartels, M., Rick, D., Lowe, E., Loizou, G., Price, P., Spendiff, M., Arnold, S., Cocker, J., and Ball, N. (2012). Development of PK-and PBPK-based modeling tools for derivation of biomonitoring guidance values. *Computer methods and programs in biomedicine*, 108(2):773–788.

[26] Baudrot, V. and Charles, S. (2019). Recommendations to address uncertainties in environmental risk assessment using toxicokinetic-toxicodynamic models. *Scientific reports*, 9(1):11432.

[27] Belfield, S. J., Cronin, M. T., Enoch, S. J., and Firman, J. W. (2023). Guidance for good practice in the application of machine learning in development of toxicological quantitative structure-activity relationships (QSARs). *Plos one*, 18(5):e0282924.

[28] Bell, R. (1947). The use of the terms "acid" and "base". *Quarterly Reviews, Chemical Society*, 1(2):113–125.

[29] Benfenati, E., Chaudhry, Q., Gini, G., and Dorne, J. L. (2019). Integrating in silico models and read-across methods for predicting toxicity of chemicals: A step-wise strategy. *Environment international*, 131:105060.

[30] Benfenati, E., Gini, G., Piclin, N., Roncaglioni, A., and Varı, M. (2003). Predicting logP of pesticides using different software. *Chemosphere*, 53(9):1155–1164.

[31] Bertelsen, S. L., Hoffman, A. D., Gallinat, C. A., Elonen, C. M., and Nichols, J. W. (1998). Evaluation of log Kow and tissue lipid content as predictors of chemical par-

titioning to fish tissues. *Environmental Toxicology and Chemistry: An International Journal*, 17(8):1447–1455.

[32] Beyer, J., Petersen, K., Song, Y., Ruus, A., Grung, M., Bakke, T., and Tollefsen, K. E. (2014). Environmental risk assessment of combined effects in aquatic ecotoxicology: a discussion paper. *Marine environmental research*, 96:81–91.

[33] Bhattacharya, A. A., Curry, S., and Franks, N. P. (2000). Binding of the general anesthetics propofol and halothane to human serum albumin: high resolution crystal structures. *Journal of Biological Chemistry*, 275(49):38731–38738.

[34] Bifarin, O. O. (2023). Interpretable machine learning with tree-based shapley additive explanations: Application to metabolomics datasets for binary classification. *Plos one*, 18(5):e0284315.

[35] Binetti, R., Costamagna, F. M., and Marcello, I. (2008). Exponential growth of new chemicals and evolution of information relevant to risk control. *Annali-Istituto Superiore di Sanità*, 44(1):13.

[36] Bintein, S., Devillers, J., and Karcher, W. (1993). Nonlinear dependence of fish bioconcentration on n-octanol/water partition coefficient. *SAR and QSAR in Environmental Research*, 1(1):29–39.

[37] Birnbaum, L. S., Burke, T. A., and Jones, J. J. (2016). Informing 21st-century risk assessments with 21st-century science. *Environmental health perspectives*, 124(4):A60–A63.

[38] Bittermann, K., Spycher, S., and Goss, K.-U. (2016). Comparison of different models predicting the phospholipid-membrane water partition coefficients of charged compounds. *Chemosphere*, 144:382–391.

[39] Bogatova, I., Shcherbina, M., Ovinnikova, B., and Tagirova, N. (1971). Chemical

composition of some planktonic animals under different conditions of growing. *Gidrobiologiceski Zurnal*, 7(5):54–57.

[40] Bogut, I., Adamek, Z., Puškadija, Z., and Galović, D. (2010). Nutritional value of planktonic cladoceran Daphnia magna for common carp (Cyprinus carpio) fry feeding. *Croatian Journal of Fisheries: Ribarstvo*, 68(1):1–10.

[41] Bohnert, T. and Gan, L.-S. (2013). Plasma protein binding: from discovery to development. *Journal of pharmaceutical sciences*, 102(9):2953–2994.

[42] Bois, F. Y. (1999). Analysis of PBPK models for risk characterization. *Annals of the New York Academy of Sciences*, 895(1):317–337.

[43] Bois, F. Y. (2000). Statistical analysis of Clewell et al. PBPK model of trichloroethylene kinetics. *Environmental health perspectives*, 108(suppl 2):307–316.

[44] Boström, M. L., Ugge, G., Jönsson, J. Å., and Berglund, O. (2017). Bioaccumulation and trophodynamics of the antidepressants sertraline and fluoxetine in laboratory-constructed, 3-level aquatic food chains. *Environmental Toxicology and Chemistry*, 36(4):1029–1037.

[45] Bouhedjar, K., Benfenati, E., and Nacereddine, A. (2020). Modelling quantitative structure activity–activity relationships (QSAARs): auto-pass-pass, a new approach to fill data gaps in environmental risk assessment under the REACH regulation. *SAR and QSAR in Environmental Research*, 31(10):785–801.

[46] Boyden, C. (1974). Trace element content and body size in molluscs. *Nature*, 251(5473):311–314.

[47] Brendelberger, H. and Geller, W. (1985). Variability of filter structures in eight Daphnia species: mesh sizes and filtering areas. *Journal of Plankton Research*, 7(4):473–486.

[48] Briggs, G. G. (1981). Theoretical and experimental relationships between soil adsorption, octanol-water partition coefficients, water solubilities, bioconcentration factors, and the parachor. *Journal of Agricultural and Food Chemistry*, 29(5):1050–1059.

[49] Brinkmann, M., Preuss, T. G., and Hollert, H. (2017). Advancing in vitro–in vivo extrapolations of mechanism-specific toxicity data through toxicokinetic modeling. *In vitro environmental toxicology-concepts, application and assessment*, pages 293–317.

[50] Buhler, D. R. and Williams, D. E. (1988). The role of biotransformation in the toxicity of chemicals. *Aquatic Toxicology*, 11(1-2):19–28.

[51] Capecchi, A., Probst, D., and Reymond, J.-L. (2020). One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *Journal of cheminformatics*, 12:1–15.

[52] Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, 76.

[53] Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Garcia-Vallvé, S., and Pujadas, G. (2015). Molecular fingerprint similarity search in virtual screening. *Methods*, 71:58–63.

[54] Cevc, G. (2015). Partition coefficient vs. binding constant: How best to assess molecular lipophilicity. *European Journal of Pharmaceutics and Biopharmaceutics*, 92:204–215.

[55] Chalcraft, K. R., Lee, R., Mills, C., and Britz-McKibbin, P. (2009). Virtual quantification of metabolites by capillary electrophoresis-electrospray ionization-mass spectrometry: predicting ionization efficiency without chemical standards. *Analytical Chemistry*, 81(7):2506–2515.

[56] Chen, C. and Kim, S. (2013). LC-MS-based metabolomics of xenobiotic-induced toxicities. *Computational and structural biotechnology journal*, 4(5):e201301008.

[57] Chen, C. C. and Kuo, D. T. F. (2018). Bioconcentration model for non-ionic, polar, and ionizable organic compounds in amphipod. *Environmental toxicology and chemistry*, 37(5):1378–1386.

[58] Chibwe, L., Titaley, I. A., Hoh, E., and Simonich, S. L. M. (2017). Integrated framework for identifying toxic transformation products in complex environmental mixtures. *Environmental science & technology letters*, 4(2):32–43.

[59] Clewell, H. J. and Andersen, M. E. (1985). Risk assessment extrapolations and physiological modeling. *Toxicology and industrial health*, 1(4):111–134.

[60] Coecke, S., Pelkonen, O., Leite, S. B., Bernauer, U., Bessems, J. G., Bois, F. Y., Gundert-Remy, U., Loizou, G., Testai, E., and Zaldívar, J.-M. (2013). Toxicokinetics as a key to the integrated toxicity risk assessment based primarily on non-animal approaches. *Toxicology in vitro*, 27(5):1570–1577.

[61] Committee, E. S., Benford, D., Halldorsson, T., Jeger, M. J., Knutsen, H. K., More, S., Naegeli, H., Noteborn, H., Ockleford, C., Ricci, A., et al. (2018). Guidance on uncertainty analysis in scientific assessments. *Efsa Journal*, 16(1):e05123.

[62] Committee, E. S., More, S. J., Bampidis, V., Benford, D., Bragard, C., Halldorsson, T. I., Hernández-Jerez, A. F., Bennekou, S. H., Koutsoumanis, K., Lambré, C., et al. (2022). Guidance on the use of the benchmark dose approach in risk assessment. *Efsa Journal*, 20(10):e07584.

[63] Cox, D. (2006). Frequentist and Bayesian statistics: A critique (keynote address). In *Statistical problems in particle physics, astrophysics and cosmology*, pages 3–6. World Scientific.

[64] Crank, J. (1979). *The mathematics of diffusion*. Oxford university press.

[65] Creton, S., Saghir, S. A., Bartels, M. J., Billington, R., Bus, J. S., Davies, W., Dent, M. P., Hawksworth, G. M., Parry, S., and Travis, K. Z. (2012). Use of toxicokinetics to support chemical evaluation: Informing high dose selection and study interpretation. *Regulatory Toxicology and Pharmacology*, 62(2):241–247.

[66] Croom, E. (2012). Metabolism of xenobiotics of human environments. *Progress in molecular biology and translational science*, 112:31–88.

[67] Csizmadia, F., Tsantili-Kakoulidou, A., Panderi, I., and Darvas, F. (1997). Prediction of distribution coefficient from structure. 1. Estimation method. *Journal of pharmaceutical sciences*, 86(7):865–871.

[68] Dai, Z., Xia, X., Guo, J., and Jiang, X. (2013). Bioaccumulation and uptake routes of perfluoroalkyl acids in Daphnia magna. *Chemosphere*, 90(5):1589–1596.

[69] Dale, V. H., Biddinger, G. R., Newman, M. C., Oris, J. T., Suter, G. W., Thompson, T., Armitage, T. M., Meyer, J. L., Allen-King, R. M., Burton, G. A., et al. (2008). Enhancing the ecological risk assessment process. *Integrated environmental assessment and management*, 4(3):306–313.

[70] Dalhoff, K., Gottardi, M., Kretschmann, A., and Cedergreen, N. (2016). What causes the difference in synergistic potentials of propiconazole and prochloraz toward pyrethroids in Daphnia magna? *Aquatic Toxicology*, 172:95–102.

[71] De Groot, R., Brekelmans, P., Herremans, J., and Meulenbelt, J. (2010). The changes in hazard classification and product notification procedures of the new European CLP and Cosmetics Regulations. *Clinical Toxicology*, 48(1):28–33.

[72] Dearden, J., Netzeva, T., and Bibby, R. (2003). A comparison of commercially available software for the prediction of partition coefficient. *Designing Drugs and Crop Protectants: Processes, Problems and Solutions, Blackwell, Oxford*, pages 168–169.

[73] Dearden, J. and Worth, A. (2007). In silico prediction of physicochemical properties. *JRC Scientific and Technical Reports, EUR*, 23051:1–68.

[74] Debruyn, A. M. and Gobas, F. A. (2007). The sorptive capacity of animal protein. *Environmental Toxicology and Chemistry: An International Journal*, 26(9):1803–1808.

[75] del Carmen Gómez-Regalado, M., Martín, J., Santos, J. L., Aparicio, I., Alonso, E., and Zafra-Gómez, A. (2023). Bioaccumulation/bioconcentration of pharmaceutical active compounds in aquatic organisms: Assessment and factors database. *Science of the Total Environment*, 861:160638.

[76] Delabrière, A., Hohenester, U. M., Colsch, B., Junot, C., Fenaille, F., and Thévenot, E. A. (2017). proFIA: a data preprocessing workflow for flow injection analysis coupled to high-resolution mass spectrometry. *Bioinformatics*, 33(23):3767–3775.

[77] Devos, Y., Craig, W., Devlin, R. H., Ippolito, A., Leggatt, R. A., Romeis, J., Shaw, R., Svendsen, C., and Topping, C. J. (2019). Using problem formulation for fit-for-purpose pre-market environmental risk assessments of regulated stressors. *EFSA Journal*, 17:e170708.

[78] Diedenhofen, M., Eckert, F., and Terzi, S. (2023). COSMO-RS blind prediction of distribution coefficients and aqueous pKa values from the SAMPL8 challenge. *Journal of Computer-Aided Molecular Design*, 37(8):395–405.

[79] Dimitrov, S., Mekenyan, O., and Walker, J. (2002). Non-linear modeling of bioconcentration using partition coefficients for narcotic chemicals. *SAR and QSAR in Environmental Research*, 13(1):177–184.

[80] Ding, J., Lu, G., Liu, J., Yang, H., and Li, Y. (2016). Uptake, depuration, and bioconcentration of two pharmaceuticals, roxithromycin and propranolol, in Daphnia magna. *Ecotoxicology and environmental safety*, 126:85–93.

[81] Ding, J., Zou, H., Liu, Q., Zhang, S., and Razanajatovo, R. M. (2017). Bioconcentration of the antidepressant fluoxetine and its effects on the physiological and biochemical status in Daphnia magna. *Ecotoxicology and environmental safety*, 142:102–109.

[82] Diouf, A., Camara, B. I., Ngom, D., Toumi, H., Felten, V., Masfaraud, J.-F., and Ferard, J.-F. (2018). Bayesian inference of a dynamical model evaluating Deltamethrin effect on Daphnia survival. *Biomath*, 7(2):ID–1812177.

[83] Djoumbou-Feunang, Y., Fiamoncini, J., Gil-de-la Fuente, A., Greiner, R., Manach, C., and Wishart, D. S. (2019). BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *Journal of cheminformatics*, 11:1–25.

[84] Ebert, D. (2005). *Ecology, epidemiology, and evolution of parasitism in Daphnia*. National Library of Medicine.

[85] Ebert, I., Bachmann, J., Kühnen, U., Küster, A., Kussatz, C., Maletzki, D., and Schlüter, C. (2011). Toxicity of the fluoroquinolone antibiotics enrofloxacin and ciprofloxacin to photoautotrophic aquatic organisms. *Environmental toxicology and chemistry*, 30(12):2786–2792.

[86] Egeghy, P. P., Judson, R., Gangwal, S., Mosher, S., Smith, D., Vail, J., and Hubal, E. A. C. (2012). The exposure data landscape for manufactured chemicals. *Science of the Total Environment*, 414:159–166.

[87] Ellgehausen, H., Guth, J. A., and Esser, H. O. (1980). Factors determining the bioaccumulation potential of pesticides in the individual compartments of aquatic food chains. *Ecotoxicology and environmental safety*, 4(2):134–157.

[88] Endo, S., Bauerfeind, J., and Goss, K.-U. (2012). Partitioning of neutral organic compounds to structural proteins. *Environmental science & technology*, 46(22):12697–12703.

[89] Erturan, A. M., Karaduman, G., and Durmaz, H. (2023). Machine learning-based approach for efficient prediction of toxicity of chemical gases using feature selection. *Journal of hazardous materials*, 455:131616.

[90] Escher, B. I. and Fenner, K. (2011). Recent advances in environmental risk assessment of transformation products. *Environmental science & technology*, 45(9):3835–3847.

[91] Feng, Y., Gao, K., and Lacasse, S. (2024). Bayesian partial pooling to reduce uncertainty in overcoring rock stress estimation. *Journal of Rock Mechanics and Geotechnical Engineering*, 16(4):1192–1201.

[92] Fent, K. and Looser, P. W. (1995). Bioaccumulation and bioavailability of tributyltin chloride: influence of pH and humic acids. *Water Research*, 29(7):1631–1637.

[93] Ferson, S. (2005). Bayesian methods in risk assessment. *Unpublished Report Prepared for the Bureau de Recherches Geologiques et Minieres (BRGM). New York*.

[94] Fick, A. (1855). V. on liquid diffusion. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 10(63):30–39.

[95] Fischer, F. C., Ludtke, S., Thackray, C., Pickard, H. M., Haque, F., Dassuncao, C., Endo, S., Schaider, L., and Sunderland, E. M. (2024). Binding of Per-and Polyfluoroalkyl Substances (PFAS) to Serum Proteins: Implications for Toxicokinetics in Humans. *Environmental Science & Technology*, 58(2):1055–1063.

[96] Fischer, I., Milton, C., and Wallace, H. (2020). Toxicity testing is evolving! *Toxicology Research*, 9(2):67–80.

[97] Fisher, R. A. (1970). Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*, pages 66–70. Springer.

[98] Fu, Q., Fedrizzi, D., Kosfeld, V., Schlechtriem, C., Ganz, V., Derrer, S., Rentsch, D., and Hollender, J. (2020). Biotransformation changes bioaccumulation and toxicity of diclofenac in aquatic organisms. *Environmental science & technology*, 54(7):4400–4408.

[99] Fu, W., Franco, A., and Trapp, S. (2009). Methods for estimating the bioconcentration factor of ionizable organic chemicals. *Environmental Toxicology and Chemistry: An International Journal*, 28(7):1372–1379.

[100] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

[101] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2020). *Bayesian data analysis*. Chapman and Hall/CRC.

[102] Gergs, A., Classen, S., Strauss, T., Ottermanns, R., Brock, T. C., Ratte, H. T., Hommen, U., and Preuss, T. G. (2016). Ecological recovery potential of freshwater organisms: Consequences for environmental risk assessment of chemicals. *Reviews of Environmental Contamination and Toxicology Volume 236*, pages 259–294.

[103] Gobas, F. A., Wilcockson, J. B., Russell, R. W., and Haffner, G. D. (1999). Mechanism of biomagnification in fish under laboratory and field conditions. *Environmental science & technology*, 33(1):133–141.

[104] Golosovskaia, E., Örn, S., Ahrens, L., Chelcea, I., and Andersson, P. L. (2024). Studying mixture effects on uptake and tissue distribution of PFAS in zebrafish (Danio rerio) using physiologically based kinetic (PBK) modelling. *Science of the Total Environment*, 912:168738.

[105] Grech, A., Brochot, C., Dorne, J.-L., Quignot, N., Bois, F. Y., and Beaudouin, R. (2017). Toxicokinetic models and related tools in environmental risk assessment of chemicals. *Science of the Total Environment*, 578:1–15.

[106] Groff, L. C., Grossman, J. N., Kruve, A., Minucci, J. M., Lowe, C. N., McCord, J. P., Kapraun, D. F., Phillips, K. A., Purucker, S. T., Chao, A., et al. (2022). Uncertainty estimation strategies for quantitative non-targeted analysis. *Analytical and bioanalytical chemistry*, 414(17):4919–4933.

[107] Hack, C. E. (2006). Bayesian analysis of physiologically based toxicokinetic and toxicodynamic models. *Toxicology*, 221(2-3):241–248.

[108] Haith, D. A. (2010). Ecological risk assessment of pesticide runoff from grass surfaces. *Environmental science & technology*, 44(16):6496–6502.

[109] Hall, L. H., Mohney, B., and Kier, L. B. (1991). The electrotopological state: an atom index for QSAR. *Quantitative Structure-Activity Relationships*, 10(1):43–51.

[110] Hampel, F. R. (1998). On the foundations of statistics: A frequentist approach. In *Research report/Seminar für Statistik, Eidgenössische Technische Hochschule Zürich*, volume 85. Seminar für Statistik, ETH.

[111] Hatsis, P., Waters, N. J., and Argikar, U. A. (2017). Implications for metabolite quantification by mass spectrometry in the absence of authentic standards. *Drug Metabolism and Disposition*, 45(5):492–496.

[112] Hebert, P. D. (1978). The population biology of daphnia (crustacea, daphnidae). *Biological Reviews*, 53(3):387–426.

[113] Hendriks, A. J. and Heikens, A. (2001). The power of size. 2. Rate constants and equilibrium ratios for accumulation of inorganic substances related to species weight. *Environmental Toxicology and Chemistry: An International Journal*, 20(7):1421–1437.

[114] Hendriks, A. J., van der Linde, A., Cornelissen, G., and Sijm, D. T. (2001). The power of size. 1. Rate constants and equilibrium ratios for accumulation of organic substances related to octanol-water partition ratio and species weight. *Environmental Toxicology and Chemistry: An International Journal*, 20(7):1399–1420.

[115] Hermens, J. L., de Bruijn, J. H., and Brooke, D. N. (2013). The octanol–water partition coefficient: strengths and limitations. *Environmental toxicology and chemistry*, 32(4):732–733.

[116] Hird, S. J., Lau, B. P.-Y., Schuhmacher, R., and Krska, R. (2014). Liquid chromatography-mass spectrometry for the determination of chemical contaminants in food. *TrAC Trends in Analytical Chemistry*, 59:59–72.

[117] Hodges, G., Roberts, D. W., Marshall, S. J., and Dearden, J. C. (2006). The aquatic toxicity of anionic surfactants to Daphnia magna—a comparative QSAR study of linear alkylbenzene sulphonates and ester sulphonates. *Chemosphere*, 63(9):1443–1450.

[118] Hollas, B. (2003). An analysis of the autocorrelation descriptor for molecules. *Journal of mathematical chemistry*, 33:91–101.

[119] Hope, B. K. and Clarkson, J. R. (2014). A strategy for using weight-of-evidence methods in ecological risk assessments. *Human and Ecological Risk Assessment: An International Journal*, 20(2):290–315.

[120] Huang, A., van den Brink, N. W., Buijse, L., Roessink, I., and van den Brink, P. J. (2021). The toxicity and toxicokinetics of imidacloprid and a bioactive metabolite to two aquatic arthropod species. *Aquatic Toxicology*, 235:105837.

[121] Hubble, J. (2001). Stochastic modeling of affinity adsorption. *Biotechnology progress*, 17(3):565–567.

[122] Huntscha, S., Hofstetter, T. B., Schymanski, E. L., Spahr, S., and Hollender, J. (2014). Biotransformation of benzotriazoles: insights from transformation product identification and compound-specific isotope analysis. *Environmental science & technology*, 48(8):4435–4443.

[123] Jeon, J., Kurth, D., Ashauer, R., and Hollender, J. (2013). Comparative toxicokinetics of organic micropollutants in freshwater crustaceans. *Environmental science & technology*, 47(15):8809–8817.

[124] Jordão, R., Casas, J., Fabrias, G., Campos, B., Piña, B., Lemos, M. F., Soares, A. M., Tauler, R., and Barata, C. (2015). Obesogens beyond vertebrates: lipid per-

turbation by tributyltin in the crustacean Daphnia magna. *Environmental Health Perspectives*, 123(8):813–819.

[125] Jorge, T. F., Mata, A. T., and António, C. (2016). Mass spectrometry as a quantitative tool in plant metabolomics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2079):20150370.

[126] Kah, M. and Brown, C. D. (2008). Log D: Lipophilicity for ionisable compounds. *Chemosphere*, 72(10):1401–1408.

[127] Karaduman, G. and Kelleci Çelik, F. (2023). 2D-Quantitative structure–activity relationship modeling for risk assessment of pharmacotherapy applied during pregnancy. *Journal of Applied Toxicology*, 43(10):1436–1446.

[128] Kastritis, P. L. and Bonvin, A. M. (2013). On the binding affinity of macromolecular interactions: daring to ask why proteins interact. *Journal of The Royal Society Interface*, 10(79):20120835.

[129] Kavlock, R. J., Bahadori, T., Barton-Maclaren, T. S., Gwinn, M. R., Rasenberg, M., and Thomas, R. S. (2018). Accelerating the pace of chemical risk assessment. *Chemical research in toxicology*, 31(5):287–290.

[130] Khabazbashi, S., Engelhardt, J., Möckel, C., Weiss, J., and Kruve, A. (2022). Estimation of the concentrations of hydroxylated polychlorinated biphenyls in human serum using ionization efficiency prediction for electrospray. *Analytical and Bioanalytical Chemistry*, 414(25):7451–7460.

[131] Kim, H. Y., Jeon, J., Hollender, J., Yu, S., and Kim, S. D. (2014). Aqueous and dietary bioaccumulation of antibiotic tetracycline in D. magna and its multigenerational transfer. *Journal of hazardous materials*, 279:428–435.

[132] Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker,

B. A., Thiessen, P. A., Yu, B., et al. (2023). PubChem 2023 update. *Nucleic acids research*, 51(D1):D1373–D1380.

[133] Kleiven, O. T., Larsson, P., and Hobæk, A. (1992). Sexual reproduction in Daphnia magna requires three stimuli. *Oikos*, pages 197–206.

[134] Knight, K. (2000). *Mathematical Statistics*. Chapman and Hall/CRC.

[135] Koivisto, S. (1995). Is Daphnia magna an ecologically representative zooplankton species in toxicity tests? *Environmental Pollution*, 90(2):263–267.

[136] Kosjek, T., Heath, E., Petrović, M., and Barceló, D. (2007). Mass spectrometry for identifying pharmaceutical biotransformation products in the environment. *TrAC Trends in Analytical Chemistry*, 26(11):1076–1085.

[137] Krantzberg, G. (1989). Metal accumulation by chironomid larvae: the effects of age and body weight on metal body burdens. In *Environmental Bioassay Techniques and their Application: Proceedings of the 1st International Conference held in Lancaster, England, 11–14 July 1988*, pages 497–506. Springer.

[138] Kratochwil, N. A., Huber, W., Müller, F., Kansy, M., and Gerber, P. R. (2002). Predicting plasma protein binding of drugs: a new approach. *Biochemical pharmacology*, 64(9):1355–1374.

[139] Kretschmann, A., Ashauer, R., Preuss, T. G., Spaak, P., Escher, B. I., and Hollender, J. (2011). Toxicokinetic model describing bioconcentration and biotransformation of diazinon in Daphnia magna. *Environmental science & technology*, 45(11):4995–5002.

[140] Krishnan, K. and Peyret, T. (2009). Physiologically based toxicokinetic (PBTK) modeling in ecotoxicology. *Ecotoxicology modeling*, pages 145–175.

[141] Kruve, A. and Kaupmees, K. (2017). Predicting ESI/MS signal change for anions in different solvents. *Analytical chemistry*, 89(9):5079–5086.

[142] Kruve, A., Kaupmees, K., Liigand, J., and Leito, I. (2014). Negative electrospray ionization via deprotonation: predicting the ionization efficiency. *Analytical chemistry*, 86(10):4822–4830.

[143] Kruve, A., Kiefer, K., and Hollender, J. (2021). Benchmarking of the quantification approaches for the non-targeted screening of micropollutants and their transformation products in groundwater. *Analytical and Bioanalytical Chemistry*, 413:1549–1559.

[144] Kukkonen, J. and Oikari, A. (1988). Sulphate conjugation is the main route of pentachlorophenol metabolism in Daphnia magna. *Comparative Biochemistry and Physiology Part C: Comparative Pharmacology*, 91(2):465–468.

[145] Kuo, D. T. and Di Toro, D. M. (2013). A reductionist mechanistic model for bioconcentration of neutral and weakly polar organic compounds in fish. *Environmental toxicology and chemistry*, 32(9):2089–2099.

[146] Kuo, D. T. and Di Toro, D. M. (2022). Determination of In Vivo Biotransformation Kinetics Using Early-Time Biota Concentrations. *Environmental Toxicology and Chemistry*, 41(1):148–158.

[147] Kuo, D. T. F. and Chen, C. C. (2016). Deriving in vivo biotransformation rate constants and metabolite parent concentration factor/stable metabolite factor from bioaccumulation and bioconcentration experiments: An illustration with worm accumulation data. *Environmental toxicology and chemistry*, 35(12):2903–2909.

[148] Lampert, W. (2006). Daphnia: model herbivore, predator and prey. *Polish journal of ecology*, 54(4):607–620.

[149] Landrum, P. F., Lydy, M. J., and Lee, H. (1992). Toxicokinetics in aquatic systems: model comparisons and use in hazard assessment. *Environmental Toxicology and Chemistry*, 11(12):1709–1725.

[150] Laplace, P. S. (1774). Mémoire sur la Probabilité des Causes par les évènemens. *Mémoires de Mathématique et de Physique, Présentés à l'Académie Royale des Sciences, Par Divers Savans & Lus Dans ses Assemblées*, 6:621–656.

[151] Latour, R. A. (2015). The Langmuir isotherm: a commonly applied but misleading approach for the analysis of protein adsorption behavior. *Journal of biomedical materials research part A*, 103(3):949–958.

[152] Lei, Z., Huhman, D. V., and Sumner, L. W. (2011). Mass spectrometry strategies in metabolomics. *Journal of Biological Chemistry*, 286(29):25435–25442.

[153] Leito, I., Herodes, K., Huopolainen, M., Virro, K., Künnapas, A., Kruve, A., and Tanner, R. (2008). Towards the electrospray ionization mass spectrometry ionization efficiency scale of organic compounds. *Rapid Communications in Mass Spectrometry: An International Journal Devoted to the Rapid Dissemination of Up-to-the-Minute Research in Mass Spectrometry*, 22(3):379–384.

[154] Leonard, T. H. (2014). A personal history of Bayesian statistics. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(2):80–115.

[155] Li, H., Zhang, Q., Su, H., You, J., and Wang, W.-X. (2020). High tolerance and delayed responses of Daphnia magna to neonicotinoid insecticide imidacloprid: toxicokinetic and toxicodynamic modeling. *Environmental Science & Technology*, 55(1):458–467.

[156] Liigand, J., Wang, T., Kellogg, J., Smedsgaard, J., Cech, N., and Kruve, A. (2020). Quantification for non-targeted LC/MS screening without standard substances. *Scientific reports*, 10(1):5808.

[157] Liigand, P., Liigand, J., Cuyckens, F., Vreeken, R. J., and Kruve, A. (2018). Ionisation efficiencies can be predicted in complicated biological matrices: A proof of concept. *Analytica Chimica Acta*, 1032:68–74.

[158] Lin, H.-I., Berzins, D. W., Myers, L., George, W. J., Abdelghani, A., and Watanabe, K. H. (2004). A Bayesian approach to parameter estimation for a crayfish (Procambarus spp): bioaccumulation model. *Environmental Toxicology and Chemistry: An International Journal*, 23(9):2259–2266.

[159] Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18.

[160] Liu, C., Gin, K. Y., Chang, V. W., Goh, B. P., and Reinhard, M. (2011). Novel perspectives on the bioaccumulation of PFCs–the concentration dependency. *Environmental science & technology*, 45(22):9758–9764.

[161] Liu, W., Zhang, H., Ding, J., He, W., Zhu, L., and Feng, J. (2022). Waterborne and dietary bioaccumulation of organophosphate esters in zooplankton Daphnia magna. *International Journal of Environmental Research and Public Health*, 19(15):9382.

[162] Lotufo, G. R., Landrum, P. F., Gedeon, M. L., Tigue, E. A., and Herche, L. R. (2000). Comparative toxicity and toxicokinetics of DDT and its major metabolites in freshwater amphipods. *Environmental Toxicology and Chemistry: An International Journal*, 19(2):368–379.

[163] Mackay, D. (1982). Correlation of bioconcentration factors. *Environmental Science & Technology*, 16(5):274–278.

[164] Mackay, D., Arnot, J. A., Gobas, F. A., and Powell, D. E. (2013). Mathematical relationships between metrics of chemical bioaccumulation in fish. *Environmental toxicology and chemistry*, 32(7):1459–1466.

[165] Mackay, D. and Fraser, A. (2000). Bioaccumulation of persistent organic chemicals: mechanisms and models. *Environmental pollution*, 110(3):375–391.

[166] Maertens, A., Golden, E., Luechtefeld, T. H., Hoffmann, S., Tsaioun, K., and Har-

tung, T. (2022). Probabilistic risk assessment–the keystone for the future of toxicology. *Altex*, 39(1):3.

[167] Malm, L., Palm, E., Souihi, A., Plassmann, M., Liigand, J., and Kruve, A. (2021). Guide to semi-quantitative non-targeted screening using LC/ESI/HRMS. *Molecules*, 26(12):3524.

[168] Manahan, S. E. (1992). *Toxicological chemistry*. CRC Press.

[169] Manallack, D. T. (2009). The acid–base profile of a contemporary set of drugs: implications for drug discovery. *SAR and QSAR in Environmental Research*, 20(7-8):611–655.

[170] Manallack, D. T., Prankerd, R. J., Yuriev, E., Oprea, T. I., and Chalmers, D. K. (2013). The significance of acid/base properties in drug discovery. *Chemical Society Reviews*, 42(2):485–496.

[171] Mangold-Döring, A., Grimard, C., Green, D., Petersen, S., Nichols, J. W., Hogan, N., Weber, L., Hollert, H., Hecker, M., and Brinkmann, M. (2021). A novel multispecies toxicokinetic modeling approach in support of chemical risk assessment. *Environmental science & technology*, 55(13):9109–9118.

[172] Matuszewski, B. K., Constanzer, M., and Chavez-Eng, C. (2003). Strategies for the assessment of matrix effect in quantitative bioanalytical methods based on HPLC-MS/MS. *Analytical chemistry*, 75(13):3019–3030.

[173] Mayhew, A. W., Topping, D. O., and Hamilton, J. F. (2020). New Approach Combining Molecular Fingerprints and Machine Learning to Estimate Relative Ionization Efficiency in Electrospray Ionization. *ACS omega*, 5(16):9510–9516.

[174] McCarthy, C. J., Roark, S. A., and Middleton, E. T. (2021). Considerations for toxicity experiments and risk assessments with PFAS mixtures. *Integrated environmental assessment and management*, 17(4):697–704.

[175] Meißner, R., Feketeová, L., Ribar, A., Fink, K., Limão-Vieira, P., and Denifl, S. (2019). Electron ionization of imidazole and its derivative 2-nitroimidazole. *Journal of The American Society for Mass Spectrometry*, 30(12):2678–2691.

[176] Meredith-Williams, M., Carter, L. J., Fussell, R., Raffaelli, D., Ashauer, R., and Boxall, A. B. (2012). Uptake and depuration of pharmaceuticals in aquatic invertebrates. *Environmental pollution*, 165:250–258.

[177] Meylan, W. M., Howard, P. H., Boethling, R. S., Aronson, D., Printup, H., and Gouchie, S. (1999). Improved method for estimating bioconcentration/bioaccumulation factor from octanol/water partition coefficient. *Environmental Toxicology and Chemistry: An International Journal*, 18(4):664–672.

[178] Mitchell, J. B. (2014). Machine learning methods in chemoinformatics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 4(5):468–481.

[179] Moffett, D. B., Mumtaz, M. M., Sullivan Jr, D. W., and Whittaker, M. H. (2022). General considerations of dose-effect and dose-response relationships. In *Handbook on the Toxicology of Metals*, pages 299–317. Elsevier.

[180] Moreau, G. and Broto, P. (1980). The autocorrelation of a topological structure: A new molecular descriptor. *Nouveau Journal de Chimie*, 4:359–360.

[181] Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., and Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic bulletin & review*, 23:103–123.

[182] Morgan, H. L. (1965). The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of chemical documentation*, 5(2):107–113.

[183] Morrissey, C., Fritsch, C., Fremlin, K., Adams, W., Borgå, K., Brinkmann, M., Eulaers, I., Gobas, F., Moore, D. R., van den Brink, N., et al. (2023). Advancing exposure

assessment approaches to improve wildlife risk assessment. *Integrated Environmental Assessment and Management*.

[184] MOSAIC (2017). Bioaccumulation database. https://mosaic.univ-lyon1.fr/bioacc. Accessed Sept 2023.

[185] Moxon, T. E., Li, H., Lee, M.-Y., Piechota, P., Nicol, B., Pickles, J., Pendlington, R., Sorrell, I., and Baltazar, M. T. (2020). Application of physiologically based kinetic (PBK) modelling in the next generation risk assessment of dermally applied consumer products. *Toxicology in Vitro*, 63:104746.

[186] Nasiri, A., Jahani, R., Mokhtari, S., Yazdanpanah, H., Daraei, B., Faizi, M., and Kobarfard, F. (2021). Overview, consequences, and strategies for overcoming matrix effects in LC-MS analysis: a critical review. *Analyst*, 146(20):6049–6063.

[187] Neely, W. B., Branson, D. R., and Blau, G. E. (1974). Partition coefficient to measure bioconcentration potential of organic chemicals in fish. *Environmental Science & Technology*, 8(13):1113–1115.

[188] Neyman, J. and Pearson, E. S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337.

[189] Ng, C. A. and Hungerbühler, K. (2013). Bioconcentration of perfluorinated alkyl acids: how important is specific binding? *Environmental science & technology*, 47(13):7214–7223.

[190] Ng, C. A. and Hungerbühler, K. (2014). Bioaccumulation of perfluorinated alkyl acids: observations and models. *Environmental science & technology*, 48(9):4637–4648.

[191] Nichols, J. W., McKim, J. M., Andersen, M. E., Gargas, M. L., Clewell III, H. J., and Erickson, R. J. (1990). A physiologically based toxicokinetic model for the up-

take and disposition of waterborne organic chemicals in fish. *Toxicology and applied pharmacology*, 106(3):433–447.

[192] Nikinmaa, M. (2014). *An introduction to aquatic toxicology*. Elsevier.

[193] Nkoom, M., Lu, G., Liu, J., Dong, H., and Yang, H. (2019a). Bioconcentration, behavioral, and biochemical effects of the non-steroidal anti-inflammatory drug diclofenac in Daphnia magna. *Environmental science and pollution research*, 26:5704–5712.

[194] Nkoom, M., Lu, G., Liu, J., Yang, H., and Dong, H. (2019b). Bioconcentration of the antiepileptic drug carbamazepine and its physiological and biochemical effects on Daphnia magna. *Ecotoxicology and environmental safety*, 172:11–18.

[195] Ockleford, C., Adriaanse, P., Berny, P., Brock, T., Duquesne, S., Grilli, S., Hernandez-Jerez, A. F., Bennekou, S. H., Klein, M., et al. (2018). Scientific Opinion on the state of the art of Toxicokinetic/Toxicodynamic (TKTD) effect models for regulatory risk assessment of pesticides for aquatic organisms. *EFSA journal*, 16(8):e05377.

[196] OECD (1995). *Test No. 107: Partition Coefficient (n-octanol/water): Shake Flask Method*. OECD.

[197] OECD (2010). *Test Guideline 417: Toxicokinetics OECD Guidelines for the Testing of Chemicals*. OECD.

[198] OECD (2022). *Test No. 123: Partition Coefficient (1-Octanol/Water): Slow-Stirring Method*. OECD.

[199] Olker, J. H., Elonen, C. M., Pilli, A., Anderson, A., Kinziger, B., Erickson, S., Skopinski, M., Pomplun, A., LaLone, C. A., Russom, C. L., et al. (2022). The ECOTOXicology knowledgebase: A curated database of ecologically relevant toxicity tests to support environmental research and risk assessment. *Environmental Toxicology and Chemistry*, 41(6):1520–1539.

[200] Osman, N. A. (2023). Statistical methods for in silico tools used for risk assessment and toxicology. *Physical Sciences Reviews*, 8(9):2711–2724.

[201] Oss, M., Kruve, A., Herodes, K., and Leito, I. (2010). Electrospray ionization efficiency scale of organic compounds. *Analytical Chemistry*, 82(7):2865–2872.

[202] Oss, M., Tshepelevitsh, S., Kruve, A., Liigand, P., Liigand, J., Rebane, R., Selberg, S., Ets, K., Herodes, K., and Leito, I. (2021). Quantitative electrospray ionization efficiency scale: 10 years after. *Rapid Communications in Mass Spectrometry*, 35(21):e9178.

[203] Palm, E. and Kruve, A. (2022). Machine learning for absolute quantification of unidentified compounds in non-targeted LC/HRMS. *Molecules*, 27(3):1013.

[204] Panagopoulos Abrahamsson, D., Park, J.-S., Singh, R. R., Sirota, M., and Woodruff, T. J. (2020). Applications of machine learning to in silico quantification of chemicals without analytical standards. *Journal of chemical information and modeling*, 60(6):2718–2727.

[205] Pantic, I., Paunovic, J., Cumic, J., Valjarevic, S., Petroianu, G. A., and Corridon, P. R. (2023). Artificial neural networks in contemporary toxicology research. *Chemico-Biological Interactions*, 369:110269.

[206] Paul, R., Zeis, B., Lamkemeyer, T., Seidl, M., and Pirow, R. (2004). Control of oxygen transport in the microcrustacean Daphnia: regulation of haemoglobin expression as central mechanism of adaptation to different oxygen and temperature conditions. *Acta Physiologica Scandinavica*, 182(3):259–275.

[207] Petitpas, I., Bhattacharya, A. A., Twine, S., East, M., and Curry, S. (2001). Crystal structure analysis of warfarin binding to human serum albumin: anatomy of drug site I. *Journal of Biological Chemistry*, 276(25):22804–22809.

[208] Petrauskas, A. A. and Kolovanov, E. A. (2000). ACD/Log P method description. *Perspectives in drug discovery and design*, 19:99–116.

[209] Pieke, E. N., Granby, K., Trier, X., and Smedsgaard, J. (2017). A framework to estimate concentrations of potentially unknown substances by semi-quantification in liquid chromatography electrospray ionization mass spectrometry. *Analytica chimica acta*, 975:30–41.

[210] Pirow, R., Wollinger, F., and Paul, R. (1999). The importance of the feeding current for oxygen uptake in the water flea Daphnia magna. *Journal of experimental biology*, 202(5):553–562.

[211] Pletz, J., Blakeman, S., Paini, A., Parissis, N., Worth, A., Andersson, A.-M., Frederiksen, H., Sakhi, A. K., Thomsen, C., and Bopp, S. K. (2020). Physiologically based kinetic (PBK) modelling and human biomonitoring data for mixture risk assessment. *Environment International*, 143:105978.

[212] Preuss, T. G., Telscher, M., and Ratte, H. T. (2008). Life stage-dependent bioconcentration of a nonylphenol isomer in Daphnia magna. *Environmental pollution*, 156(3):1211–1217.

[213] Przybylak, K., Madden, J., Covey-Crump, E., Gibson, L., Barber, C., Patel, M., and Cronin, M. (2018). Characterisation of data resources for in silico modelling: benchmark datasets for ADME properties. *Expert Opinion on Drug Metabolism & Toxicology*, 14(2):169–181.

[214] Raies, A. B. and Bajic, V. B. (2016). In silico toxicology: computational methods for the prediction of chemical toxicity. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 6(2):147–172.

[215] Raths, J., Schinz, L., Mangold-Döring, A., and Hollender, J. (2023). Elimination resistance: characterizing multi-compartment toxicokinetics of the neonicotinoid thia-

cloprid in the amphipod Gammarus pulex using bioconcentration and receptor-binding assays. *Environmental Science & Technology*, 57(24):8890–8901.

[216] Ratier, A. and Charles, S. (2022). Accumulation-depuration data collection in support of toxicokinetic modelling. *Scientific Data*, 9(1):130.

[217] Ratier, A., Lopes, C., and Charles, S. (2022a). Improvements in estimating bioaccumulation metrics in the light of toxicokinetic models and Bayesian inference. *Archives of Environmental Contamination and Toxicology*, 83(4):339–348.

[218] Ratier, A., Lopes, C., Labadie, P., Budzinski, H., Delorme, N., Queau, H., Peluhet, L., Geffard, O., and Babut, M. (2019). A Bayesian framework for estimating parameters of a generic toxicokinetic model for the bioaccumulation of organic chemicals by benthic invertebrates: Proof of concept with PCB153 and two freshwater species. *Ecotoxicology and Environmental Safety*, 180:33–42.

[219] Ratier, A., Lopes, C., Multari, G., Mazerolles, V., Carpentier, P., and Charles, S. (2022b). New perspectives on the calculation of bioaccumulation metrics for active substances in living organisms. *Integrated Environmental Assessment and Management*, 18(1):10–18.

[220] Raunio, H. (2011). In silico toxicology–non-testing methods. *Frontiers in pharmacology*, 2:9488.

[221] Reale, E., Jeddi, M. Z., Paini, A., Connolly, A., Duca, R., Cubadda, F., Benfenati, E., Bessems, J., Galea, K. S., Dirven, H., et al. (2024). Human biomonitoring and toxicokinetics as key building blocks for next generation risk assessment. *Environment International*, 184:108474.

[222] Reijenga, J., Van Hoof, A., Van Loon, A., and Teunissen, B. (2013). Development of methods for the determination of pKa values. *Analytical chemistry insights*, 8:ACI–S12304.

[223] Rendal, C., Kusk, K. O., and Trapp, S. (2011a). Optimal choice of pH for toxicity and bioaccumulation studies of ionizing organic chemicals. *Environmental Toxicology and Chemistry*, 30(11):2395–2406.

[224] Rendal, C., Kusk, K. O., and Trapp, S. (2011b). The effect of pH on the uptake and toxicity of the bivalent weak base chloroquine tested on Salix viminalis and Daphnia magna. *Environmental Toxicology and Chemistry*, 30(2):354–359.

[225] Rohatgi, A. (2024). WebPlotDigitizer. https://automeris.io/WebPlotDigitizer.html. Accessed 2023.

[226] Roy, K. and Kabir, H. (2012). QSPR with extended topochemical atom (ETA) indices, 3: modeling of critical micelle concentration of cationic surfactants. *Chemical engineering science*, 81:169–178.

[227] Rubach, M. N., Ashauer, R., Maund, S. J., Baird, D. J., and Van den Brink, P. J. (2010). Toxicokinetic variation in 15 freshwater arthropod species exposed to the insecticide chlorpyrifos. *Environmental Toxicology and Chemistry*, 29(10):2225–2234.

[228] Saavedra, L. M., Martinez, J. C. G., and Duchowicz, P. R. (2024). Advances of the QSAR approach as an alternative strategy in the environmental risk assessment. In *QSAR in safety evaluation and risk assessment*, pages 117–137. Elsevier.

[229] Sand, S., Parham, F., Portier, C. J., Tice, R. R., and Krewski, D. (2017). Comparison of points of departure for health risk assessment based on high-throughput screening data. *Environmental health perspectives*, 125(4):623–633.

[230] Sanders, H. O., Huckins, J., Johnson, B. T., and Skaar, D. (1981). Biological effects of kepone and mirex in freshwater invertebrates. *Archives of environmental contamination and toxicology*, 10:531–539.

[231] Schmitt, W., Bruns, E., Dollinger, M., and Sowig, P. (2013). Mechanistic TK/TD-

model simulating the effect of growth inhibitors on Lemna populations. *Ecological modelling*, 255:1–10.

[232] Schollée, J. E., Schymanski, E. L., Stravs, M. A., Gulde, R., Thomaidis, N. S., and Hollender, J. (2017). Similarity of high-resolution tandem mass spectrometry spectra of structurally related micropollutants and transformation products. *Journal of the American Society for Mass Spectrometry*, 28(12):2692–2704.

[233] Scholz, S., Nichols, J. W., Escher, B. I., Ankley, G. T., Altenburger, R., Blackwell, B., Brack, W., Burkhard, L., Collette, T. W., Doering, J. A., et al. (2022). The Eco-Exposome concept: Supporting an integrated assessment of mixtures of environmental chemicals. *Environmental toxicology and chemistry*, 41(1):30–45.

[234] Schonlau, M. and Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal*, 20(1):3–29.

[235] Schuler, L. J., Wheeler, M., Bailer, A. J., and Lydy, M. J. (2003). Toxicokinetics of sediment-sorbed benzo [a] pyrene and hexachlorobiphenyl using the freshwater invertebrates Hyalella azteca, Chironomus tentans, and Lumbriculus variegatus. *Environmental Toxicology and Chemistry: An International Journal*, 22(2):439–449.

[236] Schultz, T., Amcoff, P., Berggren, E., Gautier, F., Klaric, M., Knight, D., Mahony, C., Schwarz, M., White, A., and Cronin, M. (2015). A strategy for structuring and reporting a read-across prediction of toxicity. *Regulatory Toxicology and Pharmacology*, 72(3):586–601.

[237] Sekula, B. and Bujacz, A. (2016). Structural insights into the competitive binding of diclofenac and naproxen by equine serum albumin. *Journal of medicinal chemistry*, 59(1):82–89.

[238] Solomon, K. R., Baker, D. B., Richards, R. P., Dixon, K. R., Klaine, S. J., La Point, T. W., Kendall, R. J., Weisskopf, C. P., Giddings, J. M., Giesy, J. P., et al. (1996).

Ecological risk assessment of atrazine in North American surface waters. *Environmental Toxicology and Chemistry: An International Journal*, 15(1):31–76.

[239] Solomon, K. R., Wilks, M. F., Bachman, A., Boobis, A., Moretto, A., Pastoor, T. P., Phillips, R., and Embry, M. R. (2016). Problem formulation for risk assessment of combined exposures to chemicals and other stressors in humans. *Critical Reviews in Toxicology*, 46(10):835–844.

[240] Spacie, A. and Hamelink, J. L. (1982). Alternative models for describing the bioconcentration of organics in fish. *Environmental Toxicology and Chemistry: An International Journal*, 1(4):309–320.

[241] Stollewerk, A. (2010). The water flea Daphnia-a'new'model system for ecology and evolution? *Journal of biology*, 9:1–4.

[242] Su, H., Zhang, Q., Huang, K., Wang, W.-X., Li, H., Huang, Z., Cheng, F., and You, J. (2023). Two-Compartmental Toxicokinetic Model Predicts Interspecies Sensitivity Variation of Imidacloprid to Aquatic Invertebrates. *Environmental Science & Technology*, 57(29):10532–10541.

[243] Szucs, D. and Ioannidis, J. P. (2017). When null hypothesis significance testing is unsuitable for research: a reassessment. *Frontiers in human neuroscience*, 11:390.

[244] Tan, Y.-M., Barton, H. A., Boobis, A., Brunner, R., Clewell, H., Cope, R., Dawson, J., Domoradzki, J., Egeghy, P., Gulati, P., et al. (2021). Opportunities and challenges related to saturation of toxicokinetic processes: implications for risk assessment. *Regulatory Toxicology and Pharmacology*, 127:105070.

[245] Tarr, B. D., Barron, M. G., and Hayton, W. L. (1990). Effect of body size on the uptake and bioconcentration of di-2-ethylhexyl phthalate in rainbow trout. *Environmental Toxicology and Chemistry: An International Journal*, 9(8):989–995.

[246] Tatarazako, N. and Oda, S. (2007). The water flea Daphnia magna (Crustacea, Cladocera) as a test species for screening and evaluation of chemicals with endocrine disrupting effects on crustaceans. *Ecotoxicology*, 16:197–203.

[247] Testa, B., Crivori, P., Reist, M., and Carrupt, P.-A. (2000). The influence of lipophilicity on the pharmacokinetic behavior of drugs: Concepts and examples. *Perspectives in Drug Discovery and Design*, 19:179–211.

[248] TGD, E. (2003). Technical guidance document on risk assessment in support of commission directive 93/67/EEC on risk assessment for new notified substances, Commission Regulation (EC) No 1488/94 on Risk Assessment for existing substances, and Directive 98/8/EC of the European Parliament and of the Council concerning the placing of biocidal products on the market. *Part I–IV, European Chemicals Bureau (ECB), JRC-Ispra (VA), Italy.*

[249] Thakare, R., Chhonker, Y. S., Gautam, N., Alamoudi, J. A., and Alnouti, Y. (2016). Quantitative analysis of endogenous compounds. *Journal of pharmaceutical and biomedical analysis*, 128:426–437.

[250] Tkaczyk, A., Bownik, A., Dudka, J., Kowal, K., and Ślaska, B. (2021). Daphnia magna model in the toxicity assessment of pharmaceuticals: A review. *Science of the Total Environment*, 763:143038.

[251] Törnqvist, E., Annas, A., Granath, B., Jalkesten, E., Cotgreave, I., and Öberg, M. (2014). Strategic focus on 3R principles reveals major reductions in the use of animals in pharmaceutical toxicity testing. *PloS one*, 9(7):e101638.

[252] Tseng, R.-L. and Wu, F.-C. (2008). Inferring the favorable adsorption level and the concurrent multi-stage process with the Freundlich constant. *Journal of hazardous materials*, 155(1-2):277–287.

[253] US EPA. CompTox Chemicals Dashboard. https://comptox.epa.gov/dashboard/. Accessed Sept 2023.

[254] US EPA. ECOTOX Database. https://cfpub.epa.gov/ecotox/search.cfm. Accessed Sept 2023.

[255] Van der Oost, R., Beyer, J., and Vermeulen, N. P. (2003). Fish bioaccumulation and biomarkers in environmental risk assessment: a review. *Environmental toxicology and pharmacology*, 13(2):57–149.

[256] van Dijk, J., Gustavsson, M., Dekker, S. C., and van Wezel, A. P. (2021). Towards 'one substance–one assessment': An analysis of EU chemical registration and aquatic risk assessment frameworks. *Journal of environmental management*, 280:111692.

[257] Van Leeuwen, S. P., Kärrman, A., Van Bavel, B., De Boer, J., and Lindström, G. (2006). Struggle for quality in determination of perfluorinated contaminants in environmental and human samples. *Environmental science & technology*, 40(24):7854–7860.

[258] Van Ravenzwaaij, D., Cassey, P., and Brown, S. D. (2018). A simple introduction to Markov Chain Monte–Carlo sampling. *Psychonomic bulletin & review*, 25(1):143–154.

[259] Veith, G. D., DeFoe, D. L., and Bergstedt, B. V. (1979). Measuring and estimating the bioconcentration factor of chemicals in fish. *Journal of the Fisheries Board of Canada*, 36(9):1040–1048.

[260] Viant, M. R., Ebbels, T. M., Beger, R. D., Ekman, D. R., Epps, D. J., Kamp, H., Leonards, P. E., Loizou, G. D., MacRae, J. I., van Ravenzwaay, B., et al. (2019). Use cases, best practice and reporting standards for metabolomics in regulatory toxicology. *Nature communications*, 10(1):3041.

[261] Vuignier, K., Schappler, J., Veuthey, J.-L., Carrupt, P.-A., and Martel, S. (2010). Drug–protein binding: a critical review of analytical tools. *Analytical and bioanalytical chemistry*, 398:53–66.

[262] Wambaugh, J. (2019). New Approach Methodologies for Chemical Risk Assessment. Presentation.

[263] Wambaugh, J. F., Bare, J. C., Carignan, C. C., Dionisio, K. L., Dodson, R. E., Jolliet, O., Liu, X., Meyer, D. E., Newton, S. R., Phillips, K. A., et al. (2019). New approach methodologies for exposure science. *Current Opinion in Toxicology*, 15:76–92.

[264] Wambaugh, J. F., Wetmore, B. A., Pearce, R., Strope, C., Goldsmith, R., Sluka, J. P., Sedykh, A., Tropsha, A., Bosgra, S., Shah, I., et al. (2015). Toxicokinetic triage for environmental chemicals. *Toxicological Sciences*, 147(1):55–67.

[265] Wanat, K. (2020). Biological barriers, and the influence of protein binding on the passage of drugs across them. *Molecular Biology Reports*, 47(4):3221–3231.

[266] Wang, J., Nolte, T. M., Owen, S. F., Beaudouin, R., Hendriks, A. J., and Ragas, A. M. (2022). A generalized physiologically based kinetic model for fish for environmental risk assessment of pharmaceuticals. *Environmental Science & Technology*, 56(10):6500–6510.

[267] Wang, Z., Walker, G. W., Muir, D. C., and Nagatani-Yoshida, K. (2020). Toward a global understanding of chemical pollution: a first comprehensive analysis of national and regional chemical inventories. *Environmental science & technology*, 54(5):2575–2584.

[268] Wenlock, M. C. and Barton, P. (2013). In silico physicochemical parameter predictions. *Molecular pharmaceutics*, 10(4):1224–1235.

[269] Wojtyniak, J.-G., Britz, H., Selzer, D., Schwab, M., and Lehr, T. (2020). Data digitizing: accurate and precise data extraction for quantitative systems pharmacology and physiologically-based pharmacokinetic modeling. *CPT: Pharmacometrics & Systems Pharmacology*, 9(6):322–331.

[270] Wu, Y. and Li, L. (2016). Sample normalization methods in quantitative metabolomics. *Journal of Chromatography A*, 1430:80–95.

[271] Xia, X., Rabearisoa, A. H., Jiang, X., and Dai, Z. (2013). Bioaccumulation of perfluoroalkyl substances by Daphnia magna in water with different types and concentrations of protein. *Environmental science & technology*, 47(19):10955–10963.

[272] Yamazaki, K. and Kanaoka, M. (2004). Computational prediction of the plasma protein-binding percent of diverse pharmaceutical compounds. *Journal of pharmaceutical sciences*, 93(6):1480–1494.

[273] Yap, C. W. (2011). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of computational chemistry*, 32(7):1466–1474.

[274] Yu, S., Gao, S., Gan, Y., Zhang, Y., Ruan, X., Wang, Y., Yang, L., and Shi, J. (2016). QSAR models for predicting octanol/water and organic carbon/water partition coefficients of polychlorinated biphenyls. *SAR and QSAR in Environmental Research*, 27(4):249–263.

[275] Yun, Y. E., Tornero-Velez, R., Purucker, S. T., Chang, D. T., and Edginton, A. N. (2021). Evaluation of quantitative structure property relationship algorithms for predicting plasma protein binding in humans. *Computational Toxicology*, 17:100142.

[276] Zhang, F., Bartels, M., Clark, A., Erskine, T., Auernhammer, T., Bhhatarai, B., Wilson, D., and Marty, S. (2018). Performance evaluation of the GastroPlusTM software tool for prediction of the toxicokinetic parameters of chemicals. *SAR and QSAR in Environmental Research*, 29(11):875–893.

[277] Zhao, S., Jones, K. C., and Sweetman, A. J. (2018). Can poly-parameter linear-free energy relationships (pp-LFERs) improve modelling bioaccumulation in fish? *Chemosphere*, 191:235–244.

[278] Zhu, M., Chen, J., Peijnenburg, W. J., Xie, H., Wang, Z., and Zhang, S. (2023). Controlling factors and toxicokinetic modeling of antibiotics bioaccumulation in aquatic organisms: A review. *Critical Reviews in Environmental Science and Technology*, 53(15):1431–1451.