Computationally Linking Chemical Exposure to Molecular Effects with Complex Data

Comparing Methods to Disentangle Chemical Drivers in Environmental Mixtures and Knowledge-based Deep Learning for Predictions in Environmental Toxicology

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

DISSERTATION

zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALIUM
(Dr. rer. nat.)

im Fachgebiet
Informatik

vorgelegt
von Master of Science *Stefan Krämer*

geboren am 15.03.1990 in Karl-Marx-Stadt

Die Annahme der Dissertation wurde empfohlen von:

1. Professor Dr. Jörg Hackermüller (Universität Leipzig, Deutschland)
2. Professor Dr. Sven Nahnsen (Universität Tübingen, Deutschland)

Die Verleihung des akademischen Grades erfolgt mit Bestehen der Verteidigung am 19.05.2022 mit dem Gesamtprädikat *magna cum laude*.

Contents

Table of Contents **I**

Abstract **V**

Acknowledgements **VII**

Prelude **IX**

1 Introduction **1**

1.1 An overview of environmental toxicology 2

1.1.1 Environmental toxicology 2

1.1.2 Chemicals in the environment 4

1.1.3 Systems biological perspectives in environmental toxicology 7

1.2 Computational toxicology 11

1.2.1 Omics-based approaches 12

1.2.2 Linking chemical exposure to transcriptional effects 14

1.2.3 Up-scaling from the gene level to higher biological organisation levels . 19

1.2.4 Biomedical literature-based discovery 24

1.2.5 Deep learning with knowledge representation 27

1.3 Research question and approaches 29

2 Methods and Data **33**

2.1 Linking environmental relevant mixture exposures to transcriptional effects . 34
2.1.1 Exposure and microarray data 34
2.1.2 Preprocessing 35
2.1.3 Differential gene expression 37
2.1.4 Association rule mining 38
2.1.5 Weighted gene correlation network analysis 39
2.1.6 Method comparison 41
2.2 Predicting exposure-related effects on molecular level 44
2.2.1 Input 44
2.2.2 Input preparation 47
2.2.3 Deep learning models 49
2.2.4 Toxicogenomic application 54

3 Method comparison to link complex stream water exposures to effects on the transcriptional level 57

3.1 Background and motivation 58
3.1.1 Workflow 61
3.2 Results 62
3.2.1 Data preprocessing 62
3.2.2 Differential gene expression analysis 67
3.2.3 Association rule mining 71
3.2.4 Network inference 78
3.2.5 Method comparison 84
3.2.6 Application case of method integration 87
3.3 Discussion 91
3.4 Conclusion 99

4 Deep learning prediction of chemical-biomolecule interactions 101

4.1 Motivation 102

4.1.1 Workflow . . . 105
4.2 Results . . . 107
4.2.1 Input preparation . . . 107
4.2.2 Model selection . . . 110
4.2.3 Model comparison . . . 118
4.2.4 Toxicogenomic application . . . 121
4.2.5 Horizontal augmentation without tail-padding . . . 123
4.2.6 Four-class problem formulation . . . 124
4.2.7 Training with CTD data . . . 125
4.3 Discussion . . . 129
4.3.1 Transferring biomedical knowledge towards toxicology. . . . 129
4.3.2 Deep learning with biomedical knowledge representation . . . 133
4.3.3 Data integration . . . 136
4.4 Conclusion . . . 141

5 Conclusion and Future perspectives **143**
5.1 Conclusion . . . 144
5.1.1 Investigating complex mixtures in the environment . . . 144
5.1.2 Complex knowledge from literature and curated databases predict chemical-biomolecule interactions . . . 145
5.1.3 Linking chemical exposure to biological effects by integrating CTD . . . 146
5.2 Future perspectives . . . 147

S1 Supplement Chapter 1 **153**
S1.1 Example of an estrogen bioassay . . . 154
S1.2 Types of mode of action . . . 154
S1.3 The dogma of molecular biology. . . . 157
S1.4 Transcriptomics . . . 159

S2 Supplement Chapter 3 **161**

S3 Supplement Chapter 4 **175**

S3.1 Hyperparameter tuning results 176

S3.2 Functional enrichment with predicted chemical-gene interactions and CTD reference pathway genesets 179

S3.3 Reduction of learning rate in a model with large word embedding vectors. . . 183

S3.4 Horizontal augmentation without tail-padding 183

S3.5 Four-relationship classification 185

S3.6 Interpreting loss observations for SemMedDB trained models. 187

List of Abbreviations **i**

List of Figures **vi**

List of Tables **x**

Bibliography **xii**

Curriculum scientiae **xxxix**

Selbständigkeitserklärung **xliii**

Abstract

Chemical exposures affect the environment and may lead to adverse outcomes in its organisms. Omics-based approaches, like standardised microarray experiments, have expanded the toolbox to monitor the distribution of chemicals and assess the risk to organisms in the environment. The resulting complex data have extended the scope of toxicological knowledge bases and published literature. A plethora of computational approaches has been applied in environmental toxicology considering systems biology and data integration. Still, the complexity of environmental and biological systems given in data challenge investigations of exposure-related effects. This thesis aimed at computationally linking chemical exposure to biological effects on the molecular level considering sources of complex environmental data.

The first study employed data of an omics-based exposure study considering mixture effects in a freshwater environment. We compared three data-driven analyses in their suitability to disentangle mixture effects of chemical exposures to biological effects and their reliability in attributing potentially adverse outcomes to chemical drivers with toxicological databases on gene and pathway levels. *Differential gene expression analysis* and a *network inference* approach resulted in toxicologically meaningful outcomes and uncovered individual chemical effects — stand-alone and in combination. We developed an integrative computational strategy to harvest exposure-related gene associations from environmental samples considering mixtures of lowly concentrated compounds. The applied approaches allowed assessing the hazard of chemicals more systematically with correlation-based compound groups.

This dissertation presents another achievement towards a data-driven hypothesis generation for molecular exposure effects. The approach combined text-mining and deep learning. The study was entirely data-driven and involved state-of-the-art computational methods of artificial intelligence. We employed literature-based relational data and curated toxicological knowledge to predict chemical-biomolecule interactions. A word embedding neural network with a subsequent feed-forward network was implemented. Data augmentation and recurrent neural networks were beneficial for training with curated toxicological knowledge. The trained models reached accuracies of up to 94% for unseen test data of the employed knowledge base. However, we could not reliably confirm known chemical-gene interactions across selected data sources. Still, the predictive models might derive unknown information from toxicological knowledge sources, like literature, databases or omics-based exposure studies. Thus, the deep learning models might allow predicting hypotheses of exposure-related molecular effects.

Both achievements of this dissertation might support the prioritisation of chemicals for testing and an intelligent selection of chemicals for monitoring in future exposure studies.

Acknowledgements

First, I want to thank my supervisors Peter F. Stadler, Jörg Hackermüller and Jana Schor, for their constructive criticism, constant support, motivation, belief in my lifelong learning and the chance to work in an interdisciplinary and sustainable research field.

I want to express my gratitude to my collaboration partners in St.Paul for their constant support over the last three years, their supervision during my stay in February 2019 and the regular research meetings in 2021. I thank Dalma Martinović-Weigelt for providing data for the TenStreams project. In particular, I am grateful for your hospitality and for giving me the possibility to widen my scientific network. I owe thanks to Chih Lai for teaching and supporting me on my hands-on journey towards Machine Learning and Artificial Intelligence.

I want to thank all my colleagues in the PhD-Colleg *Proxies of the Eco-exposome*. Thanks to the other PhD candidates Gianina Jakobs, Theo Wernicke and Janek-Paul Dann, for scientific exchanges on the UFZ-campus and refreshing moments with beer and wine outside of work. I owe a great thanks to my mentor Wibke Busch for your scientific support, getting new perspectives, an always open door and your mental encouragement.

Next, I would like to thank all my colleagues at the formerly 'Young Investigators Group Bioinformatics & Transcriptomics' and new built Department of 'Computational Biology', for their scientific advice and support. Especially to Lisa Steinheuer, with whom I shared the most encouraging and the most struggling moments in and outside the office since our starting days of the PhD in 2018. Furthermore, to Stephan Schreiber, Sebastian Canzler, Matthias Bernt, Ali Yazbeck, Andreas Schüttler, Paul Michaelis and Kyriakos Soulios.

Last but not least, I would like to thank my friends and family (In particular, I thank you all for being patient with my moodiness). Special thanks to Anna for cheering me up when needed, the plethora of opportunities to break off the workday life on the dance floor or in the kitchen, opera and spa. Likewise, I am deeply grateful for having my supportive flatmate and friend Jessika, with whom I successfully weathered lockdowns and who kept the apartment clean (and blamed Veronika, but not me). Finally, I want to express my gratitude to my parents, who allowed, understood and supported me going my own way, not only in my career.

The following publication was related to the recent work:

Lai, C., Martinović-Weigelt, D., Serrao De Filippo, A., Krämer, Stefan and Poschen, C. (2021). **Extracting Semantics of Predicates From Millions of Bio-Medical Abstracts for Inferencing New Biological Key Events and Relationships.** The 5th International Workshop on Deep Learning in Bioinformatics, Biomedicine, and Healthcare Informatics (Accepted on 21.10.2021).

Prelude: Research scope

Organisms are exposed to exogenous chemicals that lead to endogenous processes altering or producing biochemicals (see figure 1) [Escher et al. 2017]. Exposure to exogenous and endogenous chemicals lead to molecular alterations potentially inducing cellular toxicity pathways or effects on higher levels of biological organisation. In a human health context, the exposome encompasses the entire environmental exposure in an individual's lifetime [Wild 2012, Miller and Jones 2014]. Identifying chemical effects on the cellular level may help integrate toxicity pathways to adverse health outcomes and ecosystem-level effects. The PhD-colleg *Proxies of the eco-exposome* aimed to investigate external and internal exposures and link them to molecular biological effects in aquatic species to bridge the concept to environmental health issues. Thus, the exposome narrative was transformed to environmental toxicological questions considering the *eco-exposome* [Escher et al. 2017, Lioy and Smith 2013, Scholz et al. 2021] focusing on the aquatic environment and chosen proxies of aquatic model organisms. Integrating data of exposure and molecular response is at the core of the eco-exposome concept. However, computational methods and integration approaches with available knowledge were lacking or had not been sufficiently evaluated for this purpose so far. In this thesis, we * presented the project's bioinformatics and data scientific investigations to associate chemical exposure to biological effects computationally.

Chapter 1 described the essential concepts and toxicological approaches to understand the scientific background encompassing multiple domains in (environmental and computational) toxicology. The chapter resulted in the motivation of this dissertation.
In **Chapter 2**, we comprised the applied methods with a more general technical introduction and implementation and described the selected data sets.
In **Chapter 3**, we described the evaluation of methods for linking environmental relevant chemical exposures to transcriptional effects with respect to comparing them based on their applicability and biological reliability. The identified chemical-gene interactions were compared to external references to validate the biological meaning and chemical representation on gene and pathway levels. The two strategies of differential gene expression analysis and network inference resulted in biologically meaningful results. As stand-alone and in an integrative strategy, they disentangled chemical stressors with reliable endocrine disruptive effects, especially when considering an exposure scenario of correlated compound groups.

* Research projects are performed in teams with multiple experts and heads with various ideas, suggestions and questions. This also influenced and formed my work during the last three years, which built the base of my dissertation. Therefore, it seemed inappropriate to present the processes, results and opinions in this work with *'I'*. Consequently, here and in the following the personal pronoun *'we'* was used.

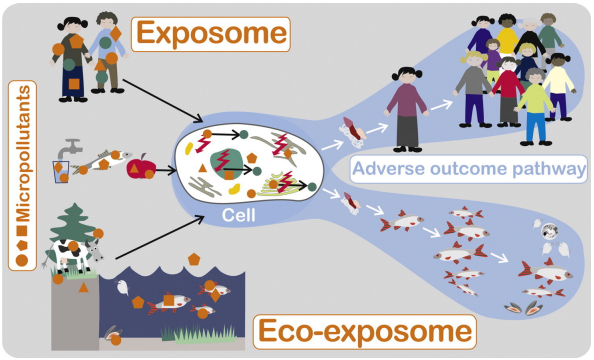

Figure 1. ***Transfer of the exposome definition to environmental toxicology.*** *An exogenous chemical exposure affects cellular processes and lead to adverse health outcomes on higher levels of biological organisation like the organism, population or even eco-system. Figure taken from Escher et al. [2017].*

In **Chapter 4**, knowledge bases were considered sources to predict exposure-associated links to biomolecules. The presented study was focused on input preparation, deep learning model selection and model evaluation. With deep learning and data integration, the presented strategy was a preliminary step towards generating potentially new toxicological hypotheses of unknown molecular key events. We showed that text-based data and toxicological databases could be employed to train a deep learning model to predict chemical-biomolecule interactions with 70% and 94% test accuracy when considering a unifying biomedical terminology.

Chapter 5 discussed the dissertation's achievements in the context of recent environmental and computational toxicology and concluded this thesis. It set the achievements of both here presented studies in a broader context of exposure studies and presented future perspectives to hazard assessment and biomonitoring, which could be also helpful when investigating *Proxies of the Eco-exposome.*

Chapter 1

Introduction

This chapter presents fundamental concepts and approaches in environmental toxicology. In this thesis, we focus on the investigation of exposure-related biological effects and toxicogenomics. First, we introduce environmental toxicology and its general aim of investigating chemicals in the environment. Furthermore, we present helpful systems biological concepts for environmental toxicology. This leads directly towards the next section of omics-based approaches transitioning towards toxicogenomics. We elaborate on recent work which links chemical exposure to transcriptional effects. Furthermore, we present data integration in the context of environmental toxicology, which leads directly to an elaboration of recent developments of adverse outcome pathways and how mechanistic knowledge can be used for predictive computational approaches. In this respect, we shed light on literature-based discovery and deep learning approaches in toxicology related areas. Finally, the chapter motivates the presented research of this dissertation and presents objective and applied approaches of the thesis.

1.1 An overview of environmental toxicology

A plethora of anthropogenic sources from industry, agriculture, and households release chemicals into the environment. Such pollutions in air, soil, or water may lead to toxic effects and adversely affect the environment on different levels of biological organisation. Moreover, the toxic exposures of exogenous chemical compounds burden the individual beings and potentially induce an exposure with endogenous chemicals based on molecular responses [Escher et al. 2017]. Thus, chemical interaction with biomolecules affects different biochemical processes and may affect a toxic effect on a cellular level. Consequently, cascades of biological processes on different levels of biological organisation lead to sub-lethal or even lethal adverse effects. For a general understanding of the research domain, the field of environmental toxicology is introduced in the following.

1.1.1 Environmental toxicology

The multidisciplinary research field of environmental toxicology (ET) has materialised in the mid 20th century due to an upcoming awareness of chemical emissions to the environment and their effects on organisms, including humans [Carson 2002]. Thus, whereas ecotoxicology is restricted to ecological endpoints, ET includes human health as an endpoint [Ragas, Ad 2021]. In ET, potentially hazardous chemicals in the environment are investigated due to their fate (environmental chemistry), their effects on living organisms (toxicology), and their impact on higher levels of biological organisation (ecology) (see figure 1.1).

The field of **environmental chemistry** assesses environmental exposures using two main approaches - chemical analytics and mathematical modelling. Chemical analytics discovers exposure patterns by measuring emissions, concentrations and behaviour of processes like biodegradation of chemicals of concern (see section 1.1.2). Moreover, environmental chemists may detect spatio-temporal or distributional exposure patterns and investigate them with the help of mathematical modelling and prediction approaches useful for environmental risk assessment.

The field of **toxicology** focuses on the chemical interaction with an organism to understand the toxicity mechanisms of specific chemical exposures in an organism and identify the exposures leading to (sub-)lethal effects. Toxicological research relies on the dose concept of Paracelsus [1965], defining that everything can be toxic and lead to adverse effects, but that it is dependent on the chemical concentration. Toxicologists investigate effects on individuals for exposures with single compounds, artificial mixtures, or environmental samples of soil, water or sediments collected from sites of interest. The field of toxicology considers two main

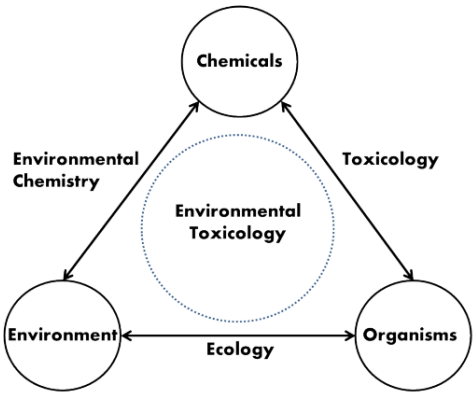

Figure 1.1. ***The triangle of environmental toxicology highlights the interplay of chemicals, the environment and organisms.*** *Environmental toxicology investigates the environmental responses of chemical emissions. The chemical fate and potentially hazardous effects are studied by their interactions with other chemicals, the environment and the organisms. The figure is taken from [Ragas, Ad 2021].*

perspectives. The *toxicokinetic* approaches study the effects of internal doses at the site of toxic actions and thus the fate of hazardous compounds in an organism. The *toxicodynamic* approaches focus on the system biological chains of events from biomolecular interaction of a chemical stressor up to an adverse outcome.

In toxicological laboratory experiments, researchers study model organisms under fixed conditions. The application of *in-vitro* measurements and *in-silico* predictions is an alternative to reduce animal testing and expand the investigation of natural environmental systems. In the effect-driven chemical analytics (see section 1.1.2), *in-vitro* test batteries of bioanalytical assays are developed to identify the risk or hazard of exposures with a small set of molecular targets for specific modes of toxic actions or adverse outcomes (see section 1.1.3). Computational and *in-silico* approaches have become important in ET with the advent of omics techniques in genetic research generating high-throughput and high-content data for systems biological investigations of biological effects on a molecular level (see section 1.2.1). In that respect, computational toxicologists curate toxicological databases [e.g. Mattingly et al. 2003, Kuhn et al. 2008] and develop data and method integrative approaches (see section 1.2.3).

Ecology focuses on interactions within the environment on organism and population levels. Especially for environmental hazard assessment, such knowledge is necessary to transfer exposure effects from individuals to the ecosystem, but it is not elaborated in this thesis in more detail. Nevertheless, sub-lethal effects on the molecular up to the individual level are relevant and can induce fatal perturbations at the ecosystem level [e.g. Kidd et al. 2007].

This thesis concentrates on the molecular interactions of chemical exposures. Therefore, various computational approaches are applied. The following section motivates why to investigate chemicals in the environment, some recent challenges, and how environmental toxicologists recently investigated toxic effects on the environment.

1.1.2 Chemicals in the environment

Frequently studied contaminants in ET are metals, organic chemicals, radioactive elements and increased nutrient concentrations, e.g., nitrates and nitrites. Environmental toxicologists classify pollutions based on release traces, chemical properties, effects on a biological entity or their use by humankind (see figure 1.2).

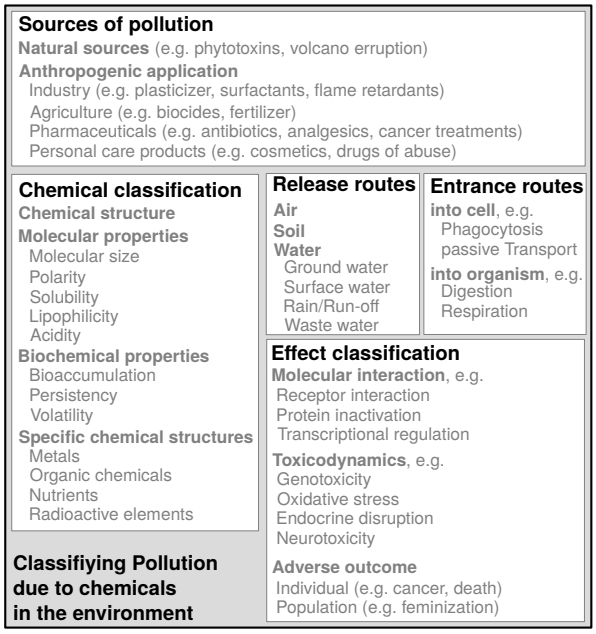

Figure 1.2. ***Classifying chemical pollution in the environment.*** *Based on the research context, chemical exposure and chemicals in the environment are grouped variously.*

For example, the sources of pollution are divided into natural — e.g. heavy metal hotspots in the earth crust, or bacterial and fungal toxins — and human-caused (*anthropogenic*) — e.g. wastewater from households or industrial sites. The majority of ET research focuses on anthropogenic pollutions. Humankind uses synthetically produced organic chemicals in the economic areas of industry, agriculture, infrastructure and households. Distinct anthropogenic origins of an environment perturbing exposure are the waste disposals and emissions into the air, the water and the soil. This dissertation was part of the PhD-colleg *Proxies of the eco-*

exposome, which focused on the adverse effects of anthropogenic pollution on the aquatic environment.

Mixture toxicity. Under environmental conditions, an organism is in general exposed to mixtures of toxicants. Chemical compounds in an environmental sample may interact through (physico-)chemical or physiological interactions. A primary challenge in ET is understanding the toxic effects of chemical mixtures in the environment [Cedergreen 2014, Kortenkamp and Faust 2018]. For example, one compound can also affect the internal xenobiotic metabolism of other compounds, or indirect interactions at the target site are possible. Interactions affect the bioavailability, the toxicokinetics, or the toxicodynamics of compounds within the mixture and thus the toxic potency (see section 2.1.2).
The components of a mixture perform either with a similar or dissimilar toxic effect. A distinction is made between *concentration addition*(CA) and *independent action* (IA). CA occurs if compounds share the same mode of action. Thus, they affect a biological endpoint on the same biological pathway by interacting with the same molecular target. The assumption for IA is that multiple compounds contribute jointly to the same biological endpoint. However, compounds act on different targets or modes of action *. One compound can not interact with the same biological entity as another. Thus, each chemical acts independently. Based on the mathematical formulations of IA and CA (see section 2.1.2), environmental toxicologists determine interaction types. In general, one compound may reduce the activity of another compound — *antagonism* — or may enhance it — *synergism*. The empirically measured mixture effect can be different from the estimated effect with single compounds for the concentrations in the mixture. For example, an empirically measured mixture toxicity can be underestimated due to antagonism.

CA is relevant, especially in the context of environmental mixtures. For example, narcotic compounds join their toxic potency to the mixture effect, albeit the single compound concentration levels may not induce toxic effects [Abernethy et al. 1988]. Comprehensive monitoring studies detect thousands of anthropogenic compounds in water bodies, mostly on shallow concentration levels [e.g. Bradley et al. 2019, Liška et al. 2015]. Subsets of hundreds to thousands of compounds in environmentally relevant mixtures share modes of action. Assuming that those compounds may not physically interact and have a similar action, this mixture may lead to toxic effects on the target organism, albeit not toxic in the concentration of a single chemical.

Chemical assessment of environmental sites. Environmental samples from ecological

* In case, the concept of modes and mechanisms of toxic actions is unfamiliar, have a look at subsection 1.1.3) and supplemental section S1.2 first.

sites of interest are collected to evaluate an organism's hazard of exposure to specific xenobiotics. For the selected site, samples of water, sediment, soil or biota are investigated. The sample consists of a mixture of up to thousands of chemical compounds. Some of them may be well known and already of emerging concern for environmental risk.
The traditional approach to monitoring the quality of the environment is a quantified analysis of chemicals of interest. A *targeted chemical analysis* measures the concentrations of a pre-selected set of compounds, expected to be an ecologically relevant set known to affect human or environmental health. However, alternatives to chemicals of emerging concern are developed, synthesised in industry, and released into the environment. These may also lead to adverse environmental outcomes, but the targeted chemical analysis is limited to the traditional compounds. Targeted chemical analysis potentially ignores derivatives, metabolites or degradation products of the chemicals of concern. Furthermore, chemicals can also be not detected, albeit present, but below an analytical detection level. An alternative is to identify all chemical compounds in a *non-targeted chemical analysis* applying mass spectrometry and chromatographic fractionation of samples. However, the non-targeted approach is more resource extensive and also the respective data analysis needs more effort than the targeted one. Thus, environmental monitoring still aims to investigate intelligent selections of chemicals of concern.

Furthermore, the assessment of biological effects is relevant and measured on selected biochemical mechanisms and environmental endpoints upon chemical exposure. Therefore, toxicologists have developed various diagnostic approaches and tools for hazard assessment and monitoring to assess environmental quality in bioanalytical assays. A so-called *bioassay* is a biological test system that measures the performance of a biological endpoint upon xenobiotic exposure. Such a tool is based on biological entities from different levels of biological organisation and thus is investigated either *in-vivo* or *in-vitro*.
The *in-vivo* bioassays determine the ecological relevance and assess the toxic potency on an organism or population level. However, for frequent and regular measurements in a regulatory and monitoring context, animal-free or -reduced alternatives are necessary, especially from a bioethical point of view. Furthermore, the approaches are resource-intensive concerning time and money.
The *in-vitro* assays investigate effects in tissues, cell(line)s or proteins and detect mechanism-specific responses. These bioassays need a small test volume, can be tested in a short time and are easier to interpret high-throughput.
One of the first highlighted environmental matters has been endocrine disruption [Carson 2002], which affects hormone regulation and induces oxidative stress [van Duursen, Majorie 2021]. For example, xenobiotically induced estrogen activity can feminise a species population

which affects reproduction. In the worst case, this lead to a collapse of a population [Kidd et al. 2007]. Nowadays, a set of *in-vitro* assays can assess endocrine disruption. Supplemental figure S1-1 presents such a cell-based bioassay detecting estrogen activity. *Equivalent activity levels* of natural or synthetic estrogens define measures of estrogen activity. Standardised estrogen references are 17-β-estradiol (E2) or 17-α-Ethinyl-estradiol (EE2). The approach allows estimating the degree of adversity without animal testing based on an equivalent value to a well-investigated estrogen. Comparing the equivalent estrogen concentration to a previously defined threshold determines the ecological risk. Other (*in-vitro*) bioassays can measure other endocrine disruptive mechanisms like anti-estrogenity, androgenity or oxidative stress. Currently, standardised batteries of bioassays build the framework of effect-based risk assessment and monitoring. For example, a battery may represent the mechanistic knowledge from a molecular interaction with a toxic compound up to the adverse biological outcome and is based on the systems biological concept of adverse outcome pathways (see section 1.1.3).

Another practical framework is the *effect-directed analysis*, which combines chemical analytics and bioassays to identify new compounds that show any activity in biological analysis *. Although such sophisticated approaches are available, the chosen set of endpoints and bioassays leads to a bias in the biological analysis of an ecological site towards a specific nature of toxicity. Next to the plethora of bioassay tools considering a small set of biological targets, high-content-screening approaches emerge in ET. *Omics-based exposure experiments* generate *Big Data*, e.g. on the transcriptomic, proteomic and metabolomic levels. The systems biological considerations (see section 1.1.3) and omics-based approaches (see section 1.2.1) expand the toolbox in ET.

1.1.3 Systems biological perspectives in environmental toxicology

Systems biology defines the integrated study of a biological entity, its properties, and its components' interactions [Yosim and Fry 2015]. The biological entities might be on different levels of biological organisation. This conceptual framework is often studied on cellular to organism levels but also on population or ecosystem levels. Iterations of assessing and enumerating all sub-entities and their interactions and predicting how the biological entity may respond to perturbations define the systems biology approach. In this respect, system

* After a liquid or solid phase extraction compounds of interest are concentrated in samples. The sample analysis is based on a chosen set of bioassays. Applying chromatographic approaches the compounds in a sample are separated in to simpler mixtures and are tested in the battery of bioassays again. For the bioactive fractions, a chemical analysis helps identify the compounds. The identified xenobiotics are confirmed, when chemical analysis and biological analysis support each other.

biologists aim to understand and predict properties by inference approaches [Garcia-Reyero and Perkins 2011], e.g. for a specific cell type, a network of interacting genes, proteins, and biochemical reactions can be integrated. Networks help characterise or understand complex biological processes. A systems biology strategy allows environmental risk and hazard assessment and the discovery of a more varied amount of biomarkers tied to environmental exposures [Yosim and Fry 2015]. This section presents a selection of such strategies useful for model generation and problem formulation in ET.

Eco-exposome. The concept of the *exposome* originates from human health and complements the concept of the genome [Miller and Jones 2014]. It defines an individual's lifelong and cumulative measure of environmental exposures and biological responses to its environment, diet, behaviour, and endogenous processes [Wild 2005, Miller and Jones 2014]. The joint investigation of genome and exposome helps understand mechanisms of toxic actions [Yosim and Fry 2015].

The National Research National Research Council [2012] has proposed the *eco-exposome* concept expanding the exposome concept to generalise exposure studies to environmental problems. However, the concept is somewhat idealistic regarding ethical and scientific limitations [Discussed in Wild 2012]. Scholz et al. [2021] defined a narrower eco-exposome concept concerning the lifelong internal exposure to individuals of a selected species and described current challenges and potential solutions in eco-exposome assessment. For example, to investigate the total (eco-)exposome of an organism or an ecosystem, respectively, is not achievable by the recent means of exposure studies. Exposure assessments investigate only a snapshot of the eco-exposome. A comprehensive study design combining various chemical analyses, bioassays and omics-based approaches may help investigate a broader spectrum of chemicals in the environment [Scholz et al. 2021]. To overcome the technical restrictions of investigating a lifelong exposure, a partial (eco-)exposome can be considered, e.g. in model organisms under fixed conditions or in human cohorts [Wild 2012]. Besides, each organism undergoes a transition through different developmental stages, and external exposures may differently affect organisms across various stages. Therefore, trans-sectional sampling is proposed as a potential solution [Scholz et al. 2021].

The eco-exposome assessment is advantageous to the exposome assessment as it allows examining entire organisms and target tissues and investigating populations and not only individuals [Scholz et al. 2021]. This has led to interdisciplinary projects for environmental health questions, e.g. the aim to investigate proxies of the aquatic eco-exposome [Escher et al. 2017].

Description of molecular interactions. The concepts of the *mode of action* (MOA) and *mechanism of action* (MeOA) describe specific and unspecific molecular interactions and the induced effect cascades in biological entities. The concepts are hardly distinguished and used ambiguously. Escher et al. [2011] has defined MOA as a

> “ *common set of physiological and behavioural signs that characterise a type of adverse biological response* ”

and MeOA as a

> “ *crucial biochemical process or xenobiotic-biological interaction, or both underlying a given mode of action.* ”

After the intake of a compound in a biological entity, it may interact with a protein as a binding ligand. It may be a receptor, enzyme or other target protein. Under normal conditions, the receptors respond to specific endogenous signalling ligands, e.g. hormones or neurotransmitters, and lead to a regulating response by, e.g. interfering with ion channels G-protein coupled receptors or nuclear receptors. However, exogenous compounds may be a chemical with a similar active group as the endogenous ligand leading to a concurrency for the receptor binding sites with the natural ligand. Consequently, xenobiotic ligands may activate the receptor protein as an agonist or inactivate the receptor as an antagonist *. Thus, the overall receptor activity level in a biological entity may be up- or downregulated by the xenobiotic influences. For example, the interactions with nuclear receptors affecting hormone regulation may induce *endocrine disruption*, or the affection of neurotransmitter interactions infer with ion channels and may induce *neurotoxicity*.

Adverse outcome pathways. Over the last decades of ET research, cost-effective and high-throughput assessments on adverse outcomes replaced traditional tests and allowed predicting xenobiotic toxicity. In current hazard assessment, frameworks allow linking *in-vitro* and *in-vivo* approaches to endpoints in human or environmental health. Ankley et al. [2010] has defined the *adverse outcome pathway* (AOP) to tackle this challenge. The AOP is a conceptual framework to describe the mechanistic knowledge of toxicology (see figure 1.3).

* Various molecular interactions and cellular processes induce toxic effects and affect the cellular system negatively. However, this would exhaust the scope of this introduction. Instead, the curious reader finds a descriptive overview of the main types of MOA in the supplement section S1.2.

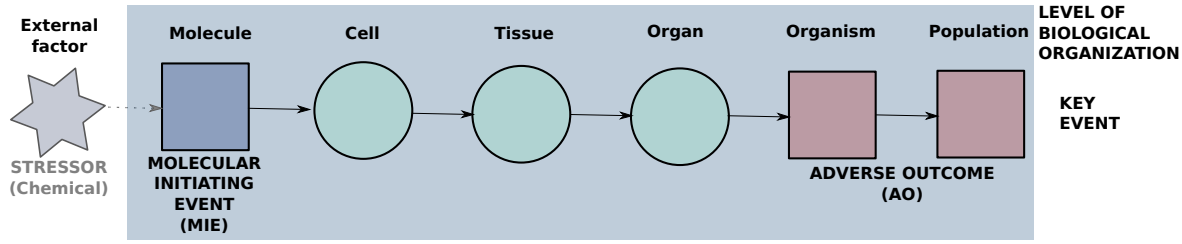

Figure 1.3. ***The adverse outcome pathway*** *is a chain of key events across different levels of biological organisation linked by key event relationships. This mechanistic knowledge representation is linear, modularised, often not species-specific, chemical-agnostic and, in the ideal case, evidence-based.*

Each *key event* (KE) represents a module or structural entity, which states a measurable change in a biological process. An AOP starts with a molecular initiating event (MIE), which an external stressor may trigger. An MIE interaction with an external stressor may be a chemical interaction with a biological receptor activating a cascade of different biological processes across different levels of biological organisation (LOBO) - the chain of KEs linked by *key event relationships* (KER). The modularised sequence may end with an adverse outcome (AO) on the population or the individual level. In a simplified manner, the AOP describes the link of molecular responses to the impacts on ecological or health endpoints [Villeneuve et al. 2014].
The assessment of evidence relies on different approaches. Initially, *in-vitro* or *in-vivo* exposure experiments assess the adverse outcome of chemical exposure. Then, high-throughput screening approaches can be used to investigate the early KEs on molecular and cellular LOBO with *in-vitro* assays or omics-based approaches. The *in-vivo* approaches are still necessary to bridge the evidence-based approaches for intermediate and late KEs regarding tissues, organs, organisms or even whole populations.
The AOP has been considered the most suitable framework for data integration approaches when linking the molecular effects to the adverse outcome in hazard assessment [Roelofs, Dick 2021]. The organisation for Economic Cooperation and Development (OECD) provides the most recent collection of postulated and proven AOPs. It is presented in an open-access database named AOPwiki and is widely used in data integration approaches.

In sum and alone, exposures in varying doses and complexities have chronic and acute effects on an organism over a lifetime. Generalised to ET, molecular interactions of an ecosystem can be aggregated and investigated due to their exposure-specific adverse outcomes over an extended period of time - the eco-exposome. In that respect, strategies have become necessary

to link chemical exposure to molecular effects. The linkage of complex chemical exposure to biological effect can be established by joint consideration of the eco-exposome concept with knowledge of chemical exposure, the MOA concept and the AOP framework [Scholz et al. 2021].
Furthermore, independent of investigating single compounds, artificial mixtures or environmental relevant samples, influences on higher levels of biological organisation are considerable and are apically measurable as adverse outcomes. The above-presented concepts in systems biology help describe mechanistic knowledge and build a foundation for computational approaches to link chemical exposure to biological effects. For example, whole bioassays and gene sets have been developed for MOA and MeOA specific molecular targets. Furthermore, an entire research field concentrating on AOP development has arisen. Stored in the AOP-wiki, a mechanistic knowledge representation has been developed.
Consequently, various data and knowledge bases are available and integrated with each other or empirically measured data. In this respect, methods (1) to investigate and link chemical analytical data with biological effect data and (2) to predict further knowledge are central for computational toxicology and the focus of this dissertation.

1.2 Computational toxicology

Initially, researchers have been interested in studying individual components, like nuclear receptors, to better understand, e.g. cellular biology or diseases. However, the biology-related sciences aim to more and more integrate the components to understand the interactions of the biological systems. The previously presented systems biological concepts have formed the basis for the computational achievements during recent decades to investigate biological effects after environmental perturbations across different levels of biological organisation.
Next to the plethora of bioassay tools considering a small set of biological targets, high-content screening approaches have emerged in ET. This section presents omics-based approaches, in particular transcriptomics, in ET (section 1.2.1) and how to link chemical exposure to transcriptional effects (section 1.2.2). Based on such inference approaches with empirical data, we shed light on the data integration approaches and systems biological perspectives upwards to higher levels of biological organisation (section 1.2.3). Furthermore, we elaborate on already used toxicological knowledge-driven approaches and how a transfer from related scientific areas are helpful in predictive computational toxicology (section 1.2.3).

1.2.1 Omics-based approaches

In recent considerations of ET, the exposome, but also the genome concept, have been regarded when aiming for understanding the interplay between environmental perturbations and biological effects. After describing the first concept in the previous section, the genome goes back to the advent of genetics in the last century (see supplemental section S1.3).

As already highlighted in section 1.1, exposures to toxicants, even mildly concentrated, may induce biochemical changes in biological entities. Such changes can affect the homeostasis of the cell-internal environment. Furthermore, the cellular responses to perturbation serve to minimise the xenobiotically induced damage. Such responses can induce a significantly measurable change in gene regulation activity and are partly also specific to toxicants (see section 1.1.3). Thus, the observation of specific stress responses may be associated with a specific exposure. The molecular biology principles (described in more detail in supplement section S1.3) are applied to assess the cellular *status quo* and understand cellular responses after perturbations. Biologists quantify and analyse the amount of a group of biomolecules in cells, tissues, organs or whole organisms at a specific time point. In an omics-based ET study, organisms or smaller biological entities are exposed to a (mixture of) environmental toxicant(s). Depending on the experiment's objectives, pools are investigated through differences in exposure due to concentrations, time-points, durations or sites. Various omics-based approaches have been developed, such as transcriptomics, proteomics, and metabolomics. All have some relevance in ET [van Straalen, Nico M. 2021b] *. This thesis concentrated on transcriptomics.

A cell's transcriptome consists of all transcripts from one experimental condition. Accordingly, it is assumed that the amount of mRNA transcript copies of one gene depends on the cell-environmental conditions. *Transcriptomics* allows determining up- or downregulation of genes and comparing different treatment groups or time points. Consequently, a transcriptomic analysis quantifies differences in gene expression between the sample pools. ET-based transcriptomics aims to gain a complete overview of all changes in mRNA abundance in a biological entity as a function of exposure to environmental chemicals [van Straalen, Nico M. 2021b].

In general, two types of transcriptomic approaches are used frequently in ET (see supplement section S1.4) — *hybridisation-* and *sequence-based*. A *hybridisation-based* analysis quantifies

* In the last decades, omics-integrative approaches and tools have emerged to computationally link and understand the cellular responses on different molecular levels [Koh and Hwang 2019, Martins et al. 2019] - the era of *multi-omics* approaches. Through the joint use of omics approaches and the applied systems biological considerations, the insights gain strength. However, these developments go beyond the presented research.

the amount of a specific set of mRNA transcript copies on *microarrays* — small glass plates with fixed labelled cDNAs. Per exposure condition, the normalised intensity of gene expression per probe is detected. A gene's response is expressed relative to the measured intensity of transcripts to a control condition. Experiments and study designs applying microarray analysis have been established to indicate gene regulation activity for diagnostic purposes in MOA studies identifying exposure-specific effects [e.g. Zare et al. 2018, Lichtensteiger et al. 2015, Snell et al. 2003]. Furthermore, high-throughput microarrays can be customised to consider a selected set of biomarker genes as a form of bioassay [van Straalen, Nico M. 2021b].
Microarrays have the advantages of being high-throughput and cost-efficient [Martins et al. 2019]. However, there are also limitations. For example, the outcome is biased to the chosen space of cDNA-labels, which may also affect the sensitivity for lowly-abundant species [e.g. Shendure 2008]. Furthermore, cross hybridisation induce noisy background levels in measurements [Wu et al. 2005], leading to problems in reproducibility, e.g. across laboratories [Feswick et al. 2017].
The *sequence-based* methods infer a cDNA sequence directly and thus are not reduced to a predefined set of RNA sequences. Therefore, the so-called RNA-seq is considered a more systematic analysis of gene expression patterns [Qian et al. 2014]. The advantages of RNA-seq are the ability to quantify a broad coverage of RNA transcripts, including unknown variants (e.g. splice variants), and its better applicability for experiments in non-model organisms. However, this may also be the main limitation, as computational analysis becomes more cumbersome [Martins et al. 2019, Qian et al. 2014]. Still, microarrays are frequently used in ET due to ostensible cost-efficiency, standardised computational and bioinformatics analysis, and prioritisation of the assessed genes [Martins et al. 2019].

Transcriptional gene expression analysis has occupied a niche in ET [van Straalen, Nico M. 2021b]. Transcriptomics contribute to hazard assessment in MOA studies by monitoring and transcriptional fingerprint imaging and is an alternative to biomarkers. In this respect, three primary advantages have to be highlighted:
First, the gene expression analysis is rapid, and respective analyses can go on in magnitudes of hours and days. Thus, they can be helpful for quick decision-making *.
Second, the gene expression is specific. A transcription profile comprises hundreds to thousands of genes — a vast space of genetic information. When comparing such profiles under different exposure conditions, distinctive patterns may be identified. Profiles have been proven to be exposure specific [e.g. Schüttler et al. 2019, Subramanian et al. 2017] and effect specific [e.g. Zare et al. 2018, Lichtensteiger et al. 2015], which is a significant advantage in

* However, it is in the same magnitudes as traditional toxicity tests with chemical analytics and bioassays [van Straalen, Nico M. 2021b].

hazard assessment. Thus, exposure-dependent profiles have been developed, which are used for toxicogenomic fingerprint imaging [e.g. Subramanian et al. 2017, Sutherland et al. 2018, Krämer et al. 2020] or to develop toxicokinetic-toxicodynamic models [e.g. Schüttler et al. 2019].
Third, the gene expression is sensitive. As gene regulation is an early biochemical response in a biological entity, xenobiotic effects and responses are expected for lower no-observed effect concentrations already. Such effects can be sub- or non-lethal and can be detected much earlier than biological endpoints like survival, growth and reproduction. However, this advantage has its limits when considering the complexity of chemical mixtures [van Straalen, Nico M. 2021b].

There are also disadvantages associated with transcriptomics in ET. Many gene expression analyses, especially with microarrays, are biased in their interpretability towards the genomic resources, like a well-annotated genome assembly. The gene expression analysis requires a knowledge-intensive infrastructure, including a high level of expertise for analysis and follow-up investigations. A plethora of computational strategies has been established dealing with the disadvantages and profiting from the advantages of transcriptomics. In the following, we describe a relevant selection of such strategies.

1.2.2 Linking chemical exposure to transcriptional effects

Researchers, who work with omics data, aim to identify gene regulation alterations associated with a treatment condition, a phenotype, or an induced perturbation. The field of *toxicogenomics* comprises the investigation with omics-based approaches and, e.g., investigates gene expression changes due to chemical exposure.
As shown in figure 1.4, scientists measure transcriptional responses applying real-time quantitative poly chain reaction, microarray or RNA-seq analysis. Based on the exact research question, various statistical analyses are considerable. However, the shown analyses are only a selection of practical approaches to understanding xenobiotic effects on the molecular level. Over the last decades of omics-based research, many gene expression analysis strategies have been developed that achieved meaningful biological outcomes.

Such approaches allow identifying single gene markers for perturbation and applying downstream analysis considering co-expressed gene sets [e.g. Schüttler et al. 2019, AbdulHameed et al. 2016]. Consequently, researchers have generated toxicogenomic fingerprints [e.g. Krämer et al. 2020, Subramanian et al. 2017, Wang et al. 2016] or have determined gene co-expression networks [e.g. Maertens et al. 2018, Ewald et al. 2020] and exposure-related classification systems [Ornostay et al. 2013, Nagata et al. 2014].

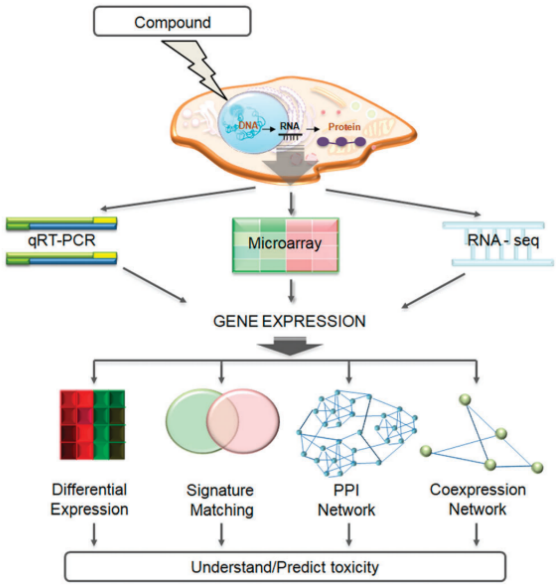

Figure 1.4. ***Overview of common computational strategies applying transcriptomics in environmental toxicology.*** *After applying a chemical exposure experiment on a biological system, the gene expression is measured with transcriptomic approaches (see section 1.2.1). Regarding differences in exposure conditions, the differences in gene expression are examined with, e.g. differential gene expression analysis or network inference. The respective outcomes help to better understand toxicity or to develop toxicity prediction tools. Taken from [Alexander-Dann et al. 2018].*

(Multivariate) Linear models of gene expression are frequently applied to describe such exposure-dependent gene expressions mathematically. The most conventional method statistically studies each gene independently. It determines whether a gene is differentially expressed under a perturbed condition compared to a control condition - the *differential gene expression analysis* [Shi and Walker 2008]. The model designs vary from simple two-group comparisons to complex models with multiple experimental factors [Ritchie et al. 2015]. Nowadays, well-established tools for microarray analysis and RNA-seq experiments are available [Ritchie et al. 2015, Love et al. 2014, Leek et al. 2006] and are frequently used also in environmental toxicogenomics research [e.g. Schüttler et al. 2017, Simões et al. 2018, Wang et al. 2016, Nair et al. 2020, Asselman et al. 2018, Limonta et al. 2019, Ewald et al. 2020]. Furthermore, other multivariate modelling approaches are relevant in ET. For example, regularisation of high-dimensional models helps retrieve comprehensive descriptors from various biological and environmental factors [e.g. Su et al. 2019, Li 2015]. Also, the partial-least-square-discriminant-analysis has been applied in omics-based ET. For example, its application has reduced high-dimensional exposures to a sparse set of exposures affecting a set of molecular markers [e.g. Skelton et al. 2014, Gandar et al. 2017, Jain et al. 2018].

Also, non-linear modelling approaches are critical in the context of ET and have found various ways to be insightful for exposure studies. *Machine learning* is applied to cluster or classify the profiles of exposures [e.g. Luechtefeld et al. 2018, Kapraun et al. 2017], gene expression [e.g. Ewald et al. 2020, Schüttler et al. 2019] or biological functionality [Ewald et al. 2020]. Thus, it is helpful in toxicity profiling or prediction. In general, the approaches can be split into supervised and unsupervised learning approaches. The supervised approaches group gene expression patterns with the help of pre-trained classifiers. They are often based on support vector machines [e.g. Tawa et al. 2014] or random forest [e.g. Antczak et al. 2013, Luechtefeld et al. 2018, Hou et al. 2020]. For example, Hou et al. [2020] have compared different non-linear machine learning approaches in their ability to estimate ecotoxicological characterisation factors and have highlighted random forest as suitable based on CompTox data *.

Unsupervised approaches are applied to cluster groups based on their expression profile without prior knowledge, using, e.g., k-means or hierarchical clustering. In (eco-)toxicogenomics, unsupervised machine learning has been already applied for, e.g. biological effects surveillance [e.g. Schroeder et al. 2016; 2017] or prediction of toxic or morphological effects [e.g. Hermsen et al. 2012, Antczak et al. 2013]. For example, the self-organizing map approach has profiled (eco-)toxicogenomic fingerprints [Wirth et al. 2011, Schüttler et al. 2019, Krämer et al.

*CompTox Dashboard: https://comptox.epa.gov

2020]. This relatively sophisticated machine learning strategy uses gene expression network representations and groups co-expressed genes based on k-means or hierarchical clustering.
The unsupervised machine learning approach of *association rule mining* (AR) has been applied in recent exposure studies for different purposes. The approach determines the rules which describe co-occurrences of frequent item sets in the data *. Initially, the approach has been used in market basket analysis. The approach has been recently considered for the association of exposures [Barrera-Gómez et al. 2017, Kapraun et al. 2017, Santos et al. 2020] and gene expression to diseases and exposures [e.g. Toti et al. 2016, Lakshmi and Vadivu 2019] and in multi-omics investigations [Mallik and Zhao 2017]. Furthermore, AR is applicable in a supervised manner as a toxicity prediction tool by constructing a classification system of association rules. For example, Nagata et al. [2014] predicted relative changes in liver weight with an association rule mining approach called classification based on association. This prediction task outperformed the application of linear discriminant analysis and identified meaningful biological results and allowed the authors develop interpretable prediction models. Although AR can help link chemical exposure to molecular effects, it has not been yet considered for eco-toxicogenomic purposes.

Network inference determines or predicts relations from toxicological knowledge networks. Consequently, network inference is a well-established method in biomedical research like single-cell omics [reviewed in Fiers et al. 2018] or cancer research [e.g. Niemira et al. 2020, Tian et al. 2020]. In (eco-)toxicogenomics, network inference approaches are also increasingly used [reviewed in Barel and Herwig 2018, Alexander-Dann et al. 2018]. Therefore, the gene expression is compared in a pairwise manner based on empirically measured data — conveniently from a transcriptomic analysis — considering at least two different treatment conditions. A gene co-expression network can be generated based on a gene-similarity matrix, e.g. an adjacency matrix based on Pearson correlation. Correlation-based networks have been widely used to understand gene regulation and infer knowledge [e.g. Ewald et al. 2020, Maertens et al. 2018, Sutherland et al. 2018, Orsini et al. 2018, Degli Esposti et al. 2019, Asselman et al. 2018]. Furthermore, other approaches have been applied like Boolean networks [e.g. Rodríguez-Jorquera et al. 2019, Kauffman et al. 2004, Akutsu et al. 2000, Jimenez et al. 2015], or reverse engineering networks [e.g. Perkins et al. 2011; 2017, Catlett et al. 2013]. The *weighted gene co-expression network analysis* (WGCNA) [Langfelder and Horvath 2008] is a state-of-the-art network inference approach Zhao et al. [2010]. WGCNA generates a network based on the pairwise correlation of the gene expression patterns across transcriptomic samples †. The constructed network consists of genes as nodes and the pairwise similar-

* A more technical introduction follows in section 2.1.4.
† A more technical introduction follows in section 2.1.5.

ity of genes as edge. Approaches, like k-means or hierarchical clustering, help cluster the constructed gene correlation network. The clusters (*modules*) are assumed to consist of co-expressed genes. Further investigations of the gene-expression (sub-)networks allow the inference of biological meaning (see figure 1.5). For example, with the help of enrichment analyses, a modular set of genes can be enriched to a previously curated gene set, which may represent biological functions, biological compartments, diseases, or (chemical) perturbations [applied in, e.g. Sutherland et al. 2018, Ewald et al. 2020]. Furthermore, the approach determines the biological meaning of genes with unknown functions, prioritises gene markers for biological entities and phenotypical endpoints. In the context of toxicogenomics, WGCNA has been applied to understand drug toxicity [Sutherland et al. 2018] or to associate gene markers to adverse outcomes in humans [AbdulHameed et al. 2014]. There are also examples of exposure studies applying WGCNA [Maertens et al. 2018, Degli Esposti et al. 2019, Ewald et al. 2020, Asselman et al. 2018].

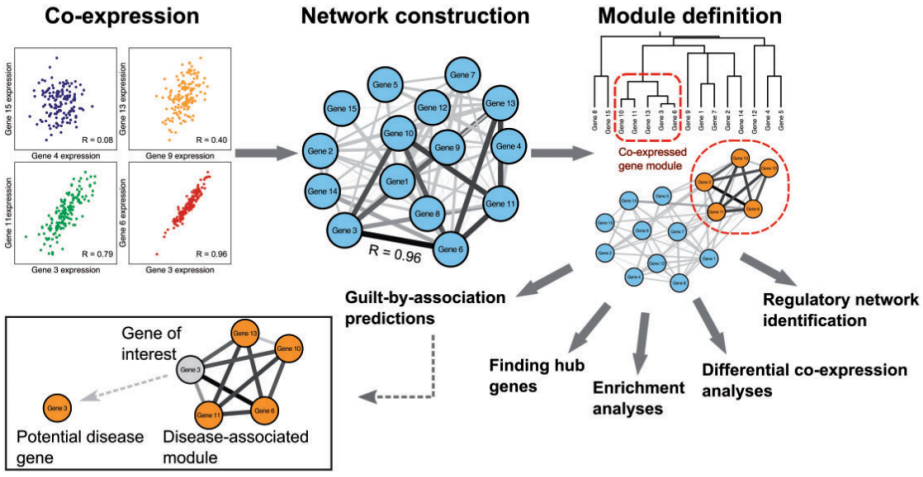

Figure 1.5. ***Overview of network inference analysis with individual tasks.*** *The pairwise correlation is calculated to determine gene co-expression in transcriptomic data. After adaption to interconnectedness measures, a gene network is generated. Groups of densely connected genes - modules - are determined. Following downstream tasks help to prove the biological reliability of exposure-related effects or to select toxicogenomic relevant targets. Taken from [van Dam et al. 2018].*

To sum up, the three introduced and frequently used methodologies link chemical exposure to transcriptional effects and are relevant for environmental toxicogenomics. Linear modelling of gene expression analysis is one of the most frequently applied omics-based approaches. In toxicogenomic analysis, researchers compare different (environmentally relevant) exposure

conditions based on differential gene expression. In that respect, transcriptomics-based approaches in environmental hazard assessment and monitoring already aim to handle complex mixtures of often mildly concentrated compounds.
However, also non-linear toxicogenomic approaches have been established in (environmental) toxicology. They are often considered in combination with differential gene expression analysis. Especially for systems biology purposes, network inference and machine learning approaches have investigated exposure-related transcriptional effects.
However, whether the strategies are similarly well-applicable stand-alone to computationally link complex mixtures of environmental samples to transcriptional expression effects has not yet been evaluated. Such an examination is uncommonly challenging when investigating environmental mixtures of lowly concentrated chemicals. In that respect, a consistent gene expression is not expected for the plethora of considered transcripts. Thus, a coarser resolution level of biological information can help, e.g. achieved by an up-scaling towards higher biological organisation levels. The following section presents such up-scaling and systems biological approaches used in the ET context.

1.2.3 Up-scaling from the gene level to higher biological organisation levels

Although the identification of endpoint- or exposure-related genes are helpful for the determination of molecular targets, these approaches are limited to individual genes with significant effects and a slight variance due to perturbation condition. However, most cell biological responses involve more subtle changes due to perturbations, especially when considering chemical exposures in the environment. For example, few or no single genes may be significant in a DEA after multiple testing corrections. The genes with minor comparative differences and high variance are not covered. In consequence, lists of significantly affected genes overlap marginally across experiments with similar study designs or even across samples of one study. Individual genes in a regulated biological pathway may not be consistent but cumulatively statistically significant across different (biological) samples. It is more challenging to identify such responses with single-gene approaches robustly. Thus, a lack of reproducibility may occur on the transcriptional level. Not only in the environmental toxicology context, an understanding of higher biological levels by associating the biochemical or metabolic pathways is crucial to diagnose or prevent adverse outcomes. In that respect, descriptive data from molecular biology or biomedicine and toxicity- and chemistry-related data are publicly accessible. These data help link levels of biological organisation in a systems biology and data integration manner.

Functional enrichment approaches overcome the limitations of conventional single-gene approaches. It is assumed a complementary and crucial analysis in omics-based studies to associate biological functions with measured gene expression and inferred co-expressed gene clusters. Therefore, a list of determined genes is annotated or statistically associated with curated or predefined gene sets like biological pathways, biological entities, diseases, adverse outcomes or perturbations.

In general, two primary approaches of functional enrichment are used to statistically determine if a predefined set of genes is associated with the gene expression analysis results. First, the *overrepresentation analysis* (ORA) determines whether a list of genes is enriched for a curated gene set using a cumulative hypergeometric statistic. In an ET context, the approach is often based on differential analysis results comparing different conditions of phenotypes, treatment or perturbation.

The *geneset enrichment analysis* (GSEA) [Shi and Walker 2008, Mootha et al. 2003] considers all genes in an experiment and not only those above a significance threshold by using the Kolmogorov-Smirnov statistic for the enrichment score. Therefore, a gene expression analysis help determine ranked gene lists. If a gene set is related to an investigated perturbation, the gene expressions may have high association scores to the related biological pathways [Shi and Walker 2008] and the gene set is enriched to higher ranks in the list [Shi and Walker 2008] *.

Functional enrichment is an integral and state-of-the-art approach to gain more systems biological insights in omics-based analysis outcomes on higher levels of biological organisation from exposure-related gene expression results. Frequently used resources for gene sets related to cell biological pathways, biological compartments, or xenobiotic responses are available in, e.g. Gene Ontology (GO) [Carbon et al. 2019], KEGG [Kanehisa and Goto 2000, Kanehisa et al. 2004], WikiPathways [Pico et al. 2008], Reactome [Jassal et al. 2020, Fabregat et al. 2018], knowledge within `Ingenuity Pathway Analysis` Tool [Krämer et al. 2014] and MsigDB [Orešič et al. 2020, Subramanian et al. 2005]. Environmental toxicologists have applied GO-annotation [e.g. Vidal-Dorsch et al. 2013, Rodríguez-Jorquera et al. 2019], ORA [e.g. Martinović-Weigelt et al. 2014, Ewald et al. 2020, Krämer et al. 2020], GSEA [e.g. Thomas et al. 2011, Schroeder et al. 2017, Zare et al. 2018, Perkins et al. 2017, Martinović-Weigelt et al. 2014] or the Ingenuity pathway analysis [e.g. Loughery et al. 2019, Feswick et al. 2016, Perkins et al. 2017] to associate transcriptional responses after xenobiotic exposures to freshwater or wastewater samples. Functional enrichment can be performed with established web interfaces, R-packages and commercial software [e.g. Krämer et al. 2014, Sergushichev 2016, Wang and Liao 2020, Huang et al. 2009].

* A more technical introduction to gene set enrichment analysis follows in section 2.1.6.

Data integration in environmental toxicogenomics. Over the decades, many ET-related data have been produced, curated and stored systematically in publicly available databases. Whole research program initiatives have generated publicly available databases [e.g. Dix et al. 2007, Pallocca and Leist 2021, U.S. Environmental Protection Agency 2021, Barron et al. 2015] used for environmental monitoring and hazard assessment.
With high-throughput chemical effect databases, prior knowledge predictions of single chemical molecular effects are available. Resources like ToxCast [Dix et al. 2007] allow associating biological endpoints to single compounds directly. One way of applying this prior knowledge to assess the environmental risk is based on site-specific measurements of chemicals when only chemistry data are available [Schroeder et al. 2016]. However, when investigating environmental sites and their samples, compound do not occur alone. Furthermore, mixtures of chemical compounds have to be investigated, which still is a significant challenge in ET. The public availability of knowledge bases curating chemical interacting genes, protein, and pathways from literature and high-throughput or high-content biological assessments allows the current ET research to address complex mixture uncertainties [Schroeder et al. 2016]. As a result, the biological effects of mixtures have to be understood from a systems toxicological perspective, which requires integrating *in-vitro* and *in-vivo* data with descriptive statistical models. Furthermore, computational network approaches are needed for knowledge representation to link chemical exposure effect data with particular phenotypes or adverse outcomes [Hartung et al. 2017].

Recent data integration approaches also imply applications to validate the biological meaning and reliability of empirically measured toxicogenomic data across different levels of biological organisation [Martins et al. 2019]. The available resources cover a plethora of chemicals and biological endpoints and allow assessing chemical exposures in their chance to interact with molecular targets with strong weight-of-evidence. For example, knowledge-inferred interaction networks for proteins [e.g. Kuhn et al. 2008, Szklarczyk et al. 2016], genes and diseases [Mattingly et al. 2003, Davis et al. 2019, Krämer et al. 2014] or AOPs [Aguayo-Orozco et al. 2019, Jornod et al. 2021, Pittman et al. 2018, Martens et al. 2021, Pollesch et al. 2019] are used to, e.g. identify the effect-driven chemical compounds of exposed sites [e.g. Berninger et al. 2014, Garcia-Reyero et al. 2009, Schroeder et al. 2017, Perkins et al. 2017]. Schroeder et al. [2016] lists a variety of *in-vitro* and *in-vivo* chemical interaction databases. In the ET context, these have been considered for predictive toxicology and computational approaches for hypothesis generation [e.g. Schroeder et al. 2016, Luechtefeld et al. 2018, Schroeder et al. 2017]. A plethora of databases contains exposure-related toxicity data from empirical studies. Schroeder et al. [2016] highlighted the importance of chemical exposure associated with toxicogenomic relations from resources like ToxCast, CTD or STITCH for environmental

surveillance, hazard assessment and monitoring. Such databases may help understand the molecular mechanisms of toxicity in respect of various environmental toxicology problems.

The *search tool for interacting chemicals* (STITCH) [Szklarczyk et al. 2016, Kuhn et al. 2008] integrates disparate data sources for 430 000 chemicals linked to genes/proteins across various species. The STRING database [Szklarczyk et al. 2019; 2021] comprises interactions used for STITCH. These interactions contain information from metabolic pathways, crystal structures, binding experiments and drug-target relationships. For each listed species, STITCH offers an inferred interaction network. Based on the inferred knowledge, STITCH also gathers predictions of relations between chemicals or associated binding proteins [Kuhn et al. 2008]. The value of the resources for ecotoxicogenomics is already acknowledged [Martins et al. 2019, Perkins et al. 2017, Kongsbak et al. 2014] — however, only a few ET-related studies integrated STITCH in practice. For example, Taboureau et al. [2020] predicted human biological systems affected by endocrine disruptive perturbations. Therefore, xenobiotically perturbed systems have been determined by integrating STITCH with the Human Protein Atlas * and the Registry of Toxic Effects of Chemical Substances [Thul et al. 2017].

The publicly available *Comparative toxicogenomic database* (CTD) [Mattingly et al. 2003] helps understand how environmental exposures affect biological systems. The database has a focus on human health, and provides manually curated information from exposure studies and literature for different types of exposure-related interactions. Furthermore, CTD provides literature-based and manually curated interactions, allowing harmonising cross-species heterogeneous data for exposures and associated biological responses. In this respect, it also comprises environmental toxicology studies, e.g. on relevant aquatic vertebrate model organisms like zebrafish or fathead minnow.
Perkins et al. [2017] have linked specific chemicals to biological effects applying a combination of a knowledge-based approach and a gene expression analysis based on covariance with measured surface water chemistry. The chemical exposure has been associated with transcriptional changes in a workflow comprising DEA, correlation approaches, and Context Likelihood of Relatedness. Estrogenic effects on gene expression in caged fathead minnows have been detected and linked to the presence of bisphenol A by integrating CTD information.
Schroeder et al. [2017] have performed an integrated analysis of transcriptomic data considering site-specific knowledge assembly models. They have evaluated the chance of contribution of detected chemical compounds to the observed biological effect. These hypothesis models of chemically associated biological effects to compounds have been detected in the investigated sites based on CTD knowledge.

* http://accelrys.com/products/ collaborative-science/databases/bioactivity-databases/rtecs.html

Gu et al. [2019] have automatically extracted chemical-induced disease relation with an attention-based distant supervision paradigm capturing local and global attention features simultaneously. Based on CTD, training relations from biomedical literature have been collected indirectly.

ToxCast [Dix et al. 2007] is frequently used in combination with the curated knowledge in CTD for purposes of AOP development. The database gathers the high-throughput measures of single compound exposures in *in-vitro* and *in-vivo* assessments of thousands of chemicals. For example, Nymark et al. [2018] has developed a six-step workflow for integrating of toxicological knowledge and databases to develop an AOP scheme. As an example of an AOP-linked molecular pathway, they have identified a network of 64 CTD-derived and pulmonary fibrosis-associated genes. The possibility of CTD to link associations to literature and gene expression studies has strengthened the direct evidence of the resulted pathway. Furthermore, Doktorova et al. [2020] has proposed a workflow for AOP development integrating gene-pathway-disease relation data from ToxCast and CTD. Relations retrieved from TG-Gates [Igarashi et al. 2015] have filled the knowledge gaps within the generated linear AOP-sequences. As another example, Oki and Edwards [2016] applied frequent itemset mining to identify frequently occurring gene-disease pairs from ToxCast endpoints on the gene level or gene expression and disease information from CTD. This work has resulted in a computationally predicted AOP network with 18 283 assay-disease interactions and 110 253 gene-disease associations. The authors have shown that multiple data source integrations are beneficial to identify computationally predicted AOPs based on high throughput data.
Data integration approaches, e.g. with ToxCast, present a research focus in ET. The AOP concept (see section 1.1.3) has been considered the most suitable framework for data integration approaches when linking the molecular effects to the ecological endpoints in hazard and risk assessment [Roelofs, Dick 2021].

AOP-related computational toxicology. The *AOPwiki* [SAAOP 2021] contains the most recently contributed AOP knowledge provided by the scientific community. Furthermore, this open-source information base includes many AOPs on a hypothetical or theoretical level. The AOPwiki is frequently considered for data integration. For example, the resource allows comparative approaches across different species and thus the expansion of possible evaluation systems for KE measurements [Lalone et al. 2018]. Moreover, thousands of chemical stressors gathered in ToxCast have been linked to the current AOP knowledge Aguayo-Orozco et al. [2019].
The *AOP-DB* [Pittman et al. 2018] is related to the AOPwiki and is an exploratory database for hypothesis-generation purposes and associates the AOP framework with existing toxico-

logical databases like the CTD or STITCH allowing to generate association networks across the biological entities. In addition, various research initiatives develop sophisticated data integrative frameworks and tools [e.g. Martens et al. 2021, Mortensen et al. 2021, Nymark et al. 2018]. Such data integration approaches mine the toxicological knowledge from publicly available annotation databases to the already available knowledge in the AOPwiki.
Furthermore, in AOP development research, the intersections of different AOPs are investigated, as those may reveal new insights into biological interactions in *AOP networks* [Knapen et al. 2018, Villeneuve et al. 2018, Pollesch et al. 2019]. These networks are functional units and allow predicting endpoints based on measurements on a molecular and cellular level [e.g. Moe et al. 2021].

1.2.4 Biomedical literature-based discovery

Natural language processing (NLP) is an interdisciplinary research field between computer science and linguistics and concerns the interactions of computer and natural * languages [Kumar 2011]. For example, a natural language understanding system converts human language samples, like text, into a more formal representation that is easier to manipulate or process in a computer program. Such text analyses may comprise multiple tasks, like information retrieval, pattern recognition, tagging, data mining or literature-based discovery (LBD). In biomedical research, *LBD* has been a common approach for automated hypothesis generation [Zhao et al. 2021]. The task is to uncover previously unknown relations retrieved from existing knowledge. The LBD method originates from Swanson [1986], who assumed that knowledge in one scientific domain is related to another non-intersecting domain, albeit not known so far. The review of Zhao et al. [2021] about recent developments in biomedical literature mining highlights the challenges, purposes and limitations of LBD and related tasks. In the biomedical research context, available frameworks for knowledge representation are word embedding inference networks [e.g. Mao and Fung 2020, Choi and Lee 2019], or predictive models [e.g. Dollah and Aono 2011, Polavarapu et al. 2005, Peng et al. 2016]. As shown in figure 1.6, biomedical NLP approaches have shown a hierarchy in their tasks [Zhao et al. 2021]. Thus, to extract a hypothesis from current text-based toxicological knowledge, one of two preliminary requirements needs to be fulfilled: (1) NLP tools for *named entity recognition and normalisation* †, *text classification* and *relation extraction* are available, which can

* Languages that are human-made and naturally evolved [Kumar 2011]

† Both tasks are sequence labeling problems [Zhao et al. 2021]. Named entity recognition seeks for the location and classifies named entities in text into pre-defined categories such as person names or organisations or semantic concepts from the UMLS. The normalization maps obtained named entities into a controlled vocabulary.

be applied for the own text data, or (2) text mined relation data sets are available.

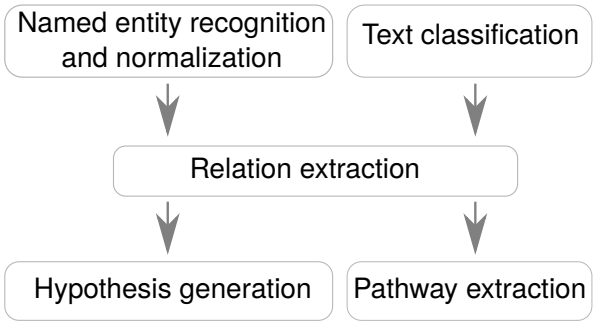

Figure 1.6. ***Natural language processing tasks*** *are relevant in biomedical research. The transfer of tasks towards environmental toxicology might help discover new exposure-related interactions (Hypothesis generation) or generate adverse outcome pathways (Pathway extraction). Adapted version of a figure in the review of Zhao et al. [2021].*

For exposure-specific AOP development, data integration has also been successfully implemented with text mining approaches [e.g. Jornod et al. 2021, Zgheib et al. 2021]. In consequence, the published literature knowledge have also become available for computational toxicology. Still, NLP approaches that consider the recent (toxicological) knowledge at once have been rare, and the respective AOP research is in its infancy. However, the current knowledge and achievements in biomedical LBD (see figure 1.6) may allow a knowledge transfer methodologically and toxicologically.

The National Library of Medicine and the National Institute of Health have generated fruitful text-based data and developed NLP tools in the human health context of toxicological research. The biomedical data infrastructure has been expanded with the *Unified Medical Language System* (UMLS) [Humphreys et al. 1998, Bodenreider 2004] and SemMedDB [Kilicoglu et al. 2012].
The UMLS contains three main knowledge entities of health and biomedicine, which combine vocabularies and standards to enable interoperability between computer systems: (1) The UMLS *Metathesaurus* contains the terms and the codes from many vocabularies with hierarchical links, definitions, and semantic relationships. (2) The UMLS *Semantic Network* contains broad categories, called semantic types and their relationships. (3) The UMLS *SPECIALIST Lexicon* is a large syntactic lexicon of biomedical and general English and comprises natural language processing tools *.
The UMLS has already been applied in different data integrative approaches or tools [e.g.

* More information follow in section 2.2.1.

Kilicoglu et al. 2012, Martens et al. 2021]. Furthermore, a comparative study applying embedding approaches [Mao and Fung 2020] has shown the ontological strengths of the UMLS. Consequently, the UMLS and their within contained NLP tools make biomedical researchers and researchers of related scientific fields an LBD approach easily accessible.
Text mining approaches may allow expanding the toxicological knowledge databases. In the NLP context, `SemRep` [Rindflesch and Fiszman 2003, Rindflesch et al. 2005] is a rather important tool, which semantically analyses biomedical texts based on UMLS resources. The tool identifies semantic predications sentence by sentence. First, the UMLS *Metathesaurus* assigns terms to a semantic concept using `MetaMap` [Aronson 2006]. Then, `SemRep` determines propositional assertions under consideration of the semantic and the syntactic constraints. It is a frequently used tool to extract semantic relations in a biomedical context and is known as a trustworthy and diverse interpreting baseline system [Kilicoglu et al. 2012; 2020].
Kilicoglu et al. [2012] generated a large-scale knowledge resource called *SemMedDB* — an expansion of the MEDLINE initiative [Ahlers et al. 2007]. The SemMedDB contains tables of semantic predications extracted from the titles and abstracts of all PubMed citations. Thus, the most current version of the SemMedDB contains the current biomedical knowledge from all PubMed citations. SemMedDB allows LBD in the biomedical context.

All these UMLS related resources have been a base for current developments in biomedical LBD and toxicology-related research. *Named entity recognition* and *relation extraction* can be performed with the UMLS tools `MetaMap` [Aronson 2006] or `SemRep` [Rindflesch et al. 2005]. According to Kilicoglu et al. [2012], SemMedDB has been proposed as applicable for advanced data-mining and to hypothesise novel relationships in the biomedical context and beyond.
The UMLS is a biomedical ontology and thus biased towards the human health domain. Ambitions have already been made to expand the UMLS ontology to further domains such as pharmacogenomics or medical informatics [e.g. Ahlers et al. 2007, Rosemblat et al. 2013b], but not ET.
The already mined and extracted relations in SemMedDB are also considerable for LBD. The data set has a unified ontology - the UMLS terminology - and is also based on current biomedical knowledge within PubMed. Recent studies have applied the SemMedDB data in the context of biomedical LBD, for example, to develop a graph of interacting semantic predications [Hristovski et al. 2015, Cong et al. 2019] or to identify causal drug-side-effect relations [Mower et al. 2017].

1.2.5 Deep learning with knowledge representation

Knowledge representation approaches lean on data integration approaches, for example, considering network inference, but also on deep-learning frameworks. Either way, knowledge representation helps rearrange, connect and predict information. Furthermore, databases and also text-based information bases contain toxicological knowledge. For example, Choi and Lee [2019] integrated the knowledge from three biomedical and toxicological databases and compared five knowledge representation models. The authors showed that the best performing model has been more accurate in inferring chemical-disease relations than the most recent approach in CTD. In consequence, data integration has been proven helpful to retrieve biologically more precise models. However, with the emergence of NLP, especially in biomedical research, knowledge represented in text format may also be relevant. Most frequently, deep learning knowledge representation approaches are utilised with word embedding models.

Word embedding approaches are used to learn a dense and low-dimensional representation from large and unlabeled corpora of text-based information and shall efficiently capture the semantics of words. In consequence, such an embedding transforms words, phrases or substrings into vectors of real numbers *. Densely distributed and low-dimensional vector representations of words are more suitable than one-hot encoded representations. Hence, the most common implementations consider the theory of distributional semantics † , such as the neural-network-based `word2vec` algorithm [Mikolov et al. 2013] and `GloVe` [Pennington et al. 2014]. In the context of human and environmental health, knowledge bases from different text-based corpora are available. There are successful implementations with word embedding and related knowledge representation approaches in recent deep learning applications for various literature mining tasks [reviewed in Zhao et al. 2021]. It has become clear that integrating domain knowledge helps improve semantic representations in a biomedical context [Zhang et al. 2019b]. The identified dependence of the representation from the corpus trained on has been one initial achievement [Wang et al. 2018]. Besides, word vector representations has been determined based on similarity and co-occurrence frequency of words [Smalheiser and Bonifield 2018]. Additionally, subword embedding models allow interpreting word-internal structures using character n-grams with implementations like fastText [Bojanowski et al. 2017]. Consequently, a recent study [Zhang et al. 2019b] has generated a biological word embedding based on the PubMed text corpus and the MESH term graph, enabling an out-of-vocabulary word consideration.

Furthermore, *contextual word embeddings* like `ELMO` [Peters et al. 2018] and BERT [Devlin

* A technical introduction follows in section 2.2.3.

† The distribution of surrounding words estimates the meaning of a target word.

et al. 2019] consider bidirectional language models with multiple attention-based transformer layers [Vaswani et al. 2017]. Such techniques allow generating multiple word embeddings for one word depending on its context. For example, BERT is trained on general English corpora and then contextualised by adding biomedical texts in a second training step. Various contextualised pre-trained biomedical-related word-embeddings have been generated, presented and made publicly available [e.g. Lee et al. 2019, Alsentzer et al. 2019, Michalopoulos et al. 2021, Peng et al. 2019]. For example, the UMLSBERT has learned semantic similarity of lexical words with the help of the UMLS semantic concepts and semantic types [Michalopoulos et al. 2021]. In this model, the word embedding vectors of words sharing the same semantic concept and the same semantic type have been adapted in training to become more similar. The model is publicly available, and it can recognise subwords and thus out-of-vocabulary words. However, considering a 'biomedical contextualised' embedding may increase the bias. Thus, recent pre-trained word embeddings may have their limitations when aiming for a toxicological hypothesis generation approach.

Recent approaches have applied mainly two ways to improve the model performance of word embeddings: Retro-fitting of pre-trained word embeddings [e.g. Zhang et al. 2019b] or combining different word embeddings [e.g. Mao and Fung 2020]. For example, Mao and Fung [2020] has combined word embeddings based on the retro-fitted BioWordVec [Zhang et al. 2019b] model and BERT-models. The combined model has performed better than single applications in selected biomedical NLP tasks. In the context of cancer pathology reports, Alawad et al. [2019] have developed a retro-fitted word embedding with UMLS-vocabularies, which has been focused on human health. However, considering a broad research scope like ET is limited by a word embedding model, that specialises in human toxicology.

Some text-based deep learning models in a biomedical context base on recurrent neural networks (RNN) [Elman 1990]. This network type enables the modelling of temporal or sequential data *. RNN has been applied to various computational tasks, e.g. handwriting recognition, activity recognition, or NLP. Long-short-term memory (LSTM) [Hochreiter and Schmidhuber 1997] is one kind of RNN and uses feedback loops of a previous step in the timeline or sequence $t-1$ for the output of the recent step t (see figure 2.3). Thus, a new cell in the recurrent network is considered for each word of a sentence sequence. For longer sequences, regular recurrent neural networks may be limited to model the dependencies between sequential steps separated by numerous others. The so-called vanishing gradient problem describes how small weights got eliminated due to multiple multiplications across time steps. Consequently, the weights of earlier layers have no significant changes, and the network forgets long-term dependencies. LSTM networks allow solving the vanishing gradient problem.

* A technical introduction follows in section 2.2.3.

In biomedical research, the deep learning approach of LSTM has been applied frequently [e.g. Chen et al. 2017, Zhao et al. 2019, Gu et al. 2019, Jimeno Yepes 2017]. For example, Jimeno Yepes [2017] have investigated biomedical word sense disambiguation in UMLS and have been shown that word embeddings improve the performance of more traditional features and allow using recurrent neural network classifiers based on LSTM nodes.

Admittedly, computational scientists widely utilise word embedding approaches. The variety of model architectures has helped contextualise information, handle long sequences, or use previous knowledge representative achievements. However, studies are very likely biased considering approaches on biomedical data as focusing on human health. Nonetheless, biomedical information also presents partly toxicological knowledge. For example, pharmacogenomics and cancer research also relies on transcriptomics studies. In addition, it may be related to at least common anthropogenic compounds such as pharmaceutical drugs, industrial or urban emissions and xenobiotics in food products. Thus, text-based biomedical data may contain at least some information relevant for exposure-related toxicology. Consequently, an exposure-related predictive task, such as determining hypotheses of links between chemical exposures and molecular biological effects, can be performed through knowledge representation applications considering deep learning or data integration approaches.

1.3 Research question and approaches

As stressed in the introduction, chemicals are released into the environment anthropogenically and may adversely affect organisms. The field of ecotoxicology monitors the chemical exposures in selected environmental sites and assesses the hazard and risk for therein living organisms. Omics-based approaches, especially microarray-based gene expression analyses, have become essential for purposes of toxicogenomic profiling and the assessment of exposure-related adverse effects. Regarding environmental sites of interest, scientists have to deal with complex mixtures and sometimes such with low chemical concentrations, although biological effects of concern may be measured. An exposure study, which investigates such conditions, deals with complex chemical analytical data and may concern high-content data of an omics-based approach. In the past, computational approaches have been applied to link chemical exposure to biological effects, such as multilinear modelling, machine learning and network inference. The plethora of computational approaches has also resulted in various combined strategies frequently considering data integration approaches to determine exposure-related molecular interactions and associate adverse effects. However, a strategy is missing to prove the stand-alone reliability of computational approaches to link chemical exposure to biological

effects, especially regarding such complex, empirically measured data. The first question in this dissertation stresses this knowledge gap: To what extent are state-of-the-art computational approaches – stand-alone and in combination – suitable to link chemical exposure to biological effects when considering environmental data concerning complex mixtures of lowly concentrated chemicals?

Moreover, a plethora of toxicologically relevant knowledge has been produced in exposure studies and is available from a variety of databases, information systems and literature. Knowledge representation approaches have been identified as a fruitful way to infer new information or connections from databases and text. Regarding data integration, also deep learning approaches become relevant, especially to harvest unknown information from the current toxicological knowledge. Inspired by the recent approaches in biomedical LBD, we believe there is a potential for a knowledge-driven prediction of exposure-related biological effects, which lead to the question: To what extent is the current and information-rich knowledge from literature and databases suitable to learn meaningful exposure-related interactions?

This dissertation stresses two issues, which are both relevant for knowledge retrieval in the context of environmental computational toxicology: (1) the assessment of the environmental status concerning exposures with complex mixtures and (2) knowledge-driven prediction from databases and literature. In concordance to developments mentioned earlier, this thesis deals with approaches to computationally link chemical exposure to biological effects employing data, which are complex structured and, thus, potentially rich in information but challenging in retrieving this information. Consequently, this thesis deals with the overall research question, whether such complex data can be applied to link chemical exposure to biological effects, and whether such associations are also biologically meaningful and to some extent reliable.

The study in chapter 3 linked complex chemical exposure of selected environmental systems to transcriptional effects and tackled the first question. We investigated three singularly applied computational approaches in their suitability to determine exposure-related effects on molecular and pathway levels, which might highlight a biologically meaningful and reliable attribution to adverse effects. Therefore, we examined empirical environmental data from an exposure study and assessed the xenobiotic effects of complex chemical exposures in ten streams in Minnesota. The challenges within this study were the retrieval of chemical-gene interactions when exposure patterns were not independent, and transcriptional effects were weak due to subtle toxic effects by chemicals. Applying different conceptual scenarios of exposure, we investigated the disentangling of chemical drivers in complex and environmental mixtures from a novel perspective of correlation-based compound groups. Finally, we deter-

mined whether the outcomes were somewhat trustworthy through a data integration approach with CTD and STITCH, and tested the biological and toxicological plausibility on gene level and on the higher resolution level of biological pathways.

The study in chapter 4 dealt with the second question. We employed semantic predications from the text-based biomedical knowledge of SemMedDB and aimed to predict the toxicogenomic relationships of chemical-biomolecule interactions by applying a deep learning model. We evaluated the model in its use to predict not-represented chemical-biomolecule interactions considering the current knowledge in CTD. An automated way of using the literature considering a knowledge representation had not yet been considered in exposure-related toxicology. The prediction of chemical-biomolecule relations considering natural language processed data and subsequent deep learning was a novel approach in ET.

Chapter 2

Methods and Data

This section describes the employed data and applied computational methodologies of sections 3 and 4, respectively.
It starts with reporting the selected data from an environmental and omics-based exposure study for the first investigation. Then, the three chosen bioinformatics approaches and the strategies to compare their outcomes are described. Furthermore, method comparative and systems biological approaches, considering functional enrichment and data integration, are reported.

The second part describes the selected biomedical resources of UMLS, SemRep, SemMedDB and toxicological reference data from CTD and how these data have been preprocessed. Then the applied deep learning prediction models for a knowledge-based discovery of chemical-biomolecule relations are presented. Lastly, the model evaluation strategies are described, including a considered toxicological application case.

2.1 Linking environmental relevant mixture exposures to transcriptional effects

Three data sets were provided by Dalma Martinović-Weigelt — a collaboration partner in Minnesota, US. Ferrey et al. [2017] * measured for ten small streams in Minnesota, (I) quantitative chemical exposure data of 146 pharmaceuticals and chemicals of concern in the US, (II) two *in-vitro* cell assay measurements for endocrine activity, and (III) gene expression data in liver tissue of fathead minnows after acute exposure to stream water ($n_{Samples,treated} = 64$; $n_{Samples,control} = 7$). Here, one transformed the chemical concentrations into toxic units and preprocessed the raw microarray data.
Applying three computational methods (Differential gene expression (DEA), Association rule mining (AR), and Network inference), exposure-related gene interaction sets were generated based on the given exposure and gene expression data. In addition, the exposure-associated gene results were validated with functional enrichment to biological pathways and chemical reference sets from external databases for each method.

2.1.1 Exposure and microarray data

Chemical compounds. Endogenous steroids had been indicated in previous surveys of waterbodies in Minnesota, but often at low concentration levels [e.g. Lee et al. 2010]. Ferrey et al. [2017] had measured 146 chemical compounds of emerging concern in the US (see supplemental table S2-1) with targeted chemical analysis in fifty streams in Minnesota. The list of compounds had comprised pharmaceuticals primarily The sites had been based on an *'internationally random selection'* [Ferrey et al. 2017] to represent a variety of streams in Minnesota, including ones that contaminants had not heavily impacted. A biological effect analysis had examined a collection of water samples from ten stream sites in Minnesota (see supplemental table S2-2) and a control sample of exposure-free water (ultraviolet-radiation-filtered Lake Superior water). The present investigation considered only the chemical measurements of these sites. If the compounds had been *not detected*, then the exposure concentrations were set to 0ng/L †. The compounds were not detected in any of the ten stream sites and were not considered as selected compounds.

* see report for detailed description on data sampling, microarray experiments and report of initial results.

† Not detected compounds were those where compound reports had been not above the respective detection limit and where compounds had been identified in samples and associated laboratory tanks. See Data Chapter3 (TenStreams)/FlaggedChemicalInput.xlsx PW: PhD_SKraemer) under consideration of supplemental table S2-3.

Bioanalytical data. To the chemical exposure data, *in-vitro* estrogen activity bio-assays * and *in-vitro* bio-assays for the activity of nitrates and nitrites † had been included to assess the exposure of selected stream waters. However, before providing the chemical exposure and bioanalytical data, Dalma Martinović-Weigelt had transformed the *in-vitro* activity levels to chemical concentrations of *EE2* and inorganic *nitrate*.

Fathead minnow microarrays. Mature fathead minnows had been exposed for 48h to either a control reference or a surface water reference collected from the ten selected sites. Five to seven fish per stream had been sacrificed, and the gene expression of respective liver tissue has been measured using a custom 60K-feature FHM DNA microarray (Agilent, GPL17 098). The provided data set contained 70 microarray samples.

In summary, the database for this study comprised ten stream locations, 146 measured compounds, EE2 and nitrate concentration equivalents of *in-vitro* activities, and 70 microarray samples.

2.1.2 Preprocessing

Transformation of chemical concentrations in toxic units. The *toxic unit* (TU) expresses the toxicity potency of a sample. In most cases, TU describes the ratio of the detected concentration c and a standardised chemical concentration for an environmental or toxic effect EC_x:

$$TU = \frac{c}{EC_x} \tag{2.1}$$

For selected chemical compounds from section 2.1.1, *toxic units* were calculated by dividing the measured concentration with an effect concentration for toxicity in fish. The effect concentrations in terms of LC50 values vary between (i) acute and chronic exposures, (ii) different fish species and (iii) QSAR-based prediction models. Therefore, for each selected compound, different available values (from ECOTOX and ECOSAR) were compared, and the effect concentration was estimated more roughly as an order of magnitude instead of taking

* The *in-vitro* estrogen activity bioassay has been chosen to capture relative levels of binding to target sites of the hormone estrogen and has been measured across the selected stream sites (see supplemental table S2-5). An androgenic *in-vitro* assessment has been performed as well (see supplemental table S2-5), but has not been quantified by any equivalent concentration and was not considered in the present study.

† Nitrates are a very good generic marker for pollution and land use in Minnesota and their activity. For sites impacted by agriculture, nitrate is a good marker for presence of a variety of contaminants that may have not been captured by the chemical analyses. Furthermore, recent work has been indicated, that nitrates may interact with endocrine function in gene expression and circulating hormones [Kellock et al. 2018, Bjerregaard et al. 2018, Pottinger 2017] and may be useful identifying transcriptional changes of chemical drivers.

an individual value (supplemental table S2-4).
An exposure pattern of one compound was viewed as the root-mean-square-scaled vector of values of its toxic unit transformed concentrations across the ten streams.

A pairwise Pearson correlation analysis (`cor(method='pairwise.complete.obs')`) was performed, and the compound wise exposure patterns were clustered in groups with similar exposure patterns - *compound groups* (CG) with `pheatmap` [version: 1.0.12 Kolde 2019].
The compound group exposure patterns were generated as the root-mean-square-scaled sum of the compound wise exposure patterns. Consequently, the mixture toxicity concept of concentration addition was applied * as low concentrated chemical concentrations may cumulatively affect one mode of action such as baseline toxicity, xenobiotic metabolism or oxidative stress (see supplemental section S1.2).
The groups were merged if the pairwise Pearson correlation of compound group exposure patterns was $cor > 0.5$.

Microarray analysis. The R-package `limma` [version 3.42.2 Ritchie et al. 2015] was applied to preprocess the microarray data. The microarray samples were loaded (`read.maimages()`), then background corrected (`backgroundCorrect()`) and normalised by a log-2 transformation. A confidence interval of $mean \pm 3 \times sd$ was used for sample wise interquartile ranges of log-2 expression to remove samples that were considered as outliers. The duplicated probes were summarised using the median of the respective expression values. Lowly expressed genes were identified relative to the average expression value of the negative control probe ($< mean + 2 \times sd$) and removed. The probe names were assigned to 11 518 gene names using the microarray platform and to 9887 unique Ensembl gene-IDs of zebrafish using Bioconductor's R package `biomaRt` [Durinck et al. 2005, version 2.42 .1] †.

* In *concentration addition* the mixture toxic potency is estimated by the sum of toxic potencies of all mixture compounds:

$$TU_{CA} = \sum_i TU_i \tag{2.2}$$

In the case of *independent action*, a correction of the additive effect is necessary, leading to the following mathematical formulation:

$$E_{IA} = 1 - \prod_i (1 - E_i) \tag{2.3}$$

,where E_i is the toxicity probability of the i-th compound when affecting a target singularly.

† see enrichment outcome: Data Chapter3 (TenStreams)/annotation_microarray.csv PW: PhD_SKraemer

2.1.3 Differential gene expression

Linear models of gene expression can mathematically describe exposure-dependent gene expressions. So, one can investigate each gene independently and determine whether it is differentially expressed under a perturbed condition compared to a control condition. Therefore, differential gene expression analysis (DEA) [Shi and Walker 2008] is conventional. DEA linear models gene expression across different pools of samples for each gene, respectively. The model designs vary from simple two-group comparisons to complex models with multiple experimental factors [Ritchie et al. 2015].

The differential gene expression analysis (DEA) was applied to the microarray samples to determine genes, which were significantly altered in their expression behaviour in sample groups under exposure-specific conditions, using the R-package `limma` [Ritchie et al. 2015, version: 3.42.2]. Four exposure scenarios reflected different xenobiotic assumptions (see figure 3.5). For the first three scenarios, stream site dummy covariates formed an initial linear design model, and the preprocessed log-2 transformed gene expression data fit those covariates (`lmFit()`). Then, the linear design models were refit (`contrasts.fit()`) based on the models of exposure scenarios with partition weights w. In the fourth scenario, the preprocessed expression data were control-normed, and the model was fit across all 69 samples (`lmFit()`) to the scaled but not normed compound group exposure patterns. Consequently, the following differential gene expression models were formulated:

1. *General treatment exposure scenario:* All stream groups ($w = 0.1$) vs. control groups ($w = -1$);
2. *Stream-wise exposure scenario:* Each stream group ($w = 1$) vs. control group ($w = -1$);
3. *Single compound exposure scenario:* Each single compound exposure pattern ($\sum_{i=1}^{10} w_i = 1$) vs. control group ($w = -1$). *
4. *Chemical group exposure scenario:* Each compound group exposure pattern ($\sum_{i=1}^{10} w_i > 1$) (control: $w = 0$)†.

* Single detected chemical compounds were considered indirectly with the respective (2.) stream-wise contrasts and not in (3.) single compound exposure scenarios.

† The given control information was compared to the treatment without considering the control group directly in the exposure pattern by control-normalisation of expression values ($\log_2 FC_{normalised} = \log_2 E_{sample} - mean(\log_2 E_{control})$). The initial idea was to consider single compound exposure scenarios without stream-wise pre-fit of gene expression. Therefore, each treated sample was weighted by a compound group exposure pattern in a gene-wise linear regression model. The highly correlated covariates and the restricted possibilities

A simple empirical Bayes model (`eBayes()`) moderated the standard errors for each of the four fit design models. The false discovery rate (FDR) was adjusted according to the method of Benjamini & Hochberg, and a significance threshold of $FDR \leq 0.05$ was used to identify significantly differentially expressed (DE) genes. The log-2 fold change of expression ($logFC$) and FDR were considered for follow-up analyses within the lists of exposure-related sets of DE genes.

2.1.4 Association rule mining

The unsupervised machine learning approach of *association rule mining* (AR) determines the rules, which describe co-occurrences of frequently combined items or item sets given the data. The approach originates from the market basket analysis.
In the following, $X \rightarrow Y$ is defined as a rule for the ANTECEDENT itemset X and the CONSEQUENT itemset Y considering a transaction set $T = \{t_1, t_2, \ldots, t_n\}$. Each transaction t_i contained a unique subset of items of one sample. The rule describes 'if itemset X is present in t_i, then the presence of itemset Y is likely.' Different frequency and co-occurrence measures may be used to filter possible association rules but the support and the confidence are utilised primarily. The *support* of a rule defines the frequency of the concatenated itemset of ANTECEDENT X and CONSEQUENT Y in the transaction set T:

$$support(X \rightarrow Y) = support(X \cup Y) = \frac{|t \in T; X \cup Y \subseteq t|}{|T|} \tag{2.4}$$

The *confidence* defines the conditional probability of CONSEQUENT Y in T given the presence of ANTECEDENT X in T:

$$confidence(X \rightarrow Y) = \frac{support(X \cup Y)}{support(X)} \tag{2.5}$$

Further measures might be employed to rank and prioritize association rules, such as the lift or the support ratio. The *lift* defines the ratio of the observed support to the expected support when X and Y were independent:

$$lift(X \rightarrow Y) = \frac{support(X \cup Y)}{support(X) \times support(Y)} \tag{2.6}$$

The *support ratio* is calculated as the relation between the supports of ANTECEDENT X and CONSEQUENT Y:

$$support\ ratio(X \rightarrow Y) = \frac{supp(X)}{supp(Y)} \tag{2.7}$$

of a linear combination of covariates made the grouping of compounds necessary.

The R-packages `aRules` [Hahsler et al. 2005, version 1.6-6], and `aRulesViz` [Hahsler 2017, version 1 .3-3], were used to perform association rule mining (AR). The itemset I comprised all expressed genes, detected compounds, compound groups, and stream sites (+ control). The microarray samples defined the transaction set $T = \{t_1, t_2, \ldots, t_n\}$. To identify the transaction subsets, the data were binarised first:

1. *Genes* were assigned to 1 if samplewise $|log_2FC_{normalised}| \geq 1$ *, else to 0;
2. The *stream* of which the respective water sample originated was assigned to 1, all others to 0;
3. *Compounds* were assigned to 1 if detected in the respective stream of a water sample, else to 0;
4. A *compound group* was assigned to 1 if ALL its compounds have been measured in the respective stream water sample, else to 0.

In this study, only pairwise association rules were computed. Thus, the ANTECEDENT represents an exposure-specific item and is either a compound, a compound group or a stream. The CONSEQUENT represents a gene item, which might be exposure-related up- or down-regulated.

The apriori algorithm was developed to effectively and time-efficiently identify frequent itemsets, which contain multiple items. Albeit the association rules had ANTECEDENT and CONSEQUENT itemsets of length one, the `aRules`-package allows considering the threshold settings for support and confidence. Furthermore, the ANTECEDENT and CONSEQUENT structure can be specified explicitly. As a result, calling the `aRules`-function `apriori()` determined association rules for co-occurring item pairs. To identify *frequent rules*, different measures of interestingness [Piatetsky-Shapiro 1991] — support, confidence and lift— were considered with respective filtering thresholds when applying the apriori algorithm. Frequent rules were grouped by exposure-specific ANTECEDENTS. The respective CONSEQUENT items comprise the set of genes related to this exposure. The measures of support, support ratio, confidence and lift were taken into account for follow-up analyses.

2.1.5 Weighted gene correlation network analysis

Network inference approaches are increasingly used in ecotoxicogenomics [Barel and Herwig 2018, Alexander-Dann et al. 2018]. Therefore, the gene expression from a transcriptomic

* control-normed expression data: $log_2FC_{normalised} = \log_2 E_{sample} - mean(\log_2 E_{control})$

analysis is compared pairwise, considering different treatment conditions.
In weighted gene correlation network analysis (WGCNA) [Langfelder and Horvath 2008], one generates a network based on the pairwise correlation of the gene expression patterns across transcriptomic samples. One assumes a scale-free topology for the network, which is given when the connectivity distribution of nodes follows a power-law - the majority of nodes has a small sum of edge weights, but a small set of nodes has vast sums of edge weights. For calculation, we chose Pearson's correlation. However, the biweight mid-correlation is known to retrieve more robust results than a Pearson correlation regarding outliers [Wilcox 2016, Langfelder and Horvath 2012]. In WGCNA, the preferred similarity measure for network generation is the topological overlap measure (TOM). A measure of adjacency considers gene pair by gene pair, whereas the TOM considers them concerning the other genes in the network. TOM counters the effects of spurious or missing edges [Yip and Horvath 2007] and is high if a gene pair shares the same neighbourhood of genes in a network.
The constructed network consists of genes as nodes and the pairwise similarity of genes as edge. Finally, a (i) fully connected, (ii) signed or unsigned, (iii) weighted or unweighted network is generated. Fully connected means that each node of the network is joint on all other edges in the neural network. The correlation value ranges between -1 and 1 in signed networks, where the negative values represent an anti-correlation, the positive values a positive correlation, and zero represents no correlation. In the case of an unsigned network, the correlation ranges between 0 and 1, and 0.5 represents no correlation. Each edge of a gene pair can be either binary-valued (existent or not) — unweighted — or continuously valued as the strength of correlation — weighted.
The constructed gene correlation network is grouped into modules based on k-means or hierarchical clustering. The module eigengene represents the modular gene expression pattern — the first principal component of gene expression across all genes. The correlation between a gene expression pattern and a module eigengene defines the module membership (MM). It describes the degree of intra-modular connectivity of a gene. A high value may indicate a hub gene that is strongly correlated to many other genes within the module.
No prior knowledge about the grouping of samples is considered for the network generation, which allows determining modules of co-expressed genes independent of *a-priori* knowledge of treatment or perturbations in the study design. Therefore, guilt-by-association correlations to gene expression determine external factors such as a chemical perturbation. One may prioritise modules via the correlation of a module eigengene and the pattern of such an external trait - the module-trait-correlation (MTC). On the other hand, the gene significance (GS) helps prioritise genes of interest for the studied perturbation, which correlates a gene expression pattern and an external trait pattern. The relevance of a gene of interest might be

exceptionally high in terms of perturbation if it is a hub gene and the module is associated with a biological functionality.

The weighted gene correlation network analysis (WGCNA) was conducted with the R-package `WGCNA` [Langfelder and Horvath 2008, version: 1.69]. A matrix of the preprocessed expression data set considered 69 columns of samples and 11 518 rows of genes. A soft threshold factor β was chosen to assume a scale-free topology in the generated network. To assume a scale-free topology, a soft threshold factor was chosen with the help of the transcriptional data set. The `WGCNA`-function `pickSoftThreshold()` calculated scale independence and mean connectivity for various soft-thresholding powers from 1 to 20. According to Zhang and Horvath [2005], an R^2-value for a power-law degree distribution should be above 0.8. The $mean(connectivity) \geq 100$ was chosen for soft threshold β. The constructed network was fully connected, signed, and weighted. The gene adjacency was calculated from expression data (`adjacency()`) using the biweight mid-correlation and the exponent β:

$$a_{ij} = \left(\frac{bicor(i,j) + 1}{2} \right)^{\beta} \tag{2.8}$$

A setting of `corOptions = list(maxPOutliers = 0.05)` was chosen as recommended in the WGCNA-tutorial. With the `WGCNA`-function `TOMsimilarity()`, the TOM was calculated:

$$TOM_{ij} = \frac{\sum a_{iu} * a_{uj} * a_{ij}}{\min \sum a_{iu}, \sum a_{uj} + 1 - a_{ij}} \tag{2.9}$$

Based on TOM, distinct groups of co-correlated genes — *modules* — were determined by applying hierarchical clustering (`hclust()`) and the R-package `dynamicTreeCut` [Langfelder et al. 2007, version 1.63-3] (`cutreeDynamic(method='tree',distM=1-TOM)` considering a minimal cluster size of $n = 30$ and a height cut-off of $c = 0.98$). Modules with highly correlated module eigengenes (`moduleEigengenes()`) were merged (`mergeCloseModules()`). The different exposure scenarios of DEA (see section 2.1.3) were considered in WGCNA as external traits. The MTC was calculated with exposure patterns and the module eigengenes. The correlation was significant if $|cor_{MTC}| \geq 0.3$ and $p_{MTC} \leq 0.05$. For individual exposure traits, the genes of a module with a significant MTC were considered as exposure-associated. Significant MTCs were investigated genewise with MM and GS. Both were determined based on biweight mid-correlation and were meaningful for visualisation and functional enrichment.

2.1.6 Method comparison

Functional enrichment analysis. Functional enrichment approaches are complementary follow-up analyses in omics-based studies to associate biological functions with measured gene

expression and inferred co-expressed gene clusters. Lists of detected genes are statistically associated with predefined gene sets like biological pathways or even chemical perturbations. Often, researchers choose overrepresentation analysis (ORA) and gene set enrichment analysis (GSEA) as functional enrichment approaches. GSEA [Shi and Walker 2008, Mootha et al. 2003] has the advantage of considering all genes in an experiment and not only those above a significance threshold.
Therefore, a ranked gene list S is retrieved from a gene expression analysis. If S is related to an investigated perturbation, the gene expressions should have high association scores to the geneset P of a related biological pathway [Shi and Walker 2008]. First, GSEA determines whether a more significant proportion of highly ranked genes in S to P occur than other genes. The enrichment score (ES) represents the extent P is overrepresented at the top or bottom of S. It is a running-sum statistic when walking down the list S Subramanian et al. [2005], increasing when a gene is in P and decreasing when not. Thus, the ES is based on a weighted Kolmogorov-Smirnov-like statistic Subramanian et al. [2005]. The maximum absolute value of ES is chosen.

Second, GSEA assesses the significance by permuting the class labels, which preserves gene-gene correlations and, thus, provides a more accurate null distribution [Subramanian et al. 2005]. If ES for S is greater than the enrichment scores for more than 95% of the randomly-permuted data sets, then the probability $p \leq 0.05$, and S is significantly enriched for P and S might occur not only by chance.

The biological meaning of the exposure-related gene sets was evaluated by conducting GSEA with the R-package `WebGestaltR` [version: 0.4.3 Liao et al. 2019, Wang et al. 2020] on KEGG [Release 88.2, Kanehisa and Goto 2000], Reactome [Version66, September 2018 Jassal et al. 2020] and Wikipathways [Release 02/10/2020 Pico et al. 2008] gene sets. The gene sets were annotated to zebrafish by the R-package `org.Dr.eg.db` [version 3.10.0 Carlson 2019] based on Ensembl gene-IDs. Previously, the log2FC- and adjusted p-values of DEA were used as the ranking measures in GSEA. In the present study, the genes were ranked by the measures of the respective approaches (DEA: $sign(\log(FC)) \times -\log_{10}(FDR)$; AR: $lift$; and $WGCNA$: $|MM| \times sign(GS) \times \log_{10}(p_{GS})$). A GSEA with all reference gene sets that contain at least ten genes was performed on the exposure-related gene sets for the respective exposure scenarios for all three approaches. An enriched gene set was viewed as significant if $FDR \geq 0.05$.

Data integration To investigate whether identified chemical-gene interactions were already known, respective data from the toxicological databases of CTD [version March 2019 Mat-

tingly et al. 2003, Davis et al. 2019] and STITCH [version5.0 Kuhn et al. 2008, Szklarczyk et al. 2016] were extracted (see table 2.1).

Table 2.1. ***List of toxicological reference data sets (including URL).*** *The following data were applied to evaluate the biological reasonability of retrieved chemical-gene interactions. Data were downloaded from CTD in* CSV *format on 2019-02-18. The chemical-protein data from STITCH for species Danio rerio had been downloaded in* TSV *format on 2019-04-17. (N: Number of samples per data set)*

Database	Relationship	N	Link
CTD	Chemical-Gene	2 127 796	ctdbase.org/downloads/CTD_chem_gene_ixns.csv.gz
	Chemical-Pathway	1 369 059	ctdbase.org/downloads/CTD_chem_pathways_enriched.csv.gz
STITCH	Chemical-Gene	74 619 879	stitch.embl.de/download/protein_chemical.links.v5.0/7955.protein_chemical.links.v5.0.tsv.gz

For each combination of a reference and a detected compound (mapped via MESH-ID), a list of exposure-associated genes was generated with all available zebrafish annotated genes. The R-package `biomaRt` [version: 2.42.1 Durinck et al. 2005] was used for annotation via ID mapping. The overlap of the reference gene sets was determined with the results of all three approaches for respective compound-dependent exposure scenarios (single compound or compound group) *.

Lists of enriched pathways per chemical from external sources.

Chemical-pathway interactions were extracted from CTD (see table 2.1) and reduced to the selected set of compounds by MESH-IDs. For each detected compound, a list of pathway-chemical interactions was prepared. The pathway-annotation was changed manually by replacing 'R-HSA-' to 'R-DRE-' for Reactome pathway-IDs and 'hsa' to 'dre' for KEGG pathway-IDs. A term was considered, if it contained at least one of the annotated genes on the microarray. A χ^2-test ($FDR \geq 0.05$) was applied to investigate statistical significance of the overlap to exposure-related results.

* From CTD 13 102 zebrafish-annotated chemical-gene interactions remained after mapping to microarray set and filtering to selected compounds. From STITCH 29 076 zebrafish-annotated chemical-gene interactions remained after protein-to-gene-reannotation, mapping and filtering.

2.2 Predicting exposure-related effects on molecular level

The second investigation is based on the knowledge represented in relations from pre-parsed PubMed abstracts stored in SemMedDB. The considered UMLS terminology is a standard ontology in biomedicine. In this thesis, the UMLS helped assign toxicological terms to levels of biological organisation and filter chemical-biomolecule interactions from SemMedDB. In this study, a deep learning model was trained to predict the relationship of a chemical-biomolecule pair based on current toxicological knowledge. The model was trained and validated in python using the deep learning packages `keras` [version:2.4.3 Chollet et al. 2015], `tensorflow` [version:2.5.0 Abadi et al. 2016], `scikit-learn` [version:0.0 Pedregosa et al. 2011], and `kerastuner` [version:1.0.3 O'Malley et al. 2019]. Model architectures for knowledge representation are considered, including layers of word embedding, and long-short-term memory, time-distributed layers and dense layers. The comparative toxicogenomic database was variously applied for data augmentation, linking chemical-gene interactions to higher biological levels and evaluating prediction results.

2.2.1 Input

The National Library of Medicine provides the **Unified medical language System (UMLS)**, which consists of three interconnected Knowledge Sources.

The **SPECIALIST Lexicon** is a syntactic and general English lexicon including many biomedical terms from a plethora of reference data and knowledge bases. The lexicon captures their syntactic, morphological, and orthographical information. These pieces of information are necessary for the SPECIALIST natural language processing tools, e.g. SemRep, which are also provided with the SPECIALIST Lexicon. Such Lexical Tools can be used, e.g., to generate the word indexes to the Metathesaurus.

The **Metathesaurus** is a huge vocabulary database applicable for multiple purposes and across multiple languages. Biomedical and health-related semantic concepts define the organisational structure of the Metathesaurus, and these are assigned to unique and permanent concept identifiers (CUI). Furthermore, the Metathesaurus contains the concept's various (synonym) names from many vocabularies *. A knowledge representation of the source vocabulary hierarchies is provided in `MRHIER.RRF`. The Metathesaurus contains non-synonymous relationships between concepts within the same source vocabulary and across different source

* The Metathesaurus is built from the electronic versions of various "source vocabularies" related to biomedicine, clinics, and health services.

vocabularies. These relations are stored in the file `MRREL.RRF`. The first-level hierarchical relations of `MRHIER.RRF` are also represented in `MRREL.RRF` as ISA relationships. Lexical terms in concept names appear in the SPECIALIST Lexicon, and the entire concept structure is represented in the file `MRCONSO.RRF`.

The **Semantic Network** contains broad subject categories — so-called *semantic types* — which categorize Metathesaurus concepts consistently. Furthermore, semantic relationships between types are provided by the Semantic Network. Therefore, all concepts in the Metathesaurus are assigned to at least one Semantic Type from the Semantic Network. This knowledge source provides a consistent concept categorisation in the Metathesaurus at a more general level. The file `MRSTY.RRF` links semantic types from the Semantic Network to the semantic concepts in the Metathesaurus.

The previously named rich release format (RRF) files were downloaded (on $2021-05-27$) from the UMLS download page * and used for UMLS annotation of lexical terms and semantic concept annotation to levels of biological organisation (LOBO) (see figure 2.1).

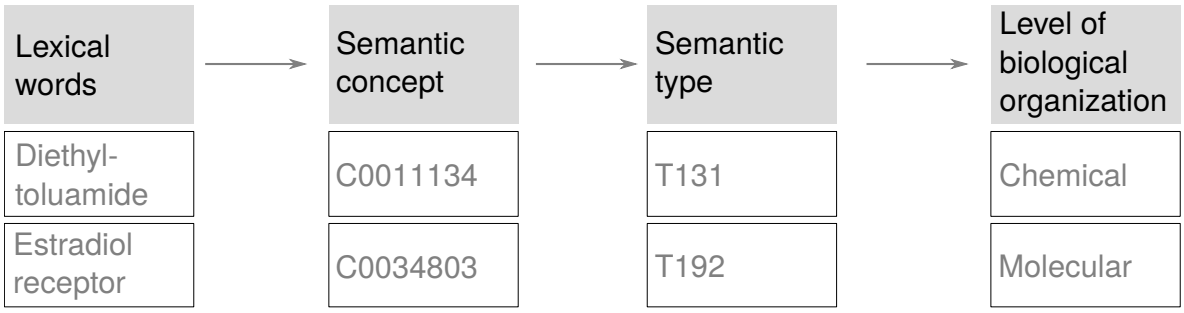

Figure 2.1. ***UMLS-Annotation of lexical words to levels of biological organisation.*** *In the UMLS Metathesaurus, biomedical words are assigned to semantic concepts. Furthermore, semantic concepts are classified and grouped into semantic types in the Semantic Network. In this study, the UMLS semantic concepts were assigned to levels of biological organisation based on their unique identifier (CUI), which structure lexical words to systems biological levels. In consequence, text-based relations can be assigned to toxicology-related terminologies.*

Kilicoglu et al. [2012] generated a large-scale knowledge resource called **SemMedDB**. It contained tables of semantic predications that were extracted from the titles and abstracts of all PubMed citations. SemMedDB contained the biomedical knowledge from all PubMed citations and was the here considered input. The *predication* data table was downloaded (on

* See also detailed data descriptions at webpage

2021−06−23) from the most recent version of the SemMedDB webpage [Kilicoglu et al. 2012]. A UMLS Terminology Services account was necessary for the download. The data set contained 112 796 186 semantic predications from 32 708 196 PubMed titles and abstracts, which had been parsed with the UMLS tool SemRep. Each predication in SemMedDB presented an UMLS-annotated triplet of the form <SUBJECT,PREDICATE, OBJECT>. The subject and the object were ontologically unified as *concept unique identifiers* (CUI). The predicate presents a relation type from the extended version of the UMLS *semantic network*. Predicates have a suffix 'NEG_' if SemRep recognizes a negation within the sentence.

The **comparative toxicogenomic database (CTD)** is frequently used in data integrative approaches in AOP-development and to extract chemical-gene or chemical-disease relations. This resource contains chemical interaction on different levels of biological organisation curated from empirical findings of omics-based exposure studies in an environmental health context and biological knowledge bases. The database consists of multiple relational data sets. The list of downloaded and selected data are shown in table 2.2.

Table 2.2. ***List of downloaded data from CTD (including URL).*** *This thesis considered relations with genes, chemicals, pathways and diseases (see top table) for input evaluation, model evaluation and data integrative analyses of prediction results. We used the downloaded vocabularies (see bottom table) to map gene and chemical names to UMLS concepts and levels of biological organisation. The following data were downloaded from the comparative toxicogenomic database (CTD) web page in* CSV *format on 2021-04-14. (N: Number of samples per data set)*

	Relationship	N	Link
T_{C2G}	Chemical-Gene	2 127 796	ctdbase.org/downloads/CTD_chem_gene_ixns.csv.gz
T_{C2P}	Chemical-Pathway	1 369 059	ctdbase.org/downloads/CTD_chem_pathways_enriched.csv.gz
T_{C2D}	Chemical-Disease	7 362 942	ctdbase.org/downloads/CTD_chemicals_diseases.csv.gz
T_{G2P}	Gene-Pathway	135 814	ctdbase.org/downloads/CTD_genes_pathways.csv.gz
T_{G2D}	Gene-Disease	84 780 962	ctdbase.org/downloads/CTD_genes_diseases.csv.gz

	Vocabularies	N	Link
T_C	Chemicals	174 328	ctdbase.org/downloads/CTD_chemicals.csv.gz
T_G	Genes	544 909	ctdbase.org/downloads/CTD_genes.csv.gz

2.2.2 Input preparation

Reduction to toxicogenomic chemical-biomolecule-relations. The input data were prepared from SemMedDB predications and were reduced to the unique and not negated triplets first.

The predication data set was reduced to relations, where the object concept represented a chemical and the subject concept a biomolecule. The UMLS Semantic Network links semantic types to all semantic concepts. For this investigation, the semantic types were additionally assigned to LOBOs manually. The file `MRSTY.RRF` contained all semantic types and was expanded by a column assigning a LOBO-like term of either *chemical* or *biomolecule* * and merged with the file `MRCONSO.RRF`. Thus, the semantic concepts of subjects and objects were assigned as chemical, biomolecule or another LOBO entity (see figure 2.1). According to the LOBO assignment, The predication data set was shrunk to predications with a chemical object and a biomolecule subject.

Furthermore, only two predicates were considered (*STIMULATES* and *INHIBITS*). Both relationships are representatives of substance interactions †. Finally, a relation was removed when *contradicting* — meaning that subject-object-pairs occurred with both relationship types in the given data. In the following, the abbreviation I designates the input data.

Input data split. The input I was split into training (I_T) and test data set (I_E). I_E comprised 10 000 relations of each predicate. Furthermore, all subjects and objects in the validation set occurred at least once in I_T. To perform 5-fold cross-validation, I_T was split into five equally sized and distinct subsets $I_{Ti}\{i \in \mathbb{N} | 1 \leq i \leq 5\}$.

Data augmentation. The hierarchical structures of the UMLS Metathesaurus were employed to determine parental terms of chemicals and biomolecules of the relationships in I.

* The terminology LOBO might be misleading when considering chemicals and biomolecules only. However, terms like *cell*, *tissue*, *organism* and *disease* were also assigned and are relevant, when considering UMLS terminology for future AOP tasks. Furthermore, some semantic types are more clinical-related and no LOBO term was assigned at all.

† The both relationship types were defined in [Kilicoglu et al. 2011]:

INHIBITS: Decreases, limits, or blocks the action or function of (substance interaction). (Example: In recent studies, the BDNF expression was reduced by typical neuroleptics; Antipsychotic Agents [Pharmacologic Substance],INHIBITS,Brain-Derived Neurotrophic Factor[Biologically Active Substance])

STIMULATES: Increases or facilitates the action or function of (substance interaction). (Example: Candesartan therapy significantly reduced inflammation and increased adiponectin levels; ¡candesartan[Pharmacologic Substance],STIMULATES,Adiponectin[Amino Acid,Peptide, or Protein]¿)

A ISA relation in `MRREL.RRF` is a directed relation edge in the hierarchical structure of semantic concepts. Such a relation $A \rightarrow^{ISA} B$ associate a semantic concept A to a superior semantic concept B — the *direct parental semantic concept*. For example, the Metathesaurus relation $Diclofenac \rightarrow^{ISA} NSAID$ shows that the chemical *Diclofenac* belongs to the pharmaceutical group of *non-steroidal anti-inflammatory agents* (NSAID). Thus, all (biomolecule) semantic concepts in I, associated with *Diclofenac*, were also associated with *NSAID*. In the Metathesaurus, semantic concepts may have multiple or also no direct parental term.

The parental semantic concepts for subjects and objects augmented information in I. Two versions of data augmentation were applied (see figure 2.2).

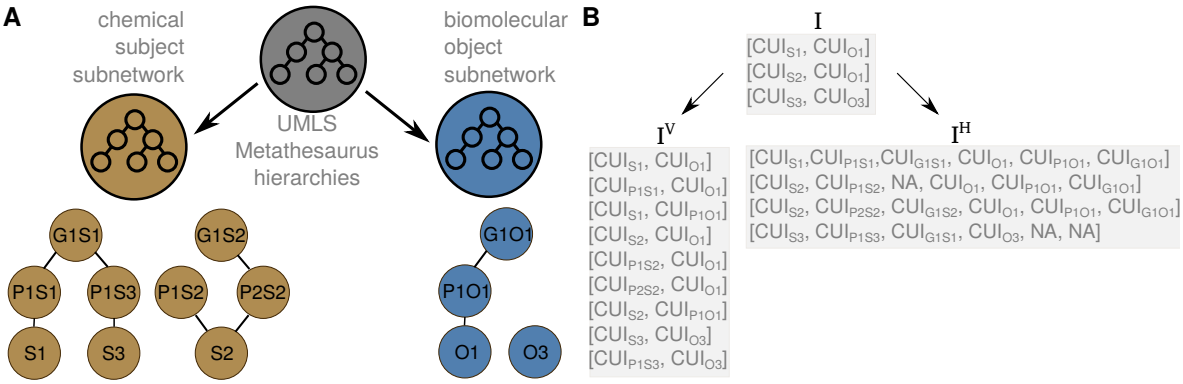

Figure 2.2. ***Input data augmentation for chemical biomolecule relations.*** *The input of UMLS-annotated chemical-biomolecule relations were expanded.* ***A:*** *The knowledge of the UMLS Metathesaurus was retrieved to determine hierarchical relations (*ISA*-relations in `MRREL.RRF`) between chemical (ocher) or biomolecular concepts (blue). Based on the respective hierarchical ordered networks, parental and grandparental terms of a concept can be determined.* ***B:*** *Considering the (grand-)parental terms allow increasing the number of relation samples (vertical augmentation) and to elongate the sequence length of samples (horizontal augmentation).*

For a **vertical augmentation**, all recombinations, with one direct parental concept of either the subject's or the object's semantic concept instead of the subject or object itself, were added to I. Thus, one chemical-biomolecule-relation in I might result in multiple relations with parental terms, e.g. one relation might occur with subject's parental term and another with object's parental term. Furthermore, due to the various sources in the Metathesaurus, each subject or object semantic concept might have multiple parental terms. After vertical augmentation, contradicting relations might occur as I might already contain some parental concepts. These contradicting relations were removed. I^V designates the vertical augmented version in the following.

The training-test-validation-split of I remained in I^V. In order to completely separate I_E^V from I_T^V, the training relations were removed if $X \in I_T^V$ & $X \in I_E^V$. As in I, separate subset splits of $I_{Ti}^V\{i \in \mathbb{N} | 1 \leq i \leq 5\}$ were considered. If identical relations were determined across subsets, all but one randomly chosen duplicate were removed.

The **horizontal augmentation** of a chemical-biomolecule relation increased the *term lengths* of subject and object from 1 to 3. For the semantic concepts of a chemical C_i and a biomolecule B_i, the direct parent P_{Ci} or P_{Bi} and the second-order parent G_{Ci} or G_{Bi} (grandparent) were determined. Each recombination $< C_i, P_{Ci}, G_{Ci} \rightarrow^{RELA} B_i, P_{Bi}, G_{Bi} >$ was considered a horizontally augmented relation of $C_i \rightarrow^{RELA} B_i$. Each semantic concept might have none or multiple parents and thus also grandparents. If a concept had no parent or grandparent, a masked token *NA* was used respectively to keep the length of all relations equal.

I^H designates the horizontal augmented version of I in the following. The training-test-validation-split of I was transmitted to I^H.

2.2.3 Deep learning models

Word embedding approaches are used to learn a dense and low-dimensional representation from a large and unlabeled corpus of text-based information and shall efficiently capture the semantics of words. Therefore, words, phrases or substrings of words are transformed into vectors of real numbers. It is assumed that a word embedding allows a generalization of semantic meanings of interest given the input data. The words that are semantically used in similar ways have similar representations, which capture their meaning. Usually, the word embedding vector length is much smaller than the size of the vocabulary. From a computational point of view, the dense and low-dimensional vector space is superior to the dimensions required for a sparse word representation (e.g. one-hot-encoding) of vocabularies with thousands or millions of words. Densely distributed and low-dimensional vector representations of words are more suitable than one-hot encoded representations. Word embedding model architectures may be expanded with further neural network layers, e.g. feed-forward networks, recurrent neural networks or convolutional neural networks.

A **feed-forward network** or multilayer perceptron may be considered as the base of many state-of-the-art neural networks (see figure 2.3 A). It contains an input, a hidden and an output layer. Each hidden and output neuron employs a non-linear activation function. The information flows through the layers and is transformed by the functions without any feedback connection – feed-forward. In this process, the non-linear activation functions have weighting parameters, which get optimised over training epochs through the supervised learning process

of backpropagation *. In `Keras` and the following of this thesis, those networks are regarded as dense neural networks or fully connected networks.

Recurrent neural networks (RNN) are frequently used in deep learning models in the biomedical context. The networks contain feedback loops allowing information to remain within the network. Thus, such models have a *'memory'* and can handle sequential data. Within an input sequence, the information from prior positions influences the current position and even the output. Thus, the network uses information of the recent position and the internal state from the previous position(s) as input. RNN can be considered in two main architectures — either with circular connections or as a deep feed-forward network (see figure 2.3 B). In the unrolled case, a new cell (see figure 2.3 B) in the recurrent network for each word of a sequence is considered. The weights are identical across RNN cells. According to `Keras`, unrolling may decrease running time for the price of higher memory usage. Furthermore, *bidirectional* RNNs can consider the future — posterior positions influence the current position — during output estimation.

For longer sequences, regular RNNs may be limited to model the dependencies between sequential steps separated by numerous others — they have a limited memory. The so-called vanishing gradient problem describes how small weights got eliminated due to manifold multiplications across time steps. In consequence, the weights of earlier layers have no significant changes, and the network forgets long-term dependencies.

Long-short-term memory (LSTM) is a kind of RNN that uses feedback loops of a previous step in the timeline or sequence $t-1$ for the output of the recent step t (see figure 2.3 C). LSTM networks allow solving the vanishing gradient problem.

In this thesis, we implemented two **model architectures** (see figure 2.4) with an initial word embedding layer (`tensorflow.python.keras.layers.embedding()`). The here applied sample wise output of the embedding layer was a sequence of two word embedded vectors. Then, a flatten layer (`tensorflow.python.keras.layers.flatten()`) followed to transform the sequential word embedding output into a one-dimensional vector. Next, the flattened output was reduced by a sequence of dense layers (`tensorflow.python.keras.layers.Dense()`) with a rectified linear unifier (ReLU) [Agarap 2018] activation function. The final dense layer ($N_{neurons=1}$) had a sigmoid activation layer (`tensorflow.python.keras.layers.Dense(1, activation='sigmoid')`) and classified the selected target — the type of relationship between a chemical and a biomolecule — in a binarised manner.

* The backpropagation is based on the least-mean-square errors between estimated and expected output. The weight parameters of the non-linear functions get changed in dependence of the error and the learning rate.

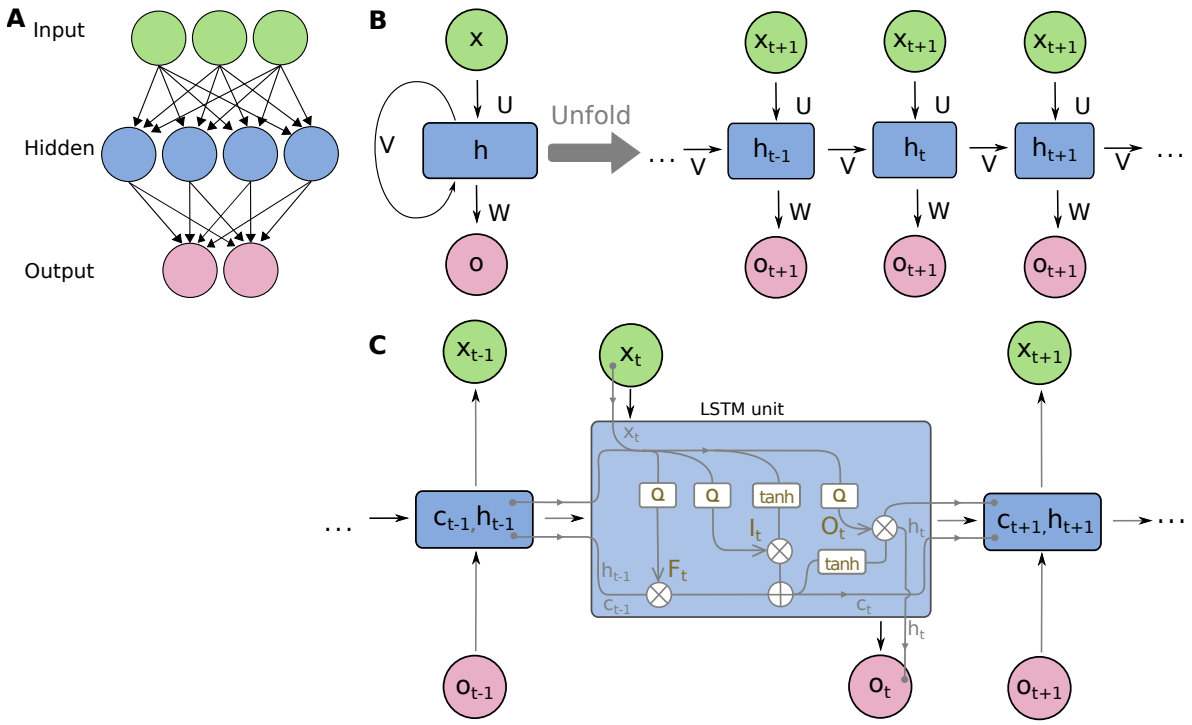

Figure 2.3. ***Selected architectures of neural networks.*** *In principle, a neural network consists of an input (green), a hidden (blue) and an output layer (red).* ***A:*** *In the feed-forward neural network, all input neurons are connected to all hidden neurons, and all hidden neurons are connected to all output neurons. During training, the weights per connection are adapted through backpropagation.* ***B:*** *A recurrent neural network contains feedback loops allowing information to remain within the network. The information of the previous part of the sequence can be used in the current parts. An unrolled architecture removes circular connections and structures them as a deep feed-forward neural network* ***C:*** *A long-short-term memory is one type of recurrent neural network. The new sequential value x_t is concatenated first to the previous output from the cell h_{t-1}, and the combined input is tanh-transformed. The forget gate (F_t) regulates the loop and helps the network learn which state to remove. The same input is sigmoid-activated through an input gate (I_t) and multiplied with the tanh-transformed input. Thus, I_t removes unnecessary elements of the combined input vector. In the following, an internal state of previous information c_{t-1} is added to the input data to create an effective recurrent layer and to reduce the risk of vanishing gradients. The final tanh-function of the output gate (O_t) determines which values to consider as output h_t. (Figure is a concatenation of adapted versions of artificial neural network (wikimedia), Recurrent neural network unfold (wikimedia), Long-short-term memory (wikimedia))*

Before the flatten layer, the second model architecture contains an additional LSTM layer (`tensorflow.python.keras.layers.LSTM()`). Furthermore, adding a time distributed dense layer (`tensorflow.python.keras.layers.TimeDistributed()`) after the RNN is recommended for `Keras` applications with LSTM. This wrapping layer allows applying the identical fully connected layer to the separate sequential output vectors. In this study, the LSTM layer and the following time distributed dense layer had the same number of neurons.

The simpler architecture with word embedding and dense layers is designated as A, whereas the one with LSTM is designated as B in the following. Different parameter settings were compared for both model architectures to determine a suitable setup for the study-specific learning task. $\bigcup_{i=2}^{5} I_{Ti}$ were chosen as training data and I_{T1} as validation data to determine the two suitable model architectures.

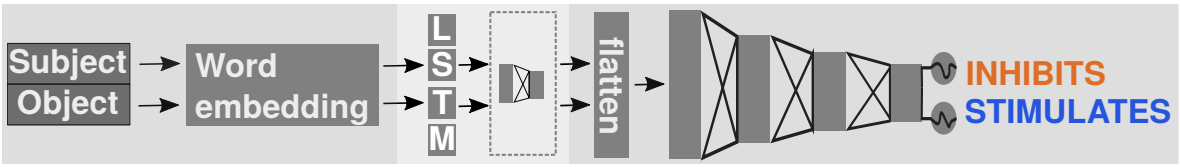

Figure 2.4. ***Deep learning model architecture with word embedding and long-short-term memory.*** *Both model architectures use subject-object-pairs of predication triplets as input and contain an initial word embedding layer. In model architecture B only (light grey area) a long-short-term memory (LSTM) and a time distributed dense layer is considered additionally. In both model architectures a flatten layer follows, before vector size gets reduced in a subsequent feed-forward series. The output gets sigmoid-transformed to finally predict the predicate target.*

The preprocessed predications in I_T were used as input (subject and object) and as targets (predicate) for training and validation. All models considered input sequences consisting of two positive integers by applying the function `sklearn.preprocessing.OneHotEncoder()` *. The predicates of the triplets were binarised with `sklearn.preprocessing.LabelEncoder()` and used as targets (*INHIBITS*: 0; *STIMULATES*: 1).

The learning objective was to minimise the loss (binary cross-entropy) and was examined

* Such sequences of indices represented a memory saving representation of a sequence of one-hot-encoded word vectors. The integers ranged from 0 to $len(vocabulary)-1$ and represented the position in the vocabulary hash, which was generated applying one-hot-encoding.

using the validation set. The binary cross-entropy was defined as:

$$Binary\ cross\ entropy = -\frac{1}{len(validation)} \sum_{i=1}^{len(validation)} (y_i * log(\hat{y_i}) + (1 - y_i) * log(1 - \hat{y}_i)) \tag{2.10}$$

The value y_i equals the label of the i-th sample in the considered validation data and $\hat{y}_i$ the predicted probability for the valid label. Next to the loss function *, we tracked the binary accuracy for validation data.
Before each epoch, the input data were shuffled to circumvent training of the sample order. Each model was trained for a maximum of 1000 epochs. The `Keras` callback-function `EarlyStopping()` was implemented to train models no longer as necessary. It allowed stopping training if the validation loss decreased less than 0.01 across a duration of 20 epochs. A batch size of 1000 was chosen.

For model architecture A, we compared different word embedding vector sizes. The sparse word representation with more than 40 000 words was reduced by word embedding to sizes of 5000, 2500, 1000, 500 and 100.
Model architecture B has an additional LSTM layer after the word embedding. We compared various parameter settings for the number of neurons equal to 100, 50, 25 and 10.
After selecting a word embedding size for model architecture A and a number for LSTM-neurons in B, a hyperparameter tuning (`kerastuner.RandomSearch()`) was applied, respectively. The model architecture of A was tuned first, and respective parameters were not tuned again in B. Various hyperparameters were tuned, e.g. L2-regularization in layers, dropouts, types of activation, number of dense layers and modifications in the LSTM layer. Though not all layer parameters were tuned or taken into further consideration, hundreds of recombinations were still considered per model architecture. A time-efficient hyperparameter tuning was performed by randomly selecting one hundred parameter recombinations. The recombination with the lowest loss value was considered as selected model architecture for A or B, respectively.

A **model evaluation** was applied with the test data I_E to compare trained models with different architectures or parameter settings. For each predication in I_E, it was determined whether they were *falsely predicted with STIMULATES* (FS), *truly predicted with STIMULATES* (TS), *falsely predicted with INHIBITS* (FI) or *correctly predicted with INHIBITS* (TI). The considered performance measures were accuracy, precision, recall and F1-score and were calculated with `sklearn.metrics.classification_report()`. The following formulas

* Note, that the default for binary cross-entropy tracking in `Keras` was applied calculating the sum of averages of the per-sample losses.

describe the measures' calculation:

$$Accuracy = \frac{n_{TI} + n_{TS}}{n_{TI} + n_{TS} + n_{FI} + n_{FS}} \tag{2.11}$$

$$Precision_R = \frac{n_{TR}}{n_{TR} + n_{FR}} \tag{2.12}$$

$$Recall_R = \frac{n_{TR}}{n_{TR} + n_{F\bar{R}}} \tag{2.13}$$

$$F1_R = 2 * \frac{Precision_R * Recall_R}{Precision_R + Recall_R} \tag{2.14}$$

The index R represents one relationship type (*STIMULATES* or *INHIBITS*), whereas $\bar{R}$ represents the contrary relationship type.

Model comparison. The selected model architectures A and B were employed for a 5-fold cross-validation training and were performed with I_T, I_T^V and I_T^H. For the i-th fold the training data comprised of $\bigcup_{j=1}^{5} I_{Tj}$, where $j \neq i$ and validation data I_{Ti}. For a validation across folds and across architectures, the performance measures of binary cross-entropy and binary accuracy were calculated per epoch for the respective validation set. From each 5-fold cross-validation, the model with the lowest loss and ideally with the highest accuracy was chosen.
All six training outcomes were comparatively validated with the unseen data of I_E, I_E^V or I_E^H. The evaluation was performed with the identical performance measures as described in model selection before. Based on the calculated measures, the performances of model architectures A and B were compared.

2.2.4 Toxicogenomic application

In contrast to the chemical-gene interactions T_{C2G} from CTD, the input I considered not only genes but all biomolecules listed in literature and annotated by the UMLS. Therefore, the coverage of chemical-gene relations from I in CTD was determined first. Two other exposure-related levels were considered (see figure 4.3). Next to the chemical interactions with genes, those with pathways and diseases were compared to adapted versions of I. For a comparison on higher LOBO, respective CTD data sets had to be preprocessed, and the chemical-biomolecule interactions in I prepared as chemical relations on higher levels.
Furthermore, the prepared and UMLS-annotated data set of chemical-gene interactions from

CTD was considered to evaluate the models based on curated toxicogenomic knowledge next to literature-retrieved predications in the test sets I_E, I_E^V and I_E^H.

The from CTD downloaded **chemical-gene interactions** were preprocessed to become comparable to SemMedDB predications. All relationships with the suffix *increased* were assigned as *STIMULATES* and with *decreased* as *INHIBITS*. Relations with the suffix *affects* were removed. After reducing the 134 relationship types in CTD to two, some chemical-gene interactions with multiple relationships resulted in contradictions. Contradictory chemical-gene pairs were removed. Then, the chemicals and genes were annotated to semantic concepts. Therefore, the chemical and gene annotations from CTD were fitted to semantic concepts with the help of the UMLS lexical data set. With this, a chemical was annotated to a concept and assigned to its CUI, not only when the GENE SYMBOL or CHEMICAL NAME directly matched to listed lexical terms, but also for the entries in columns DESCRIPTION or SYNONYM in the annotation files of CTD. With the help of the LOBO assignment, it was ensured, that a chemical concept is also annotated as CHEMICAL and a gene as BIOMOLECULE. T_{C2G} designates the preprocessed set of UMLS-annotated chemical-gene interactions in the following of this thesis.
To achieve also a possible comparison for models trained with either I_T^V or I_T^H, we also augmented T_{C2G} in the same manner as described in section 2.2.2.

To prepare **chemical-pathway interactions for the SemMedDB data**, we merged the gene-pathway relations (T_{G2P}) from CTD with the chemical-biomolecule relations in I. Therefore, the UMLS-annotated gene annotation assigned genes to UMLS CUIs, which allowed linking I. Consequently, the merge assigned to each object CUI the respectively associated pathways. The joined data set was reduced to unique pairs of chemical CUIs and pathways and is declared as I_P.
Furthermore, **chemical-pathway relations from CTD** (T_{C2P}) were considered as a reference to determine the toxicogenomic coverage in I_P. The chemicals were UMLS-annotated to CUIs with the CTD chemical annotation to make T_{C2P} comparable to I_P. The unique set of chemical-pathway interactions in T_{C2P} was furthermore reduced to relations, which contain a chemical CUI and a pathway term, that both were present at least once in I_P. Consequently, I_P was reduced vice versa.

The **chemical-disease interactions for the SemMedDB data** were prepared with I and CTD resources similar to chemical-pathway interactions before. The chemical-biomolecule relations in I were expanded with gene-disease relations (T_{G2D}) from CTD. Similar to T_{G2P}, the gene names were annotated to UMLS CUIs and merged with the UMLS-annotated version of T_{G2D}. Thus, the output assigns to each object CUI the respectively associated diseases.

The combined data set, designated as I_D, was reduced to unique pairings of chemical CUIs with diseases.
The **chemical-disease relations from CTD** (T_{C2D}) with UMLS-annotated chemical CUIs was considered as a reference to determine a toxicogenomic coverage in I_D. Both, T_{C2P} and I_D, were reduced to relations, which contain a chemical CUI and a disease term, that both were present at least once in the respectively other data set.
After generating comparable data sets from CTD and SemMedDB, the overlaps of I_P to T_{C2P} and I_D to T_{C2D} were determined (see figure 4.3).

A **toxicological evaluation** of the models trained with I was applied to compare their ability to predict recent toxicogenomic knowledge. Therefore, for all chemical-gene pairs in T_{C2G}, each trained model was applied to predict the toxicogenomic relationship type. FS, TS, FI and TI were determined and compared across models based on predictions and relationship types known in CTD. For each model, the measures of binary accuracy, precision, recall and F1-score were calculated.

We performed an **overrepresentation analysis** (ORA) to determine enriched pathways applying the R-package `WebGestaltR` [Version:0.4.3 Wang and Liao 2020]. The input and references in ORA had a unified CUI annotation. The pathway-based genesets were retrieved from the reference data T_{C2P}. A set of biomolecule concepts was significantly enriched to a CTD reference when $FDR \leq 0.05$. The enriched pathways were compared to known associations to the chemical concept in T_{C2P} by determining the coverage. In consequence, the functional enrichment allows considering predictions of chemical-gene interactions on pathway level and evaluating their coverage in recent toxicogenomic and exposure-related knowledge.

Chapter 3

Method comparison to link complex stream water exposures to effects on the transcriptional level

Researchers used omics-based approaches to understand adverse effects induced by mixtures of contaminants. In the study of this chapter, we examined transcriptomic data of hepatic tissue from fathead minnows acutely exposed to mildly polluted Minnesotan stream waters. We applied differential gene expression analysis (DEA), association rule mining, and weighted gene correlation network analysis (WGCNA) and identified potential driving (mixtures of) contaminants. In addition, we identified biologically meaningful and reliable attributions of compounds to xenobiotic effects with functional enrichment and integration of external reference bases.

The concentrations of detected contaminants inferred the expectation of mild acute toxicity in selected stream waters. Co-correlations of compounds and mixture effects occurred, and we had to deal with them — the numerical exposure data allowed us to investigate exposure from different perspectives. In particular, compound groups were determined to reduce the co-correlation of exposure patterns.

Limitations of exposure data were the size, a low toxic ratio and the small set of considered compounds. They disabled us to disentangle specific chemical effects and assignment of biological effects to particular contaminants. Nevertheless, in DEA and WGCNA, functional enrichment ranked by application-specific metrics identified biologically meaningful terms to xenobiotic stress and immune response for compound groups. These enriched sets overlapped significantly with compound-pathway associations in CTD.

3.1 Background and motivation

Thousands of chemicals from various anthropogenic origins are released into the environment and pollute all sources of influx to surface water. Thus, contaminants remain detectable in surface waters [Bradley et al. 2019] and affect the aquatic ecosystem [Blackwell et al. 2019]. Organisms in surface water bodies are affected by mixtures of often lowly concentrated contaminants. Next to more specific effects due to toxic modes of action, perturbation due to environmental relevant mixtures may induce adverse stress responses. There is a need to understand better the combined effects of chemicals in the environment Kortenkamp and Faust [2018].

Studies have investigated the chemical exposure to original surface water samples to understand sublethal effects in the aquatic environment [e.g. Perkins et al. 2017, Schroeder et al. 2017]. Effect-based examinations conducting bioanalytical *in-vitro* experiments have been used to identify biological interactions for a pre-selected but small set of biomarkers [McGovarin et al. 2018, Pérez et al. 2018, Dale et al. 2019, Calderón-Delgado et al. 2019, Perkins et al. 2017]. On the other hand, omics-based approaches help investigate many molecular endpoints on a high-throughput level and elucidate effects in complex mixtures and environmental samples. One well established environmental model organism is the fathead minnow (FHM) [Ankley and Villeneuve 2006]. For example, transcriptomic investigation of biological effects after disruptive endocrine exposures and environmental relevant mixtures have used this fish frequently [e.g. Rodríguez-Jorquera et al. 2019, Feswick et al. 2017, Zare et al. 2018, Garcia-Reyero et al. 2011]. For example, recent ecotoxicological studies have measured transcriptional changes in gene expression after an exposure treatment [Li et al. 2020, Ewald et al. 2020, Perkins et al. 2017, Martinović-Weigelt et al. 2014, Schroeder et al. 2017]. Some studies have investigated low anthropogenic exposures and have detected molecular effects before an adverse effect occurred [e.g. Perkins et al. 2017, Schroeder et al. 2017, Wiseman et al. 2013]. In transcriptomic studies, sources of background noise have to be considered, e.g. uncertainties due to low chemical perturbations below detection levels. Thus, transcriptomics helps examine surface water samples to elucidate the complex mixtures.

Different computational methodologies have been successfully applied to identify and prioritise biological effects in omics data. Identifying chemical drivers of adverse biological effects in environmental mixtures is a central challenge in ecotoxicogenomic research [Miracle et al. 2003]. One meaningful computational approach is differential gene expression analysis (DEA), comparing treated groups to control groups. A differential biological effect can be determined for a present environmental mixture perturbation in site-specific comparisons. DEA has been

frequently used in adult male FHM omics studies to investigate xenobiotic effects on hepatic tissue after exposure to surface water samples influenced by wastewater [e.g. Sellin Jeffries et al. 2012, Martinović-Weigelt et al. 2014, Rodriguez-Jorquera et al. 2015, Schroeder et al. 2017]. For example, the determination of the transcriptional changes in liver tissue has helped identify sources of environmental stress. As a result of this, studies have assessed the effectiveness of cleanup [e.g. Costigan et al. 2012, Martinović-Weigelt et al. 2014, Arstikaitis et al. 2014], and the best management practices (in treatment plants) [e.g. Rodriguez-Jorquera et al. 2015, Vidal-Dorsch et al. 2013, Berninger et al. 2014] or the remediation methodologies [Wiseman et al. 2013]. In addition, some studies applied frameworks to differentiate chemical and site-specific variations in transcriptional effects [Berninger et al. 2014] and to distinct biological effects of one stressor from another [Schroeder et al. 2017, Perkins et al. 2017]. However, DEA is only one approach to examine transcriptional expression due to chemical perturbation in environmental mixtures.

Furthermore, toxicologists applied network inference [e.g. Ewald et al. 2020] and machine learning [e.g. Krämer et al. 2020] to group genes impacted by exposure to environmental contaminants unsupervised. Other studies combined DEA with network inference [Degli Esposti et al. 2019, Sutherland et al. 2018, Barel and Herwig 2018] or machine learning [Acharjee et al. 2016, Ornostay et al. 2013]. For example, the network inference approach of weighted gene correlation network analysis [Langfelder and Horvath 2008] (WGCNA) is helpful for omics data sets to identify groups of highly correlated genes. Moreover, it correlates external information [Langfelder and Horvath 2008] like chemical exposure to modules. Toxicologists have already applied WGCNA [Sutherland et al. 2018, Orsini et al. 2018, Maertens et al. 2018, Ewald et al. 2020]. For example, Ewald et al. [2020] identified ecologically relevant co-expressed gene modules using transcriptional data of liver from adult FHM considering 38 single compounds and complex environmental mixture exposures.

Recent studies have investigated exposures [Barrera-Gómez et al. 2017, Kapraun et al. 2017, Santos et al. 2020] and determined alterations in gene expression [Creighton and Hanash 2003, Mallik and Zhao 2017, Karel and Klema 2007] with the application of association rule mining (AR). This unsupervised machine learning approach determines rules describing co-occurrences of frequently combined items in given data. Furthermore, biomedical researchers analysed omics data applying AR [e.g. Chen et al. 2019, Toti et al. 2016, Mallik and Zhao 2017, Lakshmi and Vadivu 2019]. However, researchers in ecotoxicological omics research have not yet considered AR.

When causally linking chemical exposure to biological effects with omics-based approaches, environmental toxicologists are challenged by distinguishing direct exposure effects from in-

direct xenobiotic responses Scholz et al. [2021]. Functional enrichment methodologies help associate treatments — such as exposures — to molecular effects. In eco-toxicogenomic studies, functional annotation to Gene Ontology terms [e.g. Wiseman et al. 2013], overrepresentation analysis [e.g. Rodriguez-Jorquera et al. 2015], gene set enrichment analysis [e.g. Perkins et al. 2017, Schroeder et al. 2017] or subnetwork enrichment analysis [e.g . Feswick et al. 2016, Rodríguez-Jorquera et al. 2019] were conducted alone or in combination. Also, curated gene set databases help project ontologies of biological functions, for example, gene ontology [Gene Ontology Consortium 2004, Carbon et al. 2019] or signalling pathway references like KEGG [Kanehisa et al. 2004]. Some studies integrated knowledge from toxicogenomic databases to identify known chemical-gene interactions [Holth et al. 2008, Perkins et al. 2017, Schroeder et al. 2017] or compounds as potential upstream regulators [e.g. Zare et al. 2018, Martinović-Weigelt et al. 2014]. One frequently used source is the comparative toxicogenomic database (CTD) [Mattingly et al. 2003, Davis et al. 2017], also providing toxicogenomic knowledge for aquatic model species. For example, Perkins et al. [2017] presented one way to link exposure to adverse outcomes and envision a practical approach to investigate chemical mixture interactions at low concentrations. They identified estrogen activity in Lake Superior Bay waters of Duluth by differential gene expression in exposed fish associated with CTD-references for the upstream regulator estradiol.

The present study identified links between complex chemical exposures and molecular effects on the gene and the pathway level by applying three computational approaches. The research objective was to investigate whether DEA, AR or WGCNA were suitable to uncover the individual chemical effects of complex mixture exposures and whether the approaches could determine reliable attributions of potentially adverse effects to chemical drivers.
The Minnesota Pollution Control Agency (MPCA) had contributed environmental data of chemical exposure in ten streams and biological effects with an *in-vitro* assessment and *DNA* microarrays [Ferrey et al. 2017]. Although the detected low chemical concentrations might not affect any direct acute biological adverse outcome, it might induce xenobiotic stress measurable on the transcriptional level. The present study investigated the xenobiotic effects in liver tissue of FHM after acute toxicity in mildly polluted streams. It was examined whether the three applied computational approaches were applicable in ecotoxicology and whether they allowed disentangling chemical exposure to biological effects. The application of functional enrichment validated the toxicological role of identified chemical-gene interactions on the pathway level. The trustworthiness of the selected approaches was proven by the exposure-related representation of genes and the coverage of biological terms regarding current toxicological knowledge. We examined four different exposure scenarios in the chosen approaches. Conse-

quently, this study highlighted limitations and abilities of mining, analysing and integrating omics and exposure data when considering complex mixtures of lowly concentrated chemicals.

3.1.1 Workflow

Figure 3.1 presents the workflow * performed in the R statistical programming language [version 3.6.1 R Core Team 2020].
Dalma Martinović-Weigelt, a collaboration partner in Minnesota, United States, had provided three data sets 2.1.1. For ten small streams in Minnesota, (1) quantitative chemical exposure data of 146 pharmaceuticals and chemicals of concern in the US, (2) two *in-vitro* cell assay measurements for endocrine activity, and (3) gene expression data in liver tissue of fathead minnows after acute exposure to stream water ($n_{Samples,treated} = 64$; $n_{Samples,control} = 7$) had been measured. Detailed information on data sampling, microarray experiments and reports of initial results had been described by Ferrey et al. [2017].
We transformed the measured chemical concentrations into toxic units, and preprocessed the raw microarray data (see 2.1.2). Exposure-related gene interaction sets were generated using three computational methods based on the given exposure and gene expression data (Differential gene expression: see 2.1.3; Association rule mining: see 2.1.4; Network inference: see 2.1.5). We validated the exposure-associated gene results with functional enrichment to biological pathways and chemical reference sets from external databases for each method, respectively (see 2.1.6).

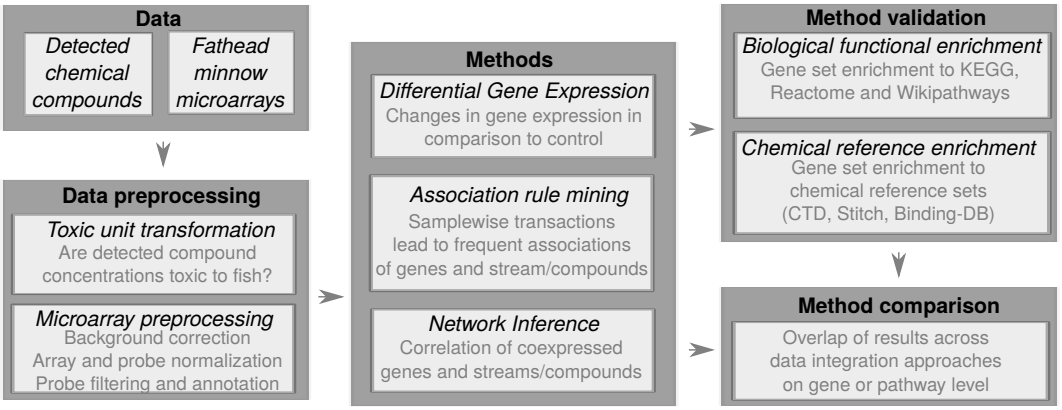

Figure 3.1. ***Project workflow.*** *The quantitative chemical exposure data and the gene expression data were preprocessed. Data were linked by applying three different bioinformatic methods. Afterwards results were functionally analysed and compared.*

* Data and code are available here: https://nc.ufz.de/s/emqxbigeWYPSnKp (*'Data Chapter3 (TenStreams)'* and *'Code Chapter3 (TenStreams)'*) with the following password *PhD_SKraemer*

3.2 Results

3.2.1 Data preprocessing

The **detected concentrations of stream water contaminants** might induce mildly acute toxicity in fish. The selected stream sites were a subset of randomly selected sites representing multiple ecological relevant regions and the diverse exposure scenarios in Minnesota. The selected compounds (see supplemental table S2-1) were common contaminants in investigated sites based on prior occurrence studies and might be present. Out of 146 screened contaminants, 27 were detected in at least one of the ten selected streams in Minnesota (see 3.2) *. Table 3.1 lists the compounds that were detected only once. Additionally, two compounds were detected via *in-vitro* assessment as described in the methods section. The chemical exposure by the number of detected compounds, and the sum of toxic units varied substantially across streams (see figure 3.2 A and B). Per stream, 1 to 18 compounds had concentrations above the compound wise detection limit. Three streams contained more than ten compounds. The stream-wise sum of toxic units was not directly related to the number of detected compounds. Consequently, the streams with many detected compounds had not necessarily high toxic effects. However, the measured and toxic unit transformed concentrations were at least four magnitudes below $TU = 1$ (see figure 3.2 D). Thus, the chemical concentrations were clearly below known toxic effect levels (based on LC50, see supplemental table S2-4) and were supposed to affect fish only on a sublethal level. The exposure in the selected streams was different in the number of detected compounds and the degree of toxic effects across streams, albeit all were on a low effect level (see figure 3.2 top). This mild chemical exposure might be limiting when investigating biological effects, even when considering alterations on the transcriptional level.

The compound-wise exposure patterns were **clustered into compound groups** (see 3.3) to overcome potential limitations due to low chemical effect levels. By the pairwise Pearson correlation analysis of compound wise exposure patterns, eight compound groups (CG) were identified. The compound wise correlation values ranged between -0.4 and 1. Each compound group had the highest sum of toxic units in another stream (see figure 3.3). When aiming at linking biological effects to exposure, it had to be considered that high correlations of exposure patterns might identify transcriptional effects due to the co-correlation of compounds. The mean pairwise correlation decreased substantially considering compound groups

* In [Ferrey et al. 2017], the number of detected chemical compounds has been 24. Considering the analytical data tables in appendix D of [Ferrey et al. 2017], for 27 compounds at least one entry without a flag, and thus considered as detected considering the ten selected streams.

Table 3.1. ***Selected single detected compounds from chemical analytical data.*** *Out of the 29 selected chemical compounds in this study, 9 were detected only once. These are listed below with their respective stream assignment. It has to be noted, that single detected compound exposure patterns are identical to the respective stream-wise exposure patterns.*

Stream	Compound
1	Carbamazepine
	Meprobamate
7	Trimethoprim
	Benzothiazole
	2-Amino-Benzothiazole
	2-Hydroxy-Benzothiazole
8	Triclosan
	Diazepam
10	Ciprofloxacin

instead of single compounds (see figure 3.3 C and D), but three compound groups contained one compound only (CG5: Amitriptyline, CG6: Sertraline, CG7: Iopamidol).

Although overall exposure suggested mild acute toxicity in fish, an endocrine activity signal had been measured, which had also been detected on the gene level (see supplemental figure S2-5) [Ferrey et al. 2017]. Thus, the mixture consideration of lowly concentrated chemicals as compound groups might be useful to detect chemical drivers for xenobiotic effects such as endocrine disruption.

Out of the seventy microarray samples of FHM hepatic tissue after 48h acute exposure to stream water, one was removed in the **preprocessing of microarray data** (see figure 3.4). The reduced set contained 11 518 unique genes after the removal of lowly expressed genes. As shown in figure 3.4C, the expression patterns of all sixty-nine microarray samples were not clustered by any exposure or location. Overall, the biological variance on the gene level seemed to be mainly driven by biological variance rather than differences in stream waters.

Four different exposure scenarios were used as an assumption to link exposure to biological effects (see figure 3.5). With the *general treatment* exposure scenario (Exposed water vs control), any xenobiotic treatment was assumed, which induced biological effects. Next to that, ten fully independent *stream-wise* exposures represented site-wise examination. Fur-

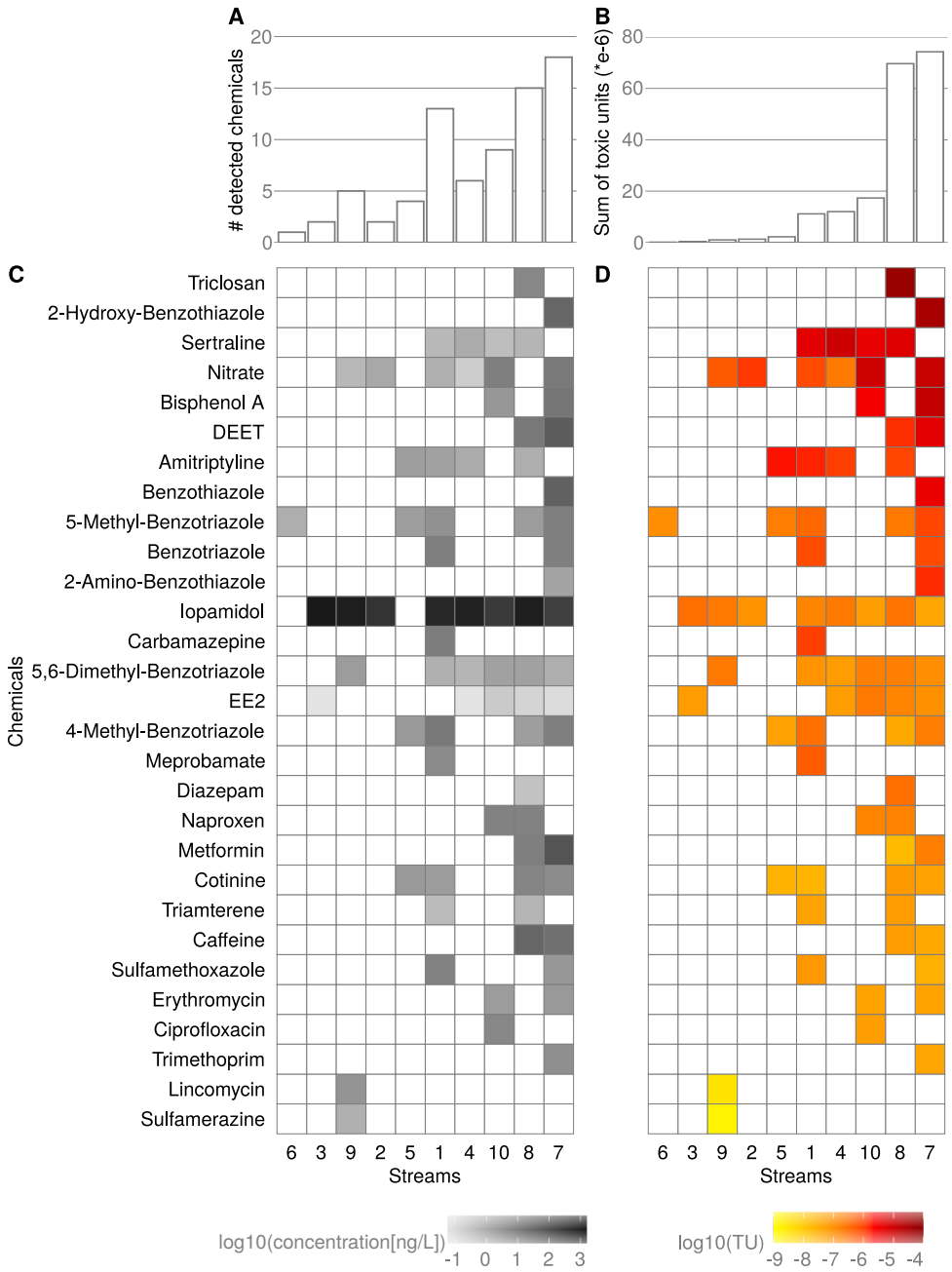

Figure 3.2. ***A mild toxic exposure by*** *29* ***detected chemicals in the selected streams in Minnesota*** *(2 compound equivalents identified via chemical in-vitro assessment; and 27 compounds by targeted chemical analysis of stream water samples).* ***A)*** *The number of detected compounds per stream.* ***B)*** *Sum of toxic units (TU) of detected compounds per stream.* ***C)*** *log10-transformed concentrations (ng/L) per detected compound and stream.* ***D)*** *log10-transformed TU per detected compound and stream. TU is defined as ratio of measured concentration to known concentration ranges based on LC50 (ECOSAR, ECOTOX and baseline) for fish. Rows and columns of both heatmaps are ordered by TU.*

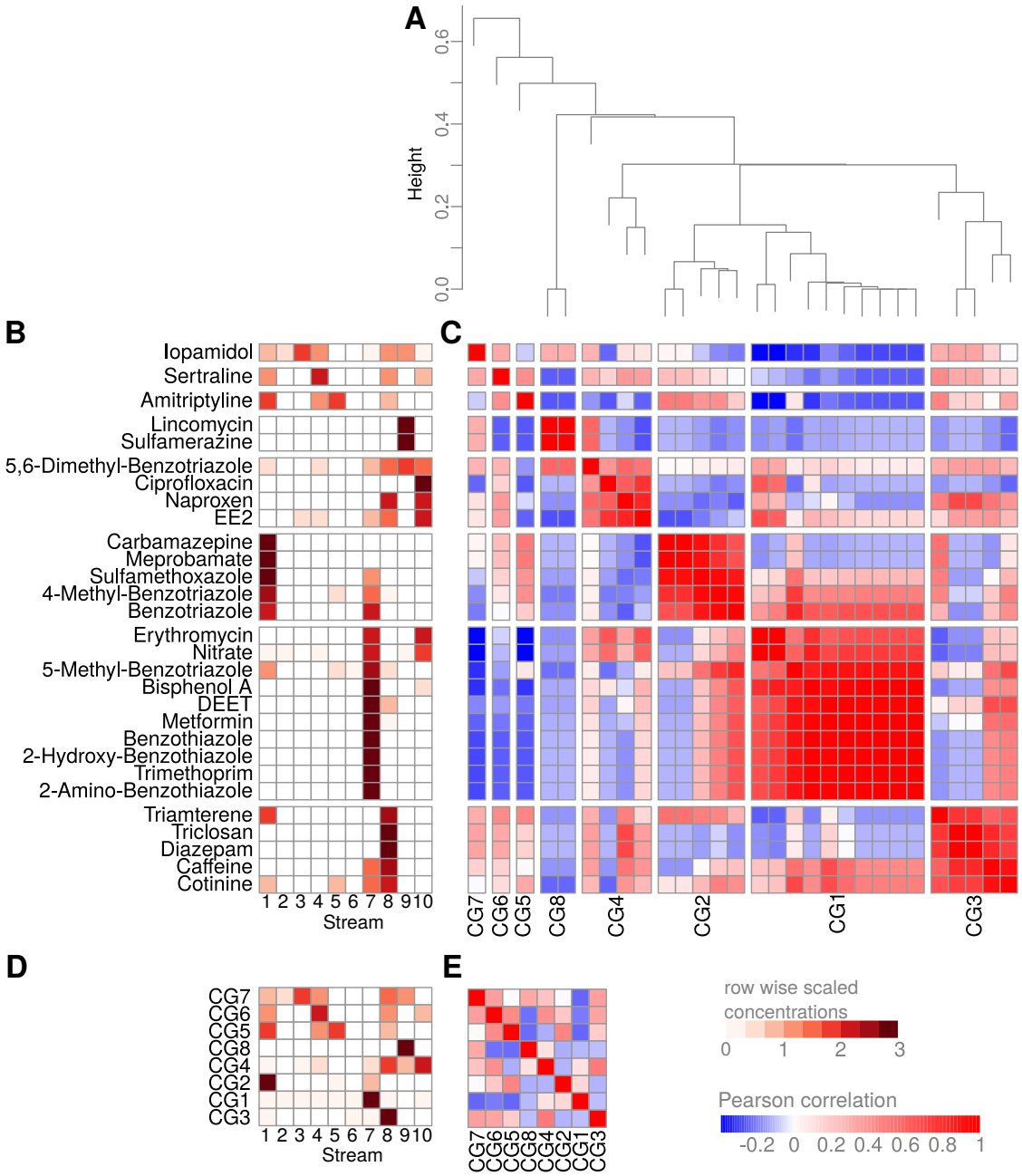

Figure 3.3. ***The exposure patterns of compound groups (CG) are less pairwise correlated, than for single compounds.*** *The heatmaps present the exposure patterns and correlation matrices of single compound and compound group exposure patterns.* ***A)*** *The dendrogram of the 29 detected compounds based on exposure pattern similarity in 8 CGs.* ***B)*** *The row wise vectors of root-mean-square scaled exposure patterns of detected single compounds grouped by correlation(see C).* ***C)*** *The pairwise Pearson-correlation matrix of compound wise exposure patterns clustered in CGs according to output of* `pheatmap` *R-function.* ***D)*** *The root-mean-square scaled exposure patterns of CGs.* ***E)*** *The pairwise Pearson-correlation of CG exposure patterns.*

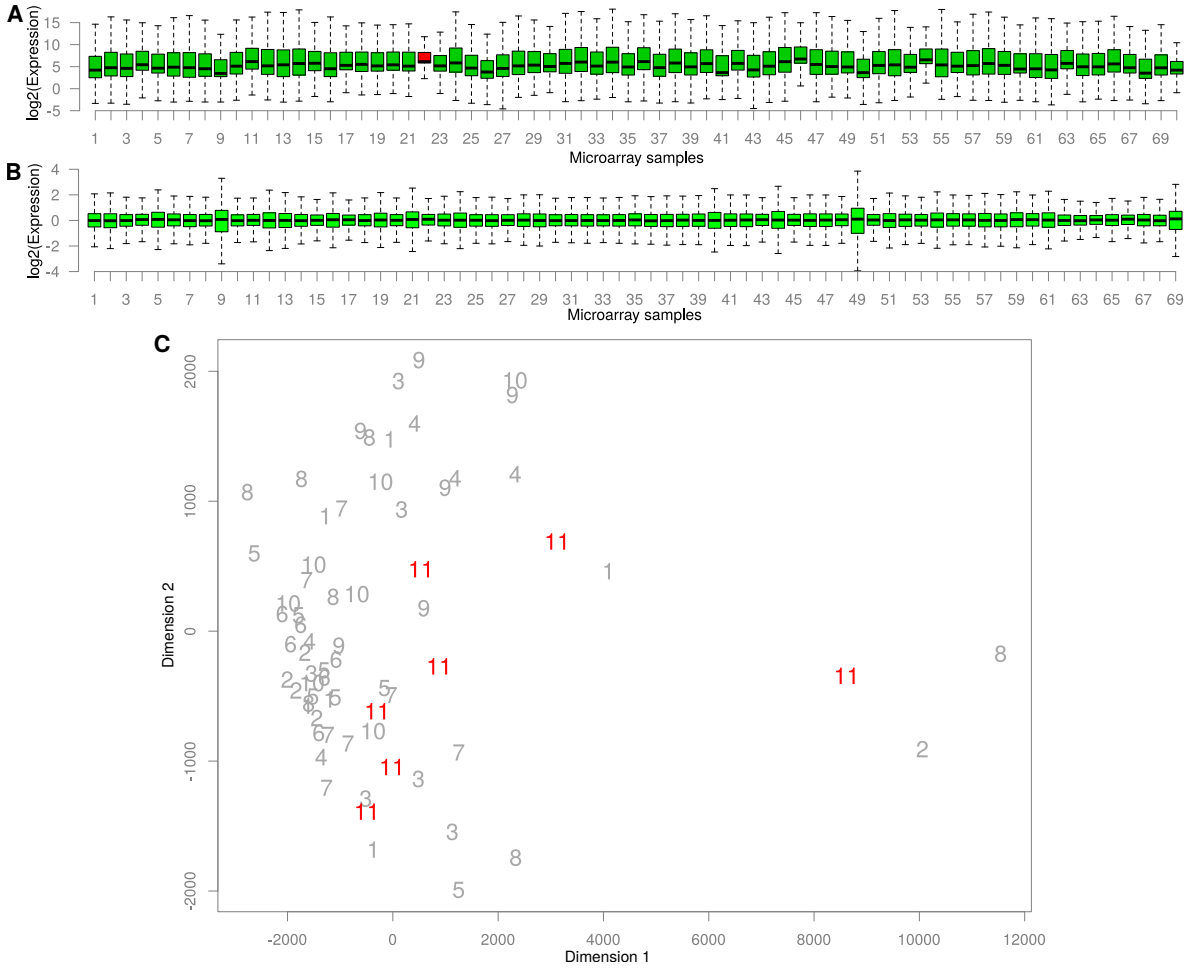

Figure 3.4. ***The preprocessed expression patterns of selected microarray samples did not cluster site-specific.*** ***A)*** *The sample wise* $\log_2$ *Expression-distribution (#samples:* 70*, #probes:* 49 905*) as boxplots with* 95 *%-interquantile range. Red colored boxes were removed based on the interquartile-range (IQR).* ***B)*** *The sample wise* $\log_2$ *Expression-distribution (# samples:* 69*, # probes:* 20 055*) as boxplot with* 95 *%-IQR after preprocessing and filtering of lowly expressed probes.* ***C)*** *Multidimensional scaling plot of distances between gene expression profiles. Red colored labels present the control samples, which are not distinct from the treated samples.*

thermore, twenty *single compound* exposure scenarios (exposure patterns of detected and not single detected compounds) and eight *compound group* exposure scenarios (exposure patterns of compound groups) were investigated as continuous models. The direct comparison allowed assessing the advantages of reducing the resolution of chemical exposure without waiving numeric values.

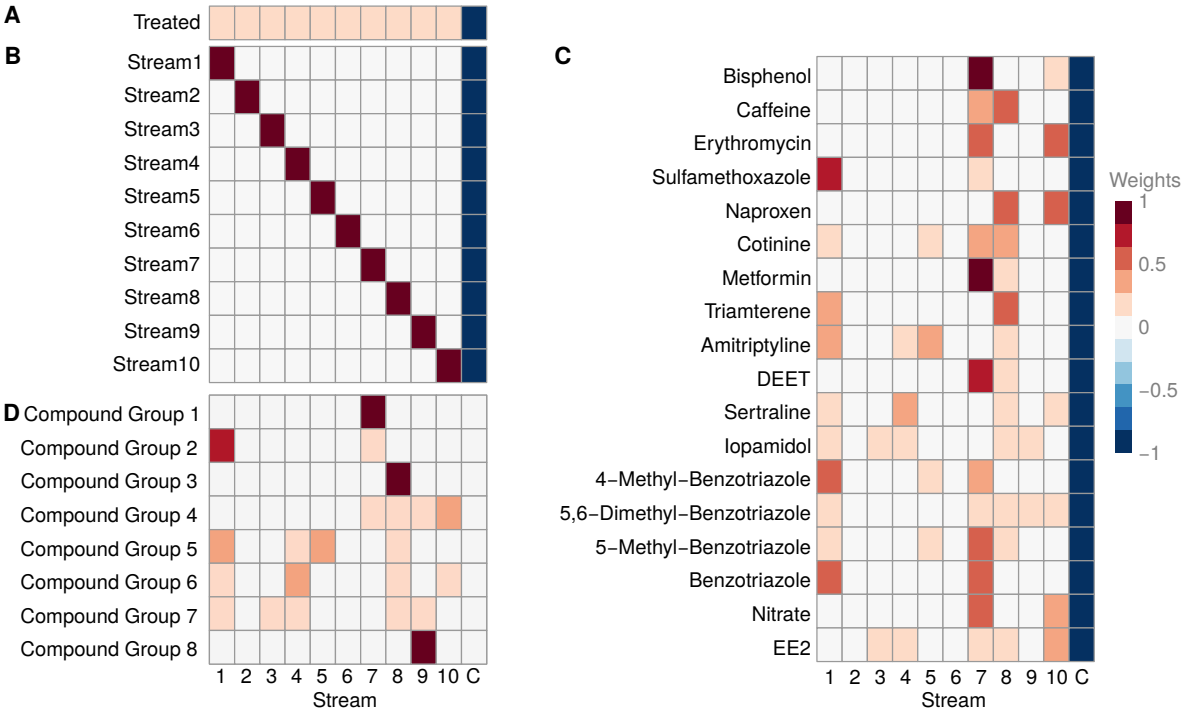

Figure 3.5. ***The four exposure scenario types in differential gene expression models across stream sites.*** ***A)*** *The general treatment model and* ***B)*** *the stream wise model are binarised exposure scenarios based on treatment-control-comparisons. Note that the stream sites of the general treatment exposure are equally weighted, and their sum is equally weighted to the control site.* ***C)*** *The single compound model and* ***D)*** *the compound group model are numeric exposure scenarios based on scaled toxic units (C) or stream-wise sum of toxic units (D).*

3.2.2 Differential gene expression analysis

Only mild acute toxicity was measured by chemical analytics and effect-based analysis. Nevertheless, the determined endocrine activities in *in-vitro* assessment led to the expectation that exposed fish should have significant transcriptional alterations due to xenobiotic effects. Based on the given exposure information, alterations of expression in FHM liver tissue for

different exposure conditions were compared to an unexposed condition. The differential gene expression analysis (DEA) was applied to model the different exposure scenarios (see figure 3.5).

General treatment exposure scenario. Although acute toxic effects were rather unexpected for an overall low TU, a significant adverse impact on pathway level was determined. For a general treatment contrast, 531 differentially expressed (DE) genes (see 3.6 A) and 75 significantly enriched biological terms were determined (see figure 3.6 B and table 3.2), which highlighted alterations of metabolism and immune responses (e.g . *Interleukin* 3 *signaling pathway*, *endogenous sterols*, *transport of small molecules*).

Highly exposure-correlated single compounds limited identification of chemical-specific induced effects. An assumption in DEA is that multi-linear contrast models consist of independent covariates. As shown in figure 3.3, the single compound and the stream covariates were highly correlated. For example, the exposure patterns of the single detected compounds were identical to respective stream-wise contrast. Therefore, the nine single detected compounds were considered only within the respective stream-wise exposure scenario. For single compound and stream-wise exposure scenarios, 97 to 694 DE genes were determined (see figure 3.6). For five of the streams, no significantly enriched term was identified. Especially these streams with a higher amount of DEs resulted in no biologically meaningful result. Also, the amount of significantly enriched terms was independent of the amount of DEs per single compound. However, 40 enriched terms were determined at a minimum for compounds detected in at least two streams.

It can be expected that co-correlated covariates are associated with a similar set of DE genes. However, only one might be responsible for the alterations in expression. In the present investigation, the DE gene sets for single compounds were similar to the general treatment (see figure 3.7, and each single compound contrast overlapped more than 50% with the DE gene set of the general treatment. The top-ranked genes of single compound exposure scenarios were also similar to those of the general treatment. For example, eight out of the ten top-ranked DE genes for the EE2 single compound exposure scenario overlapped to the top ten in general treatment.
Furthermore, the pairwise comparison of the DEA results for multiple detected compounds showed overlaps on the gene and the pathway level (see supplemental figure S2-1). Higher overlaps were mostly identified in subgroups, similar to groups of correlated single compound exposure patterns. Thus, it is likely, that the overlaps were induced by the determined co-

Table 3.2. ***Top*** 20 ***significantly functionally enriched results for the differential gene expression analysis to general treatment exposure scenario.*** *Based on the exposure scenario shown in figure 3.5 A, a* `limma` *model for differential expression was applied. Significantly differentially expressed genes were considered in a gene set enrichment analysis applying* `webGestaltR` *($FDR \leq 0.05$). The significant enrichment results with lowest FDR are shown. The enrichment outcome is available on UFZ-cloud (Path:* Data Chapter3 (TenStreams)/master_enrichment_DEA.csv *PW: PhD_SKraemer). (N(tot): Size of gene set (based on given data), N(enr): Number of genes in the enriched set, N(DEG): Number of significantly differentially expressed genes).*

Description	FDR	N(tot)	N(enr)	N(DEG)
IL-3 Signaling Pathway	< 0.0001	40	18	3
Endogenous sterols	0.0045	15	9	5
EGFR1 Signaling Pathway	0.0089	74	28	5
Metabolism of polyamines	0.0101	43	19	5
Regulation of ornithine decarboxylase (ODC)	0.0116	36	17	4
Hedgehog 'off' state	0.0141	48	17	4
CDT1 association with the CDC6:ORC:origin complex	0.0141	32	14	3
Ubiquitin-dependent degradation of Cyclin D1	0.0141	32	14	3
Ubiquitin-dependent degradation of Cyclin D	0.0141	32	14	3
Transport of small molecules	0.0142	184	58	12
GLI3 is processed to GLI3R by the proteasome	0.0147	37	14	3
Asymmetric localization of PCP proteins	0.0147	36	15	3
The role of GTSE1 in G2/M progression after G2 checkpoint	0.0153	34	14	3
FBXL7 down-regulates AURKA during mitotic entry and in early mitosis	0.0153	35	14	3
RUNX1 regulates transcription of genes involved in differentiation of HSCs	0.0157	41	18	4
Biosynthesis of amino acids	0.0160	51	15	3
Cross-presentation of soluble exogenous antigens (endosomes)	0.0160	31	14	3
Degradation of DVL	0.0161	35	15	3
Orc1 removal from chromatin	0.0163	36	14	3
Assembly of the pre-replicative complex	0.0165	33	14	3

dependencies of exposure patterns. This investigation highlights that the single compound exposure scenario has its limitations in the interpretability of DEA results.

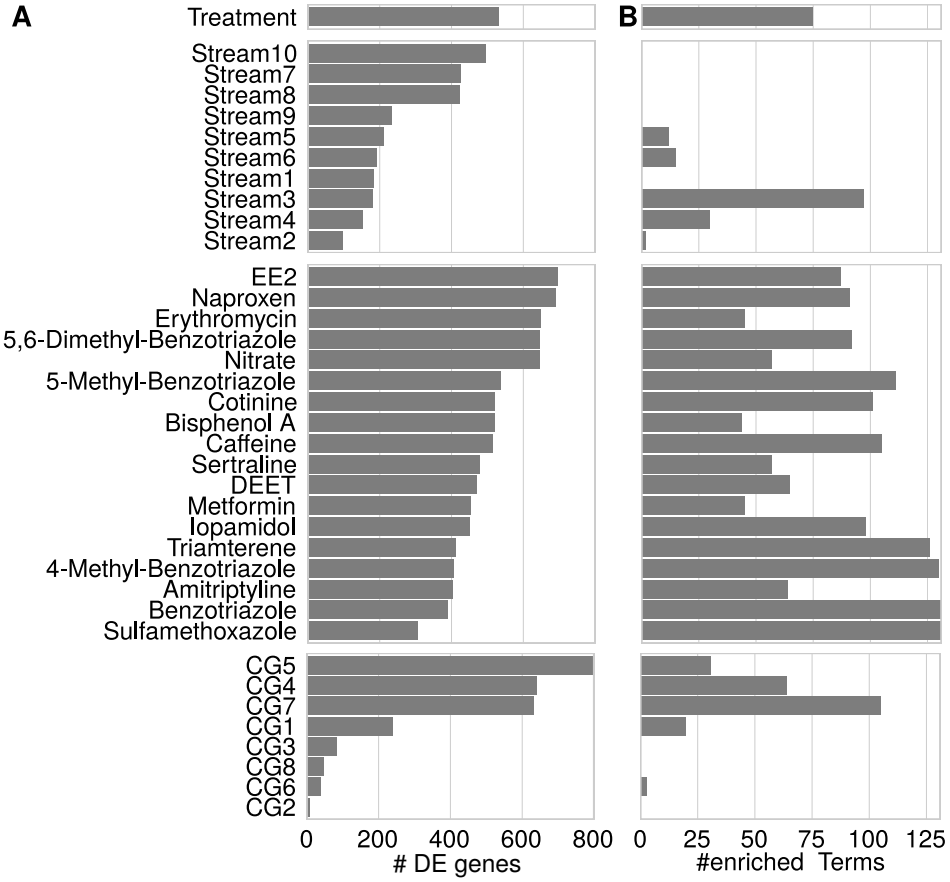

Figure 3.6. ***Identification of biologically meaningful results on pathway level applying differential gene expression analysis with geneset enrichment analysis.*** *The functional enrichment considered the total set of genes investigated in DEA with biological reference terms from KEGG, Reactome and WikiPathways applying the R-package* `webGestaltR`*.* ***A)*** *Number of significantly differentially expressed (DE) genes per exposure scenario (# samples:* 69*; total # genes:* 11 518*;* $FDR \geq 0.05$*). The contrasts are grouped in* 4 *exposure scenarios (general treatment; stream; single compound; compound group) and are decreasingly ordered by number of DE genes.* ***B)*** *Number of significantly enriched terms in gene set enrichment analysis (*$FDR \geq 0.05$*) with at least one DE gene. The enrichment outcome is available on UFZ-cloud (Path:* Data Chapter3 (TenStreams)/master_enrichment_DEA.csv *PW: PhD_SKraemer).*

Considering compound groups in DEA helped to assign biological effects to single compounds. In contrast to the single compound exposure scenario, a smaller set of covariates and weaker pairwise co-correlations were considered with the compound groups (CG). Per CG, 7 to 800 DE genes and 3 to 105 significantly enriched terms with at least one DE gene were determined (see examples in table 3.3). Similar to the stream-wise case, some compound groups had no significantly enriched biological term. Top enriched terms were associated with innate immune responses (e.g. *toll-like receptor cascades*, *T cell receptor signaling pathway*, or *neutrophil degranulation*) and regulation due to stress (e.g. *peroxisomal protein import*, *proteasome* or *endogenous sterols*).

The CG results differed in number and overlap to the general treatment. Overall, the sets of DE genes for CGs were less similar than those of a single compound scenario to the general treatment (see figure 3.7). As examples, CGs containing one compound are presented in detail (CG5 vs Amitriptyline; CG6 vs Sertraline; CG7 vs Iopamidol). Although exposure patterns were identical to single compound patterns, the covariates were less correlated in the multilinear model. In all three cases, the total and relative overlap of DE genes to general treatment was higher for a single compound scenario, although fewer DE genes were determined (see figure 3.7). The DEA outcome also had influences on the functional enrichment. In all three cases, fewer terms overlapped totally and relatively with the general treatment in the case of CGs. Thus, some advantages of a linear combination of co-dependent exposure patterns in DEA were highlighted for the examples of CGs with one compound. The disentangling of exposure-related specific effects was better feasible considering CGs, although the chemical resolution of xenobiotic effect was reduced by clustering of chemical wise exposure.

3.2.3 Association rule mining

The approach of association rule mining (AR) allows the identification of frequent associations and relationships in large data sets. It derives rules that can predict the likelihood of occurrence of one item set based on another item set's occurrence. Here, exposure patterns (streams, single compounds and compound groups) and genes were treated as items and frequently occurred exposure-gene relations were identified as rules (see figure 3.8). Compound groups are based on the clustering of exposure patterns (see figure 3.3). The itemset consisted of 11 518 genes, 11 stream sites, 29 compounds and 8 compound groups. The transaction set had a binarised format and consisted of the input information of the itemset.

A low-support ($support_{X \to Y} \geq \frac{3}{69}$) and a high-confidence threshold ($confidence_{X \to Y} \geq 0.8$) were chosen. In addition, a threshold of $lift_{X \to Y} \geq 1$ was considered, meaning that a rule is more likely to be better than a random expectation. With a selected low-support threshold,

Table 3.3. ***Top significantly functionally enriched terms for the differential gene expression results of the compound group exposure scenarios*** *Based on the exposure patterns shown in figure 3.5 D,* `limma` *models were applied for a differential gene expression analysis Significantly differentially expressed genes were considered in a gene set enrichment analysis applying* `webGestaltR` *($FDR \leq 0.05$). For five compound groups (CG), significant enrichment results with lowest FDR are shown. Not represented compound groups had no significant enrichment result. The enrichment outcome is available on UFZ-cloud (Path:* Data Chapter3 (TenStreams)/master_enrichment_DEA.csv *PW: PhD_SKraemer). (N(tot): Size of gene set (based on given data), N(enr): Amount of genes in the enriched set, N(DEG): Amount of genes, significantly differentially expressed to respective exposure scenario).*

	Description	FDR	N(tot)	N(enr)	N(DEG)
CG1	Cholesterol Biosynthesis	0.0272	10	8	1
	Toll-like Receptor Cascades	0.0279	45	9	2
	N-Glycan biosynthesis	0.0283	35	11	2
	Toll Like Receptor 4 (TLR4) Cascade	0.0288	36	8	2
	Toll Like Receptor 7/8 (TLR7/8) Cascade	0.0288	36	8	2
CG4	Lysosome	0.0021	84	43	7
	Endogenous sterols	0.0165	15	9	5
	Regulation of RUNX3 expression and activity	0.0206	34	15	3
	Regulation of RUNX2 expression and activity	0.0208	34	15	3
	T Cell Receptor Signaling Pathway	0.0215	42	12	4
CG5	Endogenous sterols	0.0158	15	9	4
	C-type lectin receptor signaling pathway	0.0261	56	17	6
	SUMOylation of intracellular receptors	0.0272	15	4	1
	IL-4 signaling Pathway	0.0277	24	10	2
	MyD88:MAL(TIRAP) cascade initiated on plasma membrane	0.0277	35	13	4
CG6	Peroxisomal protein import	0.0040	28	13	1
	Protein localization	0.0196	31	13	1
	Neutrophil degranulation	0.0393	160	47	1
CG7	Cyclin A:Cdk2-associated events at S phase entry	0.0012	37	21	3
	Cross-presentation of soluble exogenous antigens (endosomes)	0.0014	31	18	3
	SCF(Skp2)-mediated degradation of p27/p21	0.0015	34	19	3
	Proteasome	0.0015	35	19	3
	Regulation of RUNX2 expression and activity	0.0016	34	19	3

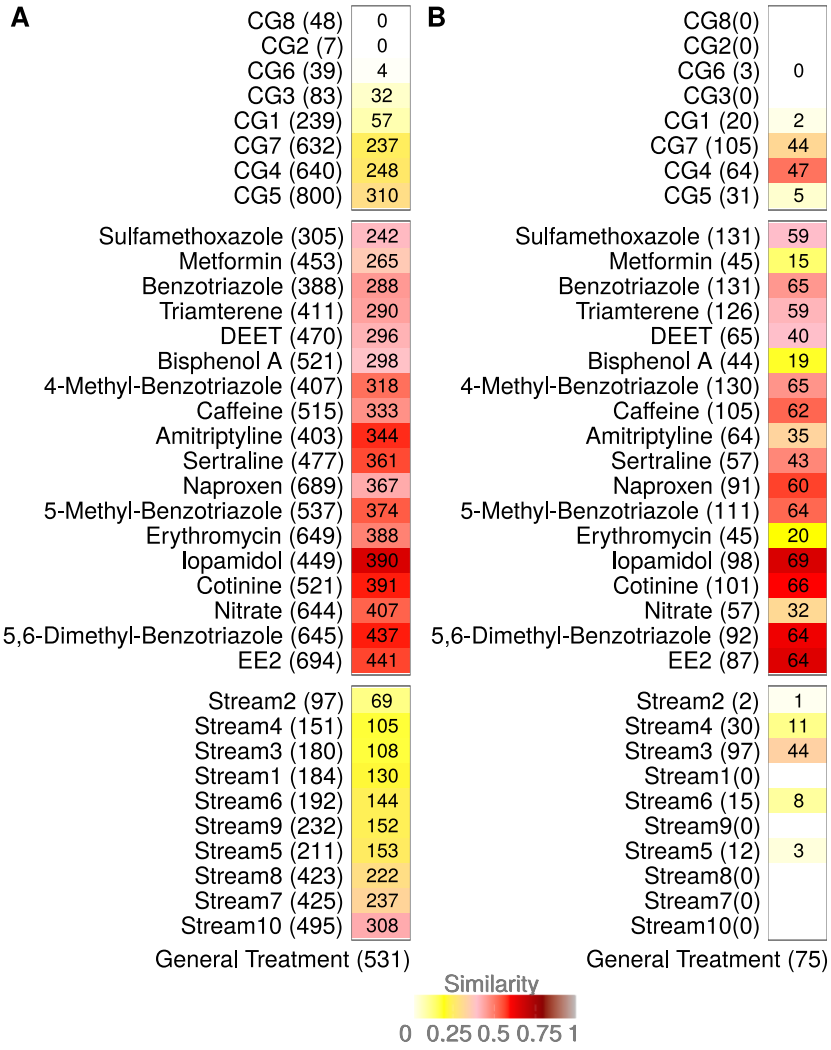

Figure 3.7. ***The sets of differentially expressed genes and enriched biological terms to the general treatment exposure scenario are highly similar to single compound exposure scenarios, but less similar to compound group exposure scenario.*** *Comparison of differential gene expression contrasts to general treatment on gene and biological pathway level.* ***A)*** *Total number of overlapping differentially expressed genes ($FDR \geq 0.05$) (value in tile) and Jaccard similarity (color code) to general treatment exposure scenario calculated with R-package* `GeneOverlap`*.* ***B)*** *Number of overlapping significantly enriched biological terms ($FDR \geq 0.05$) (value in tile) and Jaccard similarity (color code) to general treatment exposure scenario calculated with R-package* `GeneOverlap`*. The labels present the respective exposure scenario and the total number of significant differentially expressed genes (A) or significantly enriched terms (B) in brackets.*

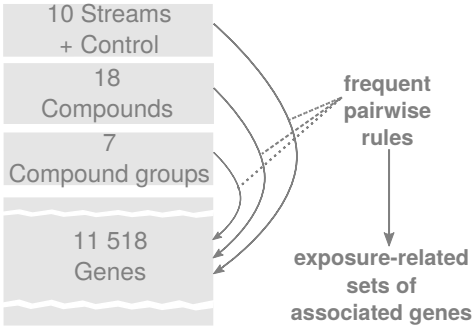

Figure 3.8. ***The binarised transaction set consisted of the input information of genes, compounds and streams*** *(see also method section 2.1.4). The total set of possible pairwise exposure-gene rules was generated from the transaction set of* 69 *microarray samples ($n_{genes} = 11518$) with samplewise regulated genes (absolute and control-normed* $\log_2(Expression) \geq 1$*) as possible* CONSEQUENT *and assigned streams, single compounds or compound groups as exposure scenario* ANTECEDENT *of the rule. Frequent pairwise association rules were determined applying* `apriori()` *of the R-package* `aRules`*. Only pairwise association rules were considered, which allows generating directly exposure-associated* CONSEQUENT *sets of genes from rules with identical* ANTECEDENT*.*

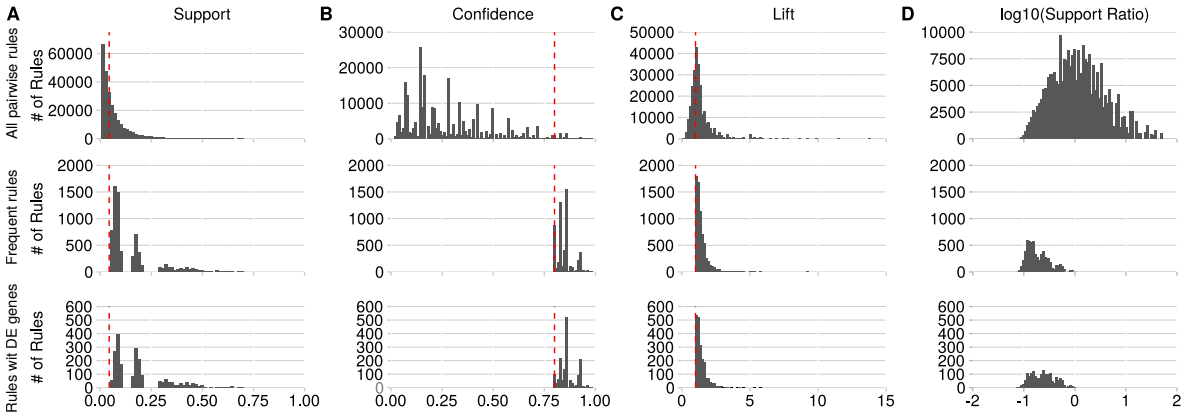

Figure 3.9. ***Distribution of AR-metrics across pairwise exposure-gene rules applying*** `aRules` *(top: All possible pairwise rules of an unique binarised exposure pattern and a regulated gene ($n = 195\,806$), middle: Pairwise rules determined with filter setting of apriori algorithm ($n = 5420$), bottom: Pairwise rules determined with filter setting of apriori algorithm and with differentially expressed gene ($n = 1948$)).* ***A)*** *Rulewise support with filtering threshold at $sup \geq 3/69$ (red dashed line).* ***B)*** *Rulewise confidence with filtering threshold at $conf \geq 0.8$ (red dashed line).* ***C)*** *Rulewise lift with filtering threshold at $lift > 1$ (red dashed line).* ***D)*** *Rulewise* $\log_{10}(supportratio)$*.*

the support ratio was also taken into account. This measure was not used as a threshold but as an interpretable tool to validate the set of frequent rules. Based on the filter thresholds for support, confidence and lift frequent pairwise rules were identified (see figure 3.9).
In total, 5420 unique pairwise association rules were determined with 17 unique binarised exposure patterns — approximately 2% of the number of possible pairwise rules. The results were summarised to sets of exposure-associated genes per ANTECEDENT (see 3.10). In the present study, the generated pairwise rules had the characteristic of a gene CONSEQUENT. Thus, each exposure-related rule set has an identical ANTECEDENT but different CONSEQUENTS.

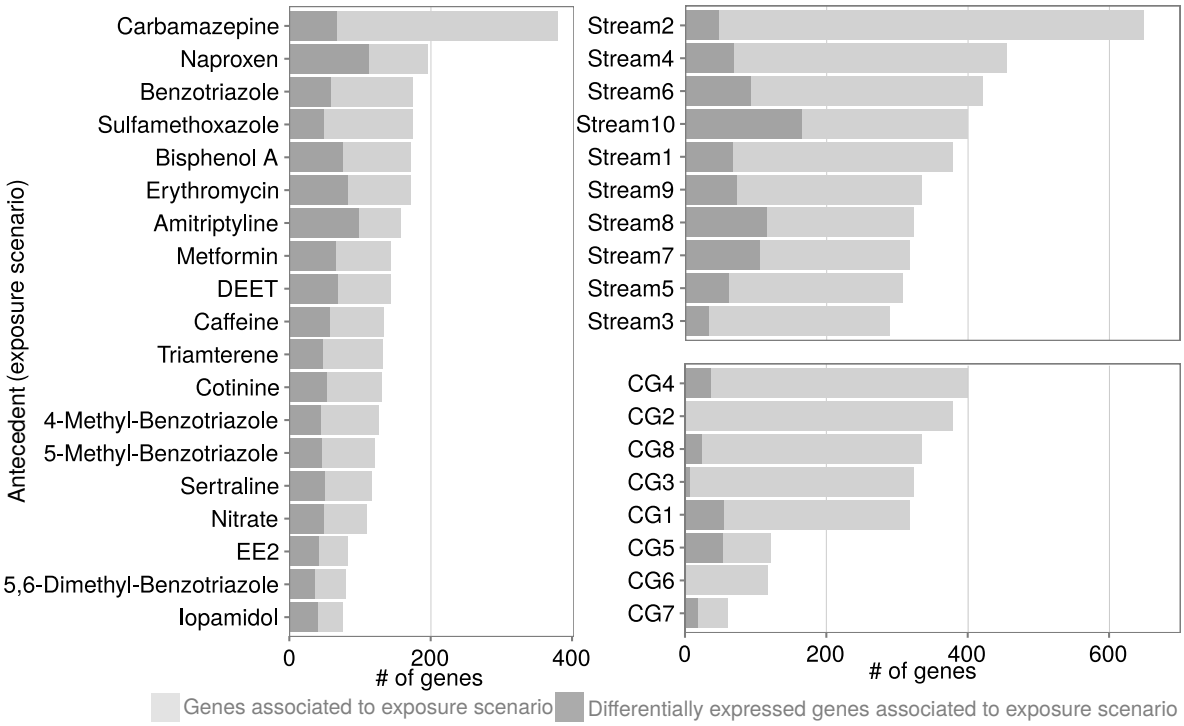

Figure 3.10. ***Identification of small to mediate overlaps of exposure-associated gene sets from association rule mining to these of differential gene expression.*** *Number of exposure-related genes grouped by* ANTECEDENT *in decreasing order. (left: Single chemicals, top right: Streams, bottom right: Compound groups). The dark grey highlighted bars present differentially expressed genes per exposure scenario. Single detected compounds (see table 3.1) are represented by the respective stream-wise exposure scenario.*

The consideration of unique exposure patterns reduced the set of 48 binarised occurrence patterns to 17. The single detected compounds had the identical set of associated genes as

the respective stream. Also, the compounds detected in the identical streams had identical sets of associated genes (see figure 3.10). No CG had a unique binarised exposure pattern. The associated genes to a CG were identical to the most extensive set of associated genes for respective single compounds (see 3.10). Thus, the CG occurrence pattern was defined by the common compound occurrence across the ten streams. The occurrences of the least often detected compound restricted the maximum occurrence pattern of a CG. In this investigation, all compounds of one CG were commonly detected in at least one stream, and with that, exposure patterns of CGs were identical to either one stream or at least one compound. Consequently, the binarisation of continuous data lead to a reduction of unique exposure patterns, and thus, to a loss of information potentially.

All selected rules had a $lift \geq 1$ and a high confidence. It was expected that a frequently occurring ANTECEDENT item dramatically increase the chance of a CONSEQUENT item, which makes rules reliable from a machine learning point of view. The support ratio helps describe some CONSEQUENT characteristics when considering a low-support threshold as in the present study ($support > 0.05$). The ANTECEDENT support was below CONSEQUENT support in all rules, except four (see figure 3.11). On average, the CONSEQUENT was six times more frequent than the ANTECEDENT, as the median support ratio valued 0.162. Each ANTECEDENT item had at least a support of 5/69, implying CONSEQUENT support of 30/69 on average. Thus, identified genes were affected frequently in the investigated microarray samples.

Applying either overrepresentation analysis or gene set enrichment analysis on exposure-related gene sets gained no significantly enriched terms for any exposure scenario. The presented AR approach with chosen filters identified an altered gene expression but supported neither xenobiotic stress nor exposure-specific effects on pathway level.

Furthermore, the exposure-associated genes in AR overlap partly with significant exposure-associated DE genes (see figure 3.10). Ciprofloxacin/Stream10 * had the highest number of associated genes (n=400) and the highest number of associated DE genes. Although the overlap to DEA results was, in general, smaller for compound groups, CG4 overlapped in parts to the DEA results. As assessed by the support ratio, AR identified mostly the frequently regulated genes. Consequently, the overlap to DEA might highlight genes representing an overall stress response instead of specific exposure-related responses. In parts, AR underlined the observation in DEA for exposure scenarios, which might primarily represent overall stress.

* Ciprofloxacin has an identical exposure pattern as CG4

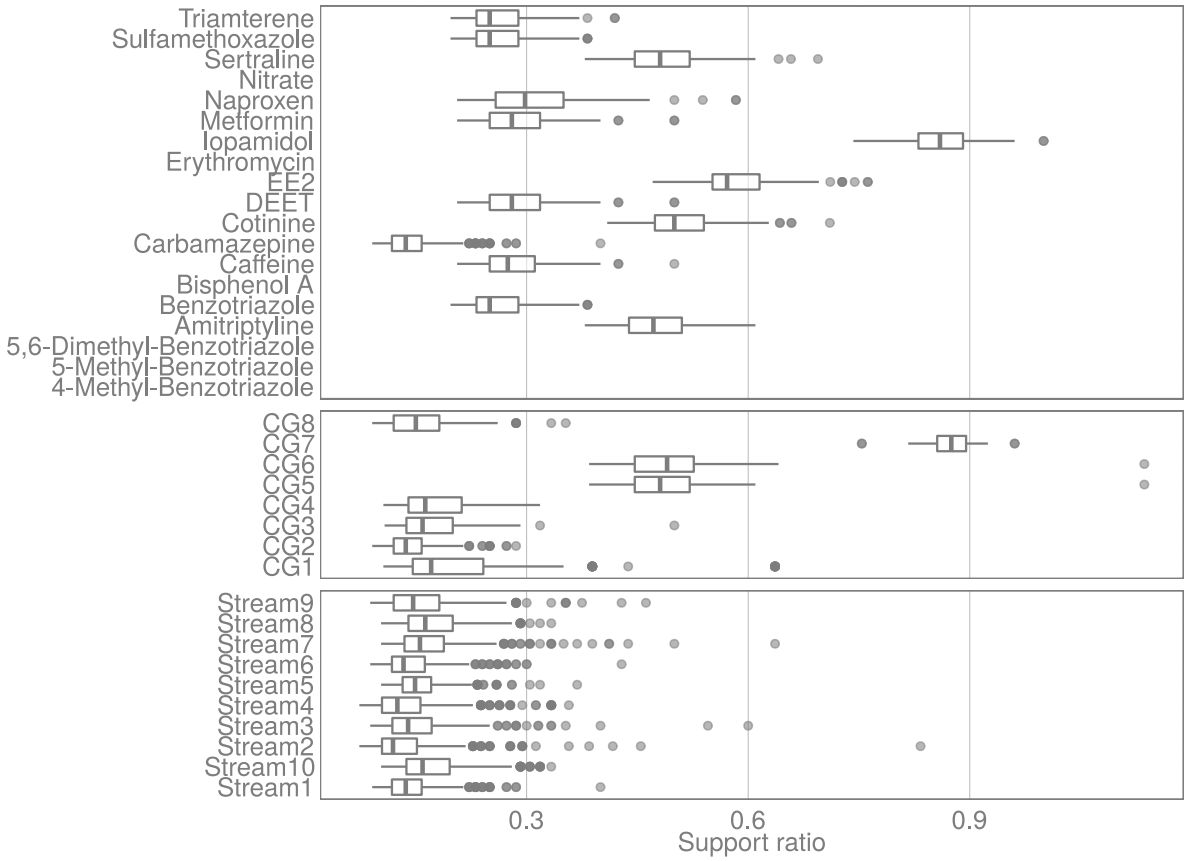

Figure 3.11. ***Antecedent items had more occurrences than consequent items in determined association rules.*** *Boxplots present the distribution of support ratio per* ANTECEDENT *item grouped by exposure scenarios (top: Single chemicals, middle: Compound groups, bottom: Streams). Single detected compounds (see table 3.1) are represented by the respective stream-wise exposure scenario. The support ratios were all below or equal 1 and thus, the* CONSEQUENT *support was greater than* ANTECEDENT *support in the determined rule set.*

3.2.4 Network inference

The weighted gene correlation network analysis (WGCNA) allows inferring a network based on the gene-wise expression patterns across all microarray samples. The preprocessed microarrays were investigated, and its 11 518 annotated genes were considered as nodes. A weighted and signed biweight mid-correlated gene network was generated with topological overlap measure (TOM) weighted edges. A soft thresholding power $\beta = 8$ was selected to give highly correlated node pairs a higher relative weight than lowly correlated node pairs. In this case, the scale-free topology model fit was above 0.8, and the mean connectivity was above 100 for the given gene expression data (see figure 3.12).

Twenty modules were determined in the generated gene dendrogram with the chosen β-parameter and after merging modules with highly correlated module eigengenes (see figure 3.13). The modules consisted of 33 to 2222 genes. The remaining set of 3150 genes was uncorrelated and was assigned to the grey module.

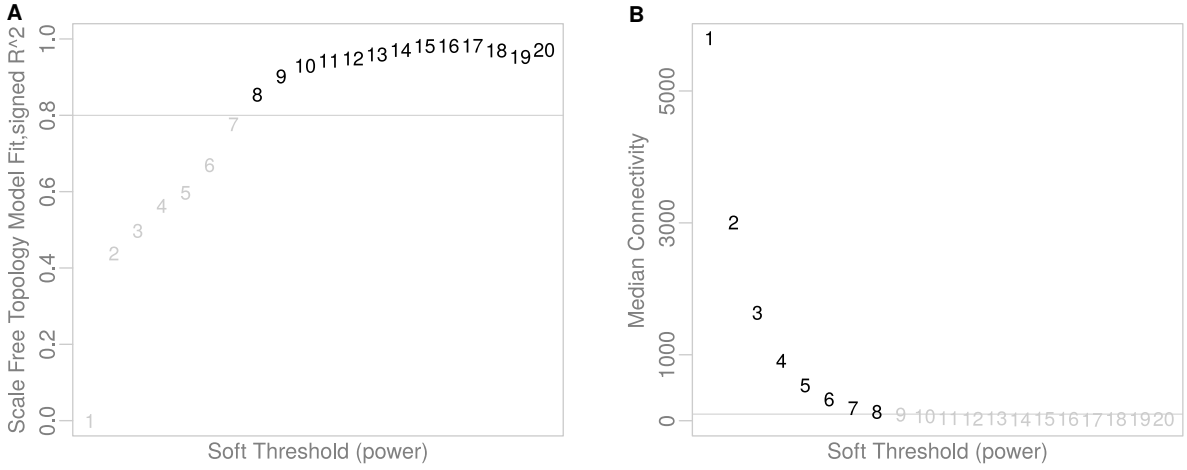

Figure 3.12. ***Scale free topology estimation for weighted gene correlation network analysis.*** *For biological networks a scale free topology characteristic is assumed. In the case of a weighted gene correlation network, a soft threshold — the power of the pairwise correlation to calculate gene-adjacency — can be estimated. For the considered expression data a soft threshold of $\beta = 8$ was chosen.* ***A)*** *Scatterplot of model fits by $|R^2|$ for power law describing scale free topology for soft thresholds from 1 to 20. According to [Zhang and Horvath 2005], $R^2 \geq 0.8$ is recommended (see black labels).* ***B)*** *Scatterplot of mean connectivity for soft thresholds from 1 to 20. A threshold of mean(connectivity) ≥ 100 was chosen (see black labels).*

The exposure patterns from DEA models (see figure 3.5) were the external (exposure) traits

to identify exposure-associated groups of co-correlated genes and calculate module-trait-correlations (MTC). The general treatment, three streams, twelve single compounds and four compound groups were significantly associated with at least one module (see figure 3.14). The entire MTC-matrix is shown in supplemental figure S2-2.

Similar to DEA, exposure traits with identical exposure patterns lead to an identical set of MTCs. For example, Stream9, Lincomycin, Sulfamerazine and CG8 had an identical exposure pattern with a singular peak in one stream and were significantly correlated to lightcyan (see figure 3.14 A).

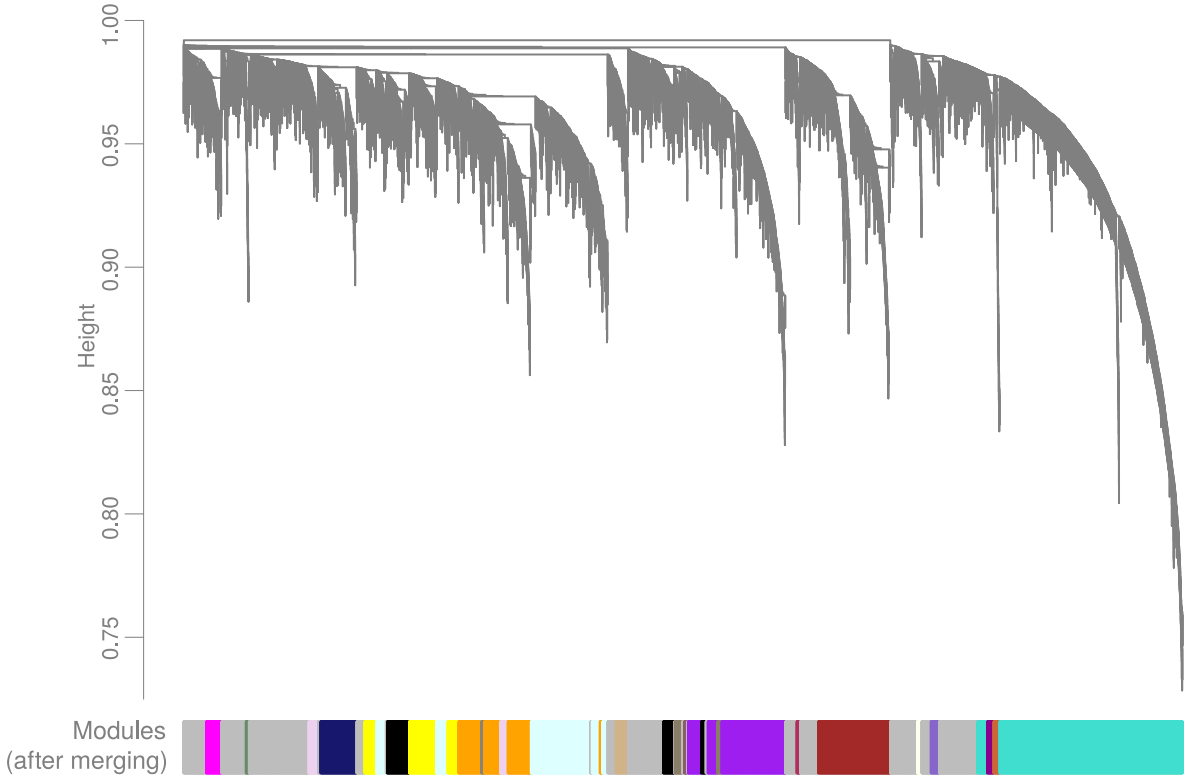

Figure 3.13. ***The gene dendrogram was clustered in*** 21 ***modules.*** *The signed and weighted gene correlation network was generated with a gene expression data set containing* 11 518 *genes and* 69 *microarray samples. Due to an improved robustness, biweight mid-correlation was chosen.*

The gene significance (GS) and module membership (MM) prioritized and ranked genes in GSEA. For each gene, we calculated the correlations to each module eigengene or each exposure pattern. Thus, non-module-member genes remained in ranking lists and became relevant in enriched gene sets. At least one gene of the enriched gene set had to be in the selected module to provide an interpretability for the specific module.

A rather unspecific xenobiotic response was expected when investigating acute exposure to lowly concentrated mixtures. In the scope of our investigation, the application of GSEA with WGCNA-metrics for gene ranking was partly meaningful. It could be valuable to interpret exposure in stream waters on the pathway level, albeit effects represented a rather unspecific xenobiotic response *.

Table 3.4. ***Significantly enriched terms for purple module with EE2 or compound group 4.*** *The purple module of the signed gene correlation network (see figure 3.13) was significantly correlated ($p_{MTC} \leq 0.05$ and $|MTC| \geq 0.3$) to the exposure patterns of EE2 and CG4 (see figure 3.14 A). The* `R`*-package* `webGestaltR` *determined functionally enriched terms ($FDR \leq 0.05$) with a gene list ranked by $MM \times -(log_{10}p_{GS})$. Enrichment results had to have at least one gene assigned to the purple module. CG4 had no further significantly enriched terms with other modules. (FDR: False discovery rate - adjusted p-value, N(tot): Size of gene set, N(enr): Number of genes in enriched set, N(mod): Number of genes in enriched set and assigned to module.*

Exposure	Description	FDR	N(tot)	N(enr)	N(mod)
EE2	Insulin signaling pathway	0.0087	80	15	9
	Regulation of actin cytoskeleton	0.0262	78	21	12
	Adipogenesis	0.0273	48	7	5
	Focal adhesion	0.0349	77	14	10
CG4	Regulation of actin cytoskeleton	0.0116	78	24	11
	Insulin signaling pathway	0.0154	80	22	9
	Vascular smooth muscle contraction	0.0185	42	9	5

Significantly enriched gene sets ($FDR \geq 0.05$) with at least one gene within the investigated module were considered biologically meaningful. The number of enriched terms varied across all MTCs with 0 to 122 biologically meaningful terms. Nine-teen MTCs contained at least one significantly enriched term applying GSEA with gene ranking by MM and GS. It was more likely to have a significant enrichment or multiple significant enrichments the greater the module was.

For example, the exposure to 5,6-Dimethyl-Benzotriazole had the identical set of significant MTCs as exposure to CG4. Both scenarios were associated with modules at nearly identical correlation values (see supplemental figure S2-2) and had significantly enriched terms (see

* see enrichment outcome: Data Chapter3 (TenStreams)/master_enrichment_WGCNA.csv PW: PhD_SKraemer

figure 3.14 and table 3.4). GSEA associated *insulin signaling pathway* to CG4, which might highlight a xenobiotic response due to endocrine disruption. The *regulation of the actin cytoskeleton* is known to be affected by oxidative stress in cells. Such xenobiotic responses might be associated with endocrine disruption. CG4 contains EE2. Out of 998 purple genes, 421 EE2-gene-associations (known in STITCH and CTD) were identified (see figure 3.16). Both enriched terms were also significantly enriched for the exposure scenario of EE2, but in the case of CG4, the enriched sets contained more genes. Although the number of enriched terms to CG4 was small, the biological meaning highlights cellular responses in the liver to xenobiotic stress. For the significant correlation to lightcyan, no significantly enriched terms were identified. However, the xenobiotic stress response in FHM might also be supported by the high coverage of EE2-associated gene interactions (352/1086). Xenobiotic responses on pathway level might genuinely be associated with a (mixture) exposure of EE2-equivalents, Ciprofloxacin, Naproxen and 5,6-Dimethyl-Benzotriazole when considering CG4.

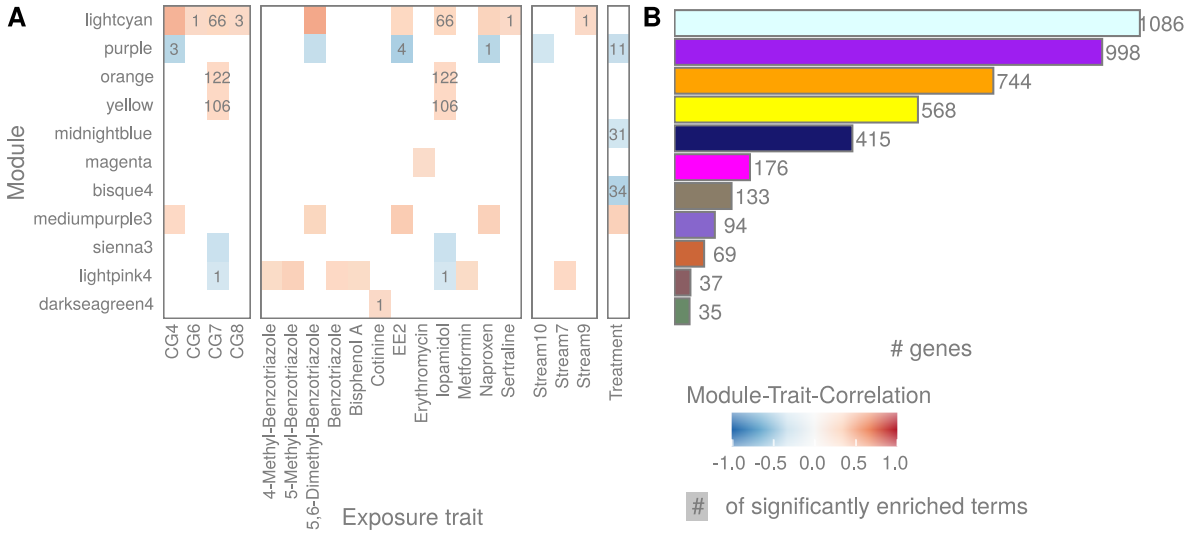

Figure 3.14. ***Significant module-trait-correlations (with number of enriched biological terms) and module sizes. A)*** *Heatmap of significant biweight mid-correlations of modules to exposure scenarios ($|MTC| \geq 0.3$ and $p_{cor} \leq 0.05$). The color of tiles represent the direction and strength of correlation. Text in tiles represent the number of significantly enriched terms ranked by module membership and gene significance of the respective module-trait-correlation with at least one gene within the module.* ***B)*** *Barplot of number of genes per module in decreasing order. The color represents the assigned name of a module.*

As highlighted above, known EE2 interactions were well covered in the purple and the lightcyan modules. Therefore, these modules might represent the known endocrine disruptive

perturbation by the exposure to the compounds in CG4. CG4 had the highest amount of exposure-related genes (see figure 3.15).

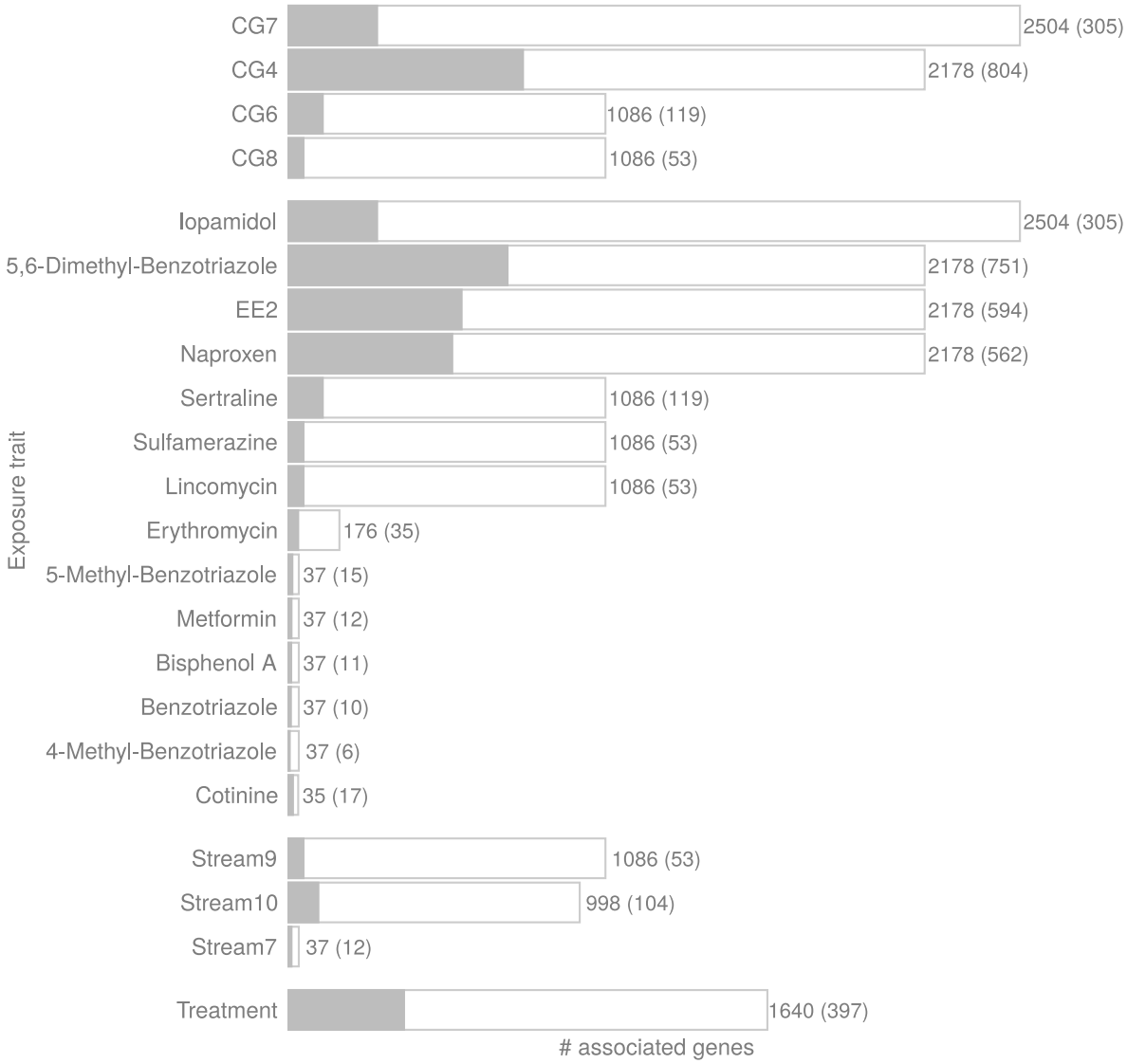

Figure 3.15. ***The number of exposure trait correlated genes, which were also differentially expressed, was the greatest for CG4.*** *Number of exposure-associated genes identified in weighted gene correlation network analysis grouped by exposure scenarios and decreasingly ordered. Filled partitions of barplot and number in brackets present the number of exposure-associated genes, which are also significantly differentially expressed.*

In this study, genes with $|GS| \geq 0.3$ and $|MM| \geq 0.3$ were considered relevant for an exposure

trait *. In both modules — lightcyan and purple –, the relation of MM and gene GS resulted for CG4 in many top-right genes that were, in parts, supported by the DEA and AR results (see figure 3.16 for the purple module).

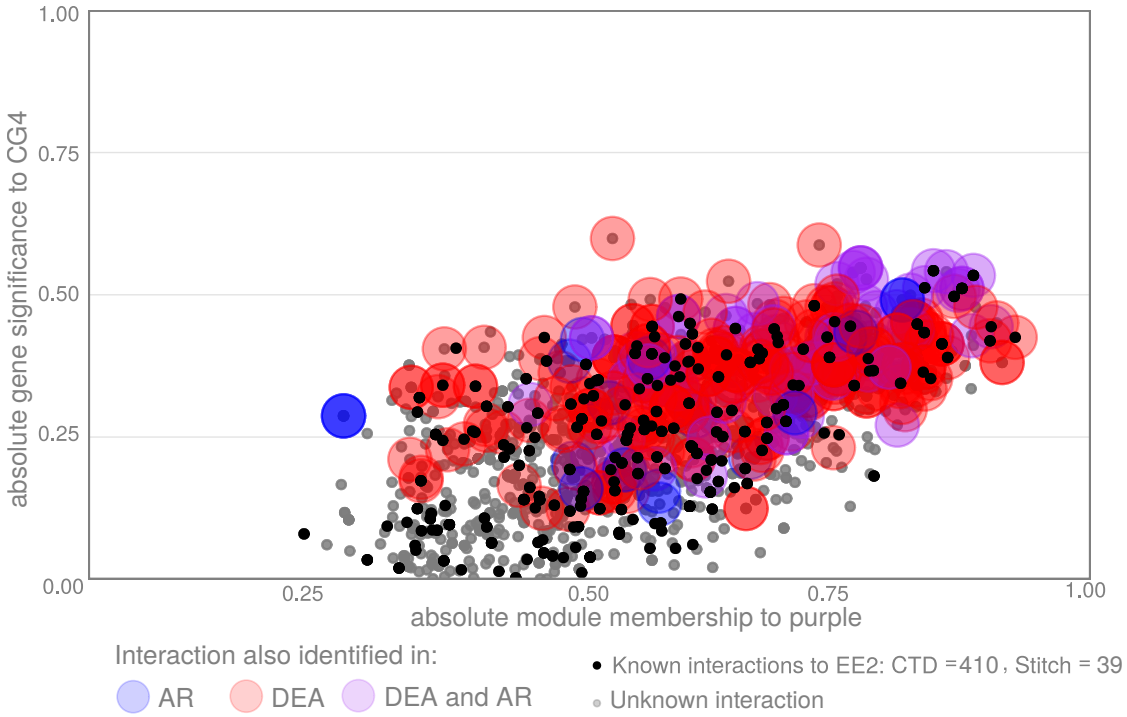

Figure 3.16. ***Many purple genes were known as EE2-interacting and other approaches supported them as CG4-associated.*** *Relation of module membership and gene significance for the example of the module-trait-correlation purple-compound group 4 (CG4). The data points represent the genes of the lightcyan module. The color assigns associations to EE2 in external reference bases. The colored halos present overlapping exposure-gene interactions in differential gene expression (DEA) or association rule mining (AR). Genes in the top-right corner of such a scatterplot are expected to be hubgene-like and their expression pattern represent the respective exposure pattern well.*

Thus, higher connected genes in the modular subnetworks projected these exposure patterns well across approaches. However, the top-right exposure-related genes were not supported by other approaches in every significant MTC.

The consideration of a method integration was essential to increase credibility on individual WGCNA results. On the other hand, a method-integrative pre-filtering of genes would reduce

* GS showed, whether a genes expression might describe the exposure pattern properly ($|GS| \geq 0.3$). GS was the biweight mid-correlation of exposure pattern and gene-wise expression pattern. MM was the biweight mid-correlation of module eigengene and gene-wise expression pattern. A high MM value presents an intra-modular hub-gene-like gene, which is highly connected to a large set of genes within the same module.

the number of genes extremely, and thus a determination of significantly enriched terms might become negatively influenced.

3.2.5 Method comparison

As omics studies have been already well established in (eco-)toxicological investigations, the current knowledge has been documented in external reference bases like the comparative toxicogenomic database (CTD) as chemical-gene interactions or STITCH as chemical-protein interactions. Such references were beneficial for the validation of the determined results. Some overlap between determined exposure-related associations and known compound-specific effects was expected and was considered on the gene and the pathway levels.

The overlap to chemical-gene interactions known in CTD and STITCH for selected compounds was determined for each represented chemical compound * (see tables S2-6, S2-7, S2-8, S2-9 and S2-10). No chemical-gene interactions were given for rarely studied compounds, like 5,6-Dimethyl-Benzotriazole. In contrast, thousands of chemical-gene interactions were known for frequently studied compounds like EE2 or Bisphenol A. Both compounds have been known as endocrine disruptors. On the one hand, it may highlight the main limitation of a study bias in human-curated toxicological databases. On the other hand, the top-ranked chemical compounds also supported the expectation to identify transcriptional responses typical for cellular stress and endocrine disruption. The total overlap to these compounds was the largest.

To proof, whether biologically meaningful results were also reliable, the overlap of enriched terms to chemical-pathway interactions listed in CTD was identified per exposure scenario and method.

In the case of **DEA**, compound wise CTD pathway associations significantly overlapped with exposure-associated enriched terms in cases of single compound and compound group consideration (see figure 3.17). DEA resulted in significantly enriched terms for 18 single compound exposure scenarios, and 15 contained at least one known interaction in CTD. The compounds Nitrate, Metformin, Bisphenol A, Caffeine, EE2 and Sertraline significantly overlapped to enriched terms (see figure 3.17). Thirteen compounds had any overlap of respective compound group exposure-associated enrichment results to CTD pathway associations (see figure 3.17).

For Nitrate, Metformin and EE2 a significant overlap to CTD's pathway association was also determined in compound group exposure scenarios. These compound-related exposures were associated with cellular stress and xenobiotic responses. A general perturbation and

* It has to be noted, that not every compound is represented in the toxicological references.

potentially endocrine disruptive effects were expected due to *in-vitro* responses of EE2- and nitrate-equivalents.

CG4 contained EE2, and more than a third of the DE genes to CG4 was listed in either CTD or STITCH as known compound-gene interaction to EE2 (see figure 3.17 A). As EE2 was not directly measured in streams but represented the estrogen activity within stream waters, the three other compounds in CG4 might be the drivers for the associated xenobiotic perturbation. The enriched terms for the DE genes hinted at cellular stress and endocrine disruption (see table 3.5).
In the case of DEA, a chemical group exposure scenario might reduce the projection of molecular effects on the transcriptional level, which a 'hidden chemical background' might drive. The *in-vitro* measured endocrine effect associated with nitrate and EE2 was supported when comparing exposure scenario related DE genes and significantly enriched terms to reference sets from CTD. With this, the biological effect for Metformin was correlated to nitrate and for both significant on the pathway level(see enrichment results table 3.6). It may highlight Metformin as a potential chemical driver of stress-responsive transcriptional effects for the *in-vitro* measured nitrate effect. Consequently, compound groups instead of single compounds, as exposure scenarios, were partly necessary to disentangle chemical driven xenobiotic effects from the overall stress response.

In the case of **WGCNA**, at least one significantly enriched term overlapped to CTD considering enrichment results for EE2 and Iopamidol. No pathway interaction overlap was significant when performing a χ^2-test, neither for a single compound nor for compound group exposure scenarios. More than 1000 enriched terms were associated with EE2 in CTD. However, the few identified enriched terms were not considered independent, although overlap to enriched results was high. Three out of four EE2-associated enriched terms in the single compound scenario and three out of three in the compound group scenario overlapped. On the contrary, Iopamidol (single compound and the only representative of CG7) had the most extensive set of significantly enriched terms ($n = 295$) considering WGCNA enrichment results. Out of 53 Iopamidol-pathway interactions known from CTD, six were identified in WGCNA (see table 3.7).

These terms were related to stress responses inducing inflammatory reactions like *signaling by interleukins* or *Fc epsilon receptor I signaling*. Although the overlap was not significant according to the $\chi^2 - test$, the result was meaningful for a xenobiotic Iopamidol exposure.

Both computational approaches - DEA and WGCNA - were able to identify pathway attributions, which were exposure-related and biologically meaningful. Xenobiotic stress-related terms were identified. However, only some DEA and no WGCNA results were reliable

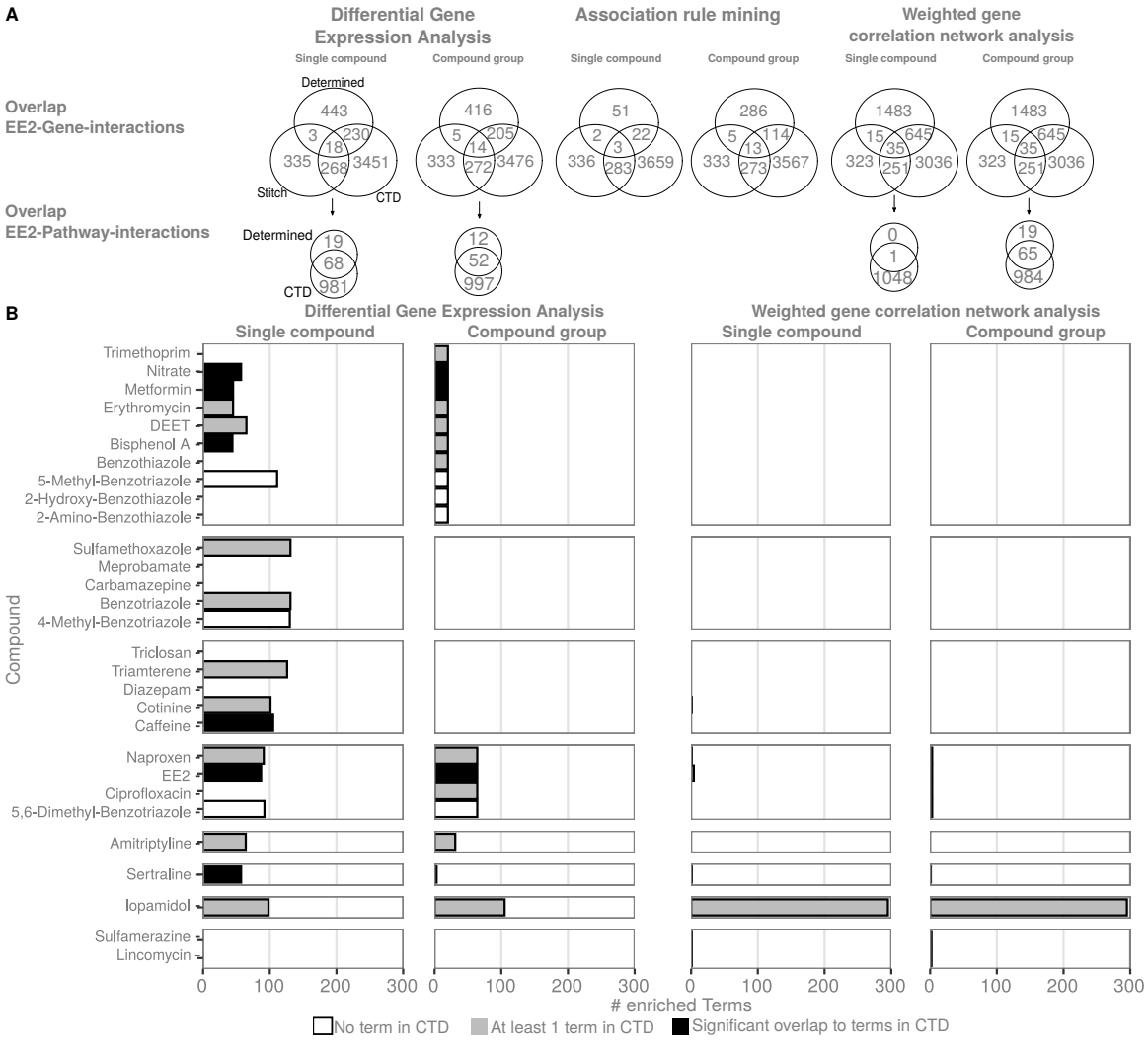

Figure 3.17. ***Identification of*** 3 ***significant overlaps of enriched terms considering differential gene expression in single compound and compound group exposure scenarios.*** *Method comparison with external reference bases.* ***A)*** *Overlap of EE2-gene interactions to known EE2-reference sets in CTD or STITCH per method (top) and overlap of significantly enriched terms to EE2 to known EE2-associated pathways in CTD per method (below).* ***B)*** *Number of significantly enriched terms* ($FDR \geq 0.05$) *in single compound or chemical group exposure scenario. Color present degree of overlap between compound wise enriched terms and known compound-pathway interactions from the comparative toxicogenomic database (CTD) (Black: Significant overlap, if* χ^2*-Test with* $p \geq 0.05$*, Grey: At least one known interaction in CTD, White: No overlap to interactions in CTD).*

considering statistical significance in a χ^2-test.

Table 3.5. ***Most significantly enriched terms for DEA results to EE2 and CG4 which are also known in CTD.*** *Based on the exposure patterns of EE2 and CG4 (see figure 3.5 C and D),* `limma` *models for differential expression were applied. A gene set enrichment analysis was performed applying* `webGestaltR` *($FDR \leq 0.05$) (see* Data Chapter3 (TenStreams) /master_enrichment_DEA.csv *PW: PhD_SKraemer). A significant overlap to CTD was determined for EE2 (single:* 68/87 *terms in CTD, CG4:* 52/64*) with χ^2-test. Table shows ten most significantly enriched terms.*

Exposure	Description
EE2	GnRH signaling pathway
EE2	RET signaling
EE2	Endogenous sterols
EE2	Insulin receptor signalling cascade
EE2	PI5P, PP2A and IER3 Regulate PI3K/AKT Signaling
EE2	Signaling by FGFR3
EE2	Signaling by FGFR4
EE2	Negative regulation of the PI3K/AKT network
EE2	Nonsense Mediated Decay independent of the Exon Junction Complex
EE2	Regulation of actin cytoskeleton
CG4	Lysosome
CG4	Endogenous sterols
CG4	Hedgehog 'on' state
CG4	Cyclin A:Cdk2-associated events at S phase entry
CG4	SCF(Skp2)-mediated degradation of p27/p21
CG4	GLI3 is processed to GLI3R by the proteasome
CG4	Visual phototransduction
CG4	CDT1 association with the CDC6:ORC:origin complex
CG4	Ubiquitin-dependent degradation of Cyclin D
CG4	Degradation of DVL

3.2.6 Application case of method integration

We performed GSEA combining the results of WGCNA and DEA. For each significant MTC, the modular gene set was considered as the reduced background of genes. The genes were

ranked by $signum(logFC) \times -log10(FDR)$ for the DEA outcome of the respective exposure scenario. In total, eight significantly enriched terms were identified across five different significant MTCs (see table 3.8).

Table 3.6. ***Significantly enriched terms for DEA results of metformin and nitrate which are also known in CTD.*** *Differential gene expression analysis was applied with* `limma` *models for exposure patterns of nitrate, metformin and CG*1 *(see figure 3.5 C and D). A gene set enrichment analysis was performed applying* `webGestaltR` *($FDR \leq 0.05$) (see* Data Chapter3 (TenStreams) /master_enrichment_DEA.csv *(PW: PhD_SKraemer)). Significant overlaps to CTD were determined in cases of metformin (single:* 21/45 *terms in CTD, CG:* 14/20*) and nitrate (single:* 17/57 *terms in CTD, CG:*13/20*) according to* χ^2*-test. Table shows five most significantly enriched terms.*

Compound	Scenario	Description
Metformin	Single compound	Toll Like Receptor 10 Cascade
		Toll Like Receptor 5 Cascade
		Toll Like Receptor TLR1:TLR2 Cascade
		Toll Like Receptor TLR6:TLR2 Cascade
		Toll Like Receptor 2 Cascade
nitrate	Single compound	Biosynthesis of amino acids
		Metabolism of amino acids and derivatives
		Toll Like Receptor 4 Cascade
		Toll Like Receptor 7/8 Cascade
		MyD88 dependent cascade initiated on endosome
Metformin/Nitrate	Compound group	Toll Like Receptor 4 Cascade
		Toll Like Receptor 7/8 Cascade
		MyD88 dependent cascade initiated on endosome
		Toll Like Receptor 9 Cascade
		Toll Like Receptor 10 Cascade

In the case of the MTC purple-CG4, toxicological reasonable enriched terms were identified with more than two module-member genes. A GSEA with DEA-ranked genes determined an endocrine-related and a stress-responsive effect. The *cytokine signaling in immune system* was also a highly ranked chemical-pathway interaction for three of the four compounds in CG4 (EE2: Rank 19 of 1337, Ciprofloxacin: Rank 1 of 173, Naproxen Rank 2 of 190). The enriched gene set for the *androgen receptor signaling pathway* contained genes that encode pro-

Table 3.7. ***Significantly functionally enriched terms for weighted gene network analysis to Iopamidol which are known in the comparative toxicogenomic database.*** *In the case of Iopamidol, the exposure pattern was identical for compound group and single compound. A gene set enrichment analysis was performed for all modules, which were significantly associated with Iopamidol (see figure 3.14) applying* `webGestaltR` *($FDR \leq 0.05$). (see enrichment outcome:* Data Chapter3 (TenStreams) /master_enrichment_WGCNA.csv *PW: PhD_SKraemer). The significant enrichment results, which were covered in CTD, are shown with their rank in CTD significant enrichment results.*

Rank	Description
5/53	Cytokine Signaling in Immune system
28/53	Toll Like Receptor 9 Cascade
30/53	Cellular responses to stress
35/53	Signaling by Interleukins
36/53	Protein processing in endoplasmic reticulum
40/53	Fc epsilon receptor I signaling

teins relevant to direct molecular interactions with the androgen receptor (*ctdp1*, *ncoa2* and *nr3c1*). The enrichment of this term was associated with steroid binding and glucocorticoid receptor activity in general, which hinted at lipid metabolism and was a reasonable xenobiotic and metabolic response due to endocrine disruption. These identified terms highlighted the endocrine-related effect, which was recognized by a mixture of endocrine activity and measured concentrations of Ciprofloxacin, 5,6-Dimethyl-Benzotriazole and Naproxen. None of the enriched terms per MTC was identified with the GSEA approach considering WGCNA-metrics. However, both applied gene rankings identified meaningful results with WGCNA outcomes.

Table 3.8. ***Significantly enriched terms for method integration of WGCNA and DEA.*** *Based on various exposure patterns (see figure 3.5),* `limma` *models for differential expression were applied. The network inferred grouping of genes applying the* `WGCNA`*-package in R was used to filter the considered set of DEA results for significant module-exposure trait correlations (MTC) (see figure 3.14). The filtered results were considered in a gene set enrichment analysis applying* `webGestaltR` *($FDR \leq 0.05$) (see enrichment outcome:* Data Chapter3 (TenStreams)/master_enrichment_WGCNAwDEA.csv *PW: PhD_SKraemer). The significantly enriched results per significant MTC are shown. (nER: normalised enrichment score, FDR: false discovery rate)*

MTC	Geneset	Description	nER	FDR	Size
purple-S10*	WP1367	TGF-β-Receptor Signaling Pathway	1.625	0.0199	10/10
purple-CG4	WP1348	Androgen Receptor Signaling Pathway	-1.830	0.0184	3/10
purple-CG4	R-DRE-1280215	Cytokine Signaling in Immune system	1.757	0.0457	14/14
lightcyan-CG7	WP467	mRNA processing	-1.736	0.0273	4/13
orange-CG7	dre00190	Oxidative phosphorylation	-2.325	< 0.0001	9/19
orange-CG7	WP1339	Electron transport chain	-2.325	0.0016	5/11
orange-CG7	R-DRE-1428517	Citric acid cycle	-1.918	0.0020	4/10
		& respiratory electron transport			
lightcyan-CG8	dre04210	Apoptosis	-1.998	< 0.0001	13/13

3.3 Discussion

The present study linked chemical exposure in ten streams in Minnesota to biological effects measured on the transcriptional level in liver tissue of adult fathead minnows. The mildly toxic acute exposures from selected surface waters were associated with xenobiotic effects with *in-vitro* bioassays and two out of three applied computational approaches. In addition, we showed that external references of biological genesets and toxicological databases helped validate the applied methods. In parts, we disentangled chemical responses from the chemical background in overall mildly toxic but existent perturbations in freshwater environments.

Responses due to xenobiotic stress and endocrine disruption as proof of concept for *in-vitro* activities. Nitrate and EE2 are markers of general pollution and endocrine disruption. Both were added to the chemical exposure data set based on their streamwise *in-vitro* assessments. The compound distribution alone did not help determine the driving (mixtures of) compounds leading to a xenobiotic adverse outcome. The occurrence pattern of EE2 and Nitrate could not be generated as a linear combination of occurrence patterns of detected compounds (see figure 3.2). If concentration addition would be assumed for mixture exposures, more compounds had to be relevant, which were not detected in the present investigation. Thus, the chemical analysis identified only a snapshot of potentially occurring chemicals. Nevertheless, by considering compound group exposure scenarios, it was possible to reliably attribute biologically meaningful effects to detected compounds concerning xenobiotic stress and endocrine disruption.

The biggest compound group (CG1) comprise ten compounds, and two — inorganic Nitrate and Metformin — are known for similar toxicological effects [Cordero-Herrera et al. 2020], such as inducing xenobiotic stress and affecting the endocrine system [Lin et al. 2020, Kellock et al. 2018, Bjerregaard et al. 2018, Pottinger 2017]. Straub et al. [2019] verified the formation of metformin ($C_4H_{11}N_5$) to nitrate in the aqueous environment, albeit bioaccumulation of Metformin in fish has not been proven. Considering both - single compound and compound group exposure scenario - Metformin and Nitrate hint to chemical-specific effects detected on the pathway level (see figure 3.17). Although not the top FDR-ranked terms to CG1, but a significant amount of enriched terms significantly overlapped to known pathway interactions in CTD to both named compounds (see table 3.6).
Toll-like receptor cascades were significant and highly ranked chemical-gene interactions to Metformin and nitrate in CTD for both exposure scenarios. Furthermore, the top-ranked enriched term for CG1 — *Cholesterol biosynthesis* — suits to exposure of Metformin, as

the compound has been used to treat diabetes mellitus type II affecting insulin sensitivity, although not listed in CTD for the respective compound. For DEA, the biological response was less influenced by chemical background noise when considering compound groups. For example, the top-ranked enriched terms for single compound Metformin (e.g. *Ribosomal proteins* and *Nonsense-mediated decay*) might be rather general and potentially driven by xenobiotic perturbation and did not overlap to known interactions from CTD. The overlapping terms to CTD were relatively similar for both exposure scenarios. However, the effects driven by Metformin and Nitrate were biologically and toxicologically meaningful considering the exposure of CG1, but also the transcriptional effect on a gene level was less similar to the general treatment exposure scenario. Therefore, a more specific response was determined when considering a group of compounds that might support a xenobiotic effect induced by Metformin as one potential and reliable main driver.

Furthermore, the exposures to the single compound EE2 and CG4 were biologically meaningful in DEA and WGCNA. For example, 87 significantly enriched terms were identified in functional enrichment of DEA results with at least one significantly DE gene to EE2 (see figure 3.6). In adddition, some biologically meaningful terms represented known EE2 interactions (see figure 3.17 and table 3.5). The terms were related to the endocrine activity (e.g. *Insulin signaling*, *gonadotropin-releasing hormone signaling*), immune responses (e.g. *interleukin-3 signaling*, *T-cell receptor signaling*) or potential tumorous processes (e.g. *Epidermal Growth Factor Receptor1 signaling*, *PI3K/AKT signaling*) *.

In WGCNA, the same three modules were significantly correlated to EE2, 5,6-Dimethyl-Benzotriazole and CG4 (see figure 3.14). The functional enrichment with the purple module to EE2 resulted in four and for CG4 in three significantly enriched terms (see table 3.4). Purple genes were significantly enriched to the *Insulin signaling pathway* in both cases, which has a central role in vertebrate endocrine regulation. Around half of the enriched genes were assigned to the purple module.
CG4 was associated with the compound group consisting of EE2, Ciprofloxacin, Naproxen and 5,6-Dimethyl-Benzotriazole. Considering this compound group allowed interpreting the detected compounds as potential chemical drivers of *in-vitro* measured endocrine activity in stream waters. However, the set of known EE2-pathway interactions in CTD lists a broad palette of cellular stress responses, not only endocrine disruption. Thus, the overlaps of

* As we consider highly correlated compounds, different interpretations might be possible. For example, the result could be a proof of concept, that DEA can be used to identify *in-vitro* measured endocrine effects of stream waters also with transcriptional omics data. Or that an EE2 exposure pattern covered streams with the highest toxic effects, leading to cellular stress due to xenobiotic perturbation and metabolic alterations partly hinting to endocrine disruptive processes.

enriched terms to known EE2-pathway-associations were high (single compound: 3/4, compound group: 3/4), but these terms might also fit a rather generic xenobiotic stress response. Another compound of CG4 - Ciprofloxacin - is an antibiotic known to be hepatotoxic but was not recently studied as an endocrine disruptor. According to CTD, the compound might induce liver injuries and stress induced by xenobiotic perturbation, leading to inflammatory processes. As Ciprofloxacin was detected only in one selected stream, the mixture of non-detected compounds or other detected compounds might induce a cumulative effect. The given data did not allow distinguishing mixture effects in more detail. Few Naproxen-associated pathways in CTD overlapped with CG4-enriched terms, but Naproxen is known for endocrine activity in fish [Kwak et al. 2018]. Of all detected compounds, 5,6-Dimethyl-Benzotriazole is the least investigated in (omics) exposure studies. This compound was not represented in CTD. A PubMed search identified two papers in total considering this compound in exposure or ecotoxicological context. Nevertheless, Benzotriazoles as a chemical group became more studied in the last years, and sublethal chronic effects leading to carcinogenicity and endocrine disruption were reported [Shi et al. 2019, Feng et al. 2020, Liang et al. 2014].
To summarise the effect, potentially driven by exposure to CG4: The biologically meaningful enrichment results were partly supported by literature and the external database CTD. Nevertheless, it remained unclear if all or even more compounds were essential to this (mixture) effect.

Different rankings for GSEA were shown as biologically meaningful for WGCNA results. Network inference approaches became popular to investigate data on the gene level in different areas of ecotoxicology [e.g. Williams et al. 2011, Perkins et al. 2011, Orsini et al. 2018, Barel and Herwig 2018]. For example, mutual-information-based modules were identified with `ARACNE` [Williams et al. 2011], or protein-protein interaction networks were investigated in exposure studies [Barel and Herwig 2018], or reverse-engineering networks helped to identify AOPs [Perkins et al. 2011]. In this study, the exposure to chemical compounds were linked to transcriptional effects applying WGCNA as generating a gene correlation network is a frequently used approach in omics analysis [e.g. Sutherland et al. 2018, Maertens et al. 2018, Degli Esposti et al. 2019]. A plethora of human-driven decisions is necessary concerning parameter settings in WGCNA. The setting decisions might be made differently by others, which would potentially lead to different results. In the present study, the biweight mid-correlation was calculated as it is more robust than the Pearson correlation [Langfelder and Horvath 2017] *. Also, the soft threshold might be chosen differently to fulfil the scale-free-

* In cases of chemical vectors, many zeros due to rare detection occurred, which would influence the calculation of biweight mid-correlations. Note that in such cases, the calculation is performed with Pearson

topology-criterion by other researchers. Here, the recommendations by WGCNA-developers were followed [Zhang and Horvath 2005].

The MTC values were generally low in this study. Therefore, the correlation threshold was set lower to consider a more extensive set of significant MTCs. In statistics, an absolute correlation value equal to or above 0.5 is usually considered a strong correlation, and $cor \geq 0.5$ is often used as meaningful filter for correlation analysis. However, this threshold might also limit the yield of potentially meaningful results. None to only a few relations would be considered in studies with overall lower correlation values. In the applied WGCNA approach, only one MTC was above 0.5 ($bicor = 0.53$ lightcyan module and 5,6-Dimethyl-Benzotriazole). However, no biologically meaningful result was identified for this MTC, which contained at least one gene of the lightcyan module. In such cases, it might be helpful to consider mediate correlations, which are commonly considered in the range of $0.3 \geq cor \geq 0.5$ in statistics. The results should be interpreted with more caution but might represent biologically meaningful. A subtle transcriptional effect was expected when measuring mild acute toxic effects in selected streams. However, the low correlation threshold helped identify weakly but present transcriptional effects.

In the case of WGCNA, the application of GSEA seemed meaningful. Gene significance and module membership prioritised and ranked genes in GSEA. For some significant enrichment results, the number of module-member genes was greater than 10 (e.g. for significant correlation to CG4-purple in table 3.4). This study was the first to consider ranking metrics of WGCNA to identify biological enrichments and with that exposure-pathway-associations. A rather unspecific xenobiotic response was expected when investigating acute exposure to lowly concentrated mixtures. The application of GSEA with WGCNA-metrics for gene ranking was partly meaningful and helped interpret exposure in stream waters on pathway level, albeit effects represented a rather unspecific xenobiotic response [*].

Regarding method integration concepts for functional enrichment, we also performed a GSEA combining the results of WGCNA and DEA. The combination of DEA and WGCNA resulted in biologically meaningful results and highlighted the importance of method integration to gain biologically meaningful and reliable results. In the case of the MTC CG4-purple, the enriched terms were related to endocrine disruption suiting to the selected compound subset EE2 Ciprofloxacin, 5,6-Dimethyl-Benzotriazole and Naproxen.

In summary, both enrichment approaches for WGCNA were meaningful, and both are rec-

correlation instead automatically in the `WGCNA`-package.

[*] see enrichment outcome: Data Chapter3 (TenStreams)/master_enrichment_WGCNA.csv PW: PhD_SKraemer

ommendable for functional analysis.

AR was not proper to disentangle chemical exposure and biological effects. Recent toxicological and biological studies presented AR as a meaningful approach [Creighton and Hanash 2003, Chen et al. 2019, Nagata et al. 2014]. On the one hand, exposure analyses were performed considering AR [Toti et al. 2016, Kapraun et al. 2017]. On the other hand, R-packages were developed to not only investigate any transaction set [Hahsler et al. 2005, Pinyaga et al. 2002] but specifically for omics data [Chen et al. 2019].

It might have different reasons that AR results under chosen filter settings were toxicologically uninformative and performed worse than DEA and WGCNA. In this study, 69 transactions were considered, which was quite large for an omics data set in an ecotoxicological scope to investigate stream water samples, but small in contrast to the scope of other AR studies [e.g. Kapraun et al. 2017, Bell et al. 2016]. At the same time, the large itemsets of biological and chemical entities in ranges of tens of thousands and greater when considering omics data lead to an immense need for memory and consumption of time. The smaller number of transactions reduced the chance to identify meaningful and non-trivial association rules, which made the transfer of AR to ecotoxicology challenging.
Furthermore, AR is generally based on arbitrary human-decided filtering thresholds, potentially influencing the number of frequent associations. Low support, high confidence and an additional lift filter reduced the set to a few frequent rules. Furthermore, only one-item-to-one-item rules were determined, which reduced the computational efforts for AR immensely — applying such a version of the apriori algorithm reduced the set of possible rules and equalised the limitation of high-dimensional itemsets.

Identical exposure patterns in stream-wise, single compound and compound group exposure scenarios also affected AR. The exposure data with too few sites for the number of compounds were limiting in the context of AR. To overcome such a limitation, the number of compounds would have to be smaller than the number of stream sites, and each compound should be measured at least two times. Furthermore, AR needed binarised input data, which reduced the information richness of numerical exposure patterns. The compounds with numerical concentrations across the identical set of streams had identical binarised exposure patterns. The exposure-associated genes became associated with multiple compounds, although only one might be truly relevant. Recent studies aimed to overcome such restrictions of binarised data and used a classified or weighted AR instead [Nagata et al. 2014, Lakshmi and Vadivu 2019]. The here chosen parameter setting identified rules with majorly frequently occurring genes. Nevertheless, neither cellular stress responses nor immune system-related terms were signif-

icantly enriched. In AR, support, lift and confidence were tested for functional enrichment, but none significant functionally enriched term was identified.

Further measures are applicable to validate frequent or strong rules [Piatetsky-Shapiro 1991]. As the present study applied a low support threshold, the measure of support ratio was used to interpret determined frequent rules. This measure helped to understand the determined rules. The exposure-associated genes tend to be regulated across the majority of microarray samples, independent of the specific exposure scenario. Furthermore, the more often a compound was detected, the higher the average support ratio was. Thus, most rules represented interactions with genes, which were affected across many samples, and a frequently occurred ANTECEDENT had an increased overlap with CONSEQUENT occurrences. Consequently, a rule did not necessarily describe the likeliness of an altered gene expression due to a specific exposure scenario. The exposure-associated genes might be rather general xenobiotic stress-regulated genes, less specific for an exposure scenario, e.g. single compound exposures. Overall stress responses instead of compound-specific biological effects might be expected to be measurable on pathway level. This outcome supports the stream-wise and single compound DEA results, where co-dependencies of exposure patterns led to high similarities to the general perturbation exposure scenario. Although AR alone might be limited in its outcomes in this ecotoxicological study, the approach might be helpful as a support system for the other approaches. It can be expected to extract potentially reliable results when identifying exposure-related gene associations with multiple approaches. In other investigations [e.g. Liu et al. 2013], a combined consideration of AR and DEA was shown as useful and identified meaningful results. However, significant results were identified neither for ORA nor for GSEA. In the present study, support, confidence and lift were all contemplated as potential ranking metrics for GSEA, but none suited for a meaningful application of GSEA. Other metrics in AR might be helpful to rank genes for GSEA. For example, Chen et al. [Chen et al. 2019] developed an R-package to investigate multi-omics data (considered as gene transaction sets) and retrieved biological meaning with a new measure. Transcriptional data alone would also be applicable in such investigations, but considering exposure would be still challenging. However, this was out of the scope of the present study to investigate this further. Although chosen parameter settings resulted in no biologically meaningful outcome, we encourage to apply AR as it was proven to be meaningful in a toxicological context by others [Barrera-Gómez et al. 2017, Kapraun et al. 2017, Santos et al. 2020, Creighton and Hanash 2003, Mallik and Zhao 2017, Karel and Klema 2007]. However, another strategy or research question might be needed.

Disentangling exposure effects with different exposure scenarios. The transcriptional effects represented alterations partly due to stress and immune responses on the tran-

scriptional level. The applied **general treatment exposure scenario** assumed any xenobiotic response in stream waters, but did not identify compound-specific effects within complex mixtures. This exposure scenario helped to understand the overall picture of chemical exposure and might describe a widespread background of non-detected but ubiquitously distributed compounds responsible for most transcriptional effects. In the case of DEA and WGCNA, xenobiotic responses like stress (e.g. *interleukin3-signaling pathway*), metabolic changes (e.g. *metabolism of poly amines*) or also endocrine effects (e.g. *Insulin signaling pathway*, *gonadotropin-releasing hormone-signaling pathway*) were identified. These findings emphasise the activities measured in the *in-vitro* bioassays on a transcriptional level with the expected biological meaning. With that, the biological effects supported the finding of mildly acute mixture toxicity induced by a smaller set of detected compounds and an unknown chemical background of lowly-concentrated compounds.

Freshwater bodies are burdened by hundreds to thousands of contaminants [Bradley et al. 2019, Busch et al. 2016]. In natural environmental settings, site-wise exposure scenarios seem to be unable to disentangle compound induced biological effects. Considering **stream-wise and single compound exposure scenarios**, the high overlaps to the general treatment were induced by the co-correlations of exposure patterns. The stream-wise exposure scenario investigated mixtures of occurred compounds per site, whether detected or not. Some compounds were detected in only one stream. For example, the exposure to Ciprofloxacin (measured in stream10 only) resulted in a large set of associated and DE genes but no significantly enriched terms. This unspecific differential expression is more likely to relate to the overall mixture response within this stream than to one detected compound. Thus, a site-wise exposure scenario did not allow differing between a single compound and a mixture exposure effect. Interferences with other compounds, detected or not, might amplify or nullify chemical-specific effects. The unknown chemical background might induce noise in the transcriptional data. A potentially increased error rate in single compound exposure scenarios led to high amounts of false positive results (see figure 3.7). Consequently, each compound scenario contained indirectly other compounds and unknown chemical background, which might induce a similar DEA outcome to the general treatment. The present study showed that the assumption of independent chemical variables in single chemical exposure scenarios was a strict simplification. Thus, a single compound scenario hardly disentangled exposure-related transcriptional effects.

Different ideas were already reviewed for exposome studies assessing correlations of exposure and biological effect [Santos et al. 2020]. The applied variation of a variable selection resulted in a consideration of **compound groups**. The chemical groups of highly correlated

and dependent compounds were set as exposure scenarios to overcome the weakness of the single compound scenario and to consider a smaller set of covariates with weaker pairwise correlations on average. We expected to model the "everywhere hidden chemicals" primarily as an intercept in such a differential expression context, wiping out the effects associated with a general treatment.

To determine chemical compound groups, no structural similarities of chemical compounds were considered but pairwise correlation of exposure patterns across streams. The clustering to exposure-related compound groups was based partly on human decisions. For example, assigning 5,6-Dimethyl-Benzotriazole to CG4 might be grouped differently by other researchers. Furthermore, the selected set of streams was small for an exposure pattern comparison *. In the case of the detected compounds, two might be highly correlated but only one compound might be toxic for fish. The overall mild exposure might justify the investigation of effects due to concentration patterns instead of similarities due to chemical structure. However, the chosen grouping improved the ability to investigate chemical exposure on the transcriptional level and resulted in meaningful and reliable outcomes.

The unspecific xenobiotic effects were potentially reasoned by lowly exposed mixtures [Yuxuan et al. 2019, Gandar et al. 2017] as non-detected compounds, which are unequal zero, add up to a mixture effect [Escher et al. 2013]. Hundreds to thousands of chemicals might be contained in the stream waters [Liška et al. 2015, Bradley et al. 2017, Kolpin et al. 2002]. Based on the chemical analysis, a significant chemical background remained unknown, which led to a vaguely estimated chemical exposure. The compound group exposure scenarios were more meaningful than a single compound approach as a linear combination of single compound exposure patterns led to reduced dependencies between chemical covariates. With that, also a part of chemical mixture effects was investigated under the assumption of concentration addition. However, the generation of those chemical groups was still artificial and based on a small set of compounds and stream sites. Technically, compound groups with only one compound could directly disentangle the chemical-specific biological effects of single compounds. For example, the compound group exposure scenario to CG7 was directly interpretable as exposure to Iopamidol and resulted in cellular stress responses on pathway level (see table 3.3). However, also multi-compound groups were biologically meaningful. For example, CG4 was reliably associated to biologically meaningful effects on gene and pathway levels.

* The fewer sites, the coarser grained are exposure patterns.

3.4 Conclusion

Stream water samples might contain thousands of compounds, and the majority is lowly concentrated or even not detectable, albeit present. Such mixture perturbations may induce rather unspecific but stress-responsive effects. This study was focused on acute effects in fathead minnow liver tissue induced by ecological mixtures of freshwater samples. The investigated mixtures comprised lowly affecting chemical compounds, and, according to the *in-vitro* bioassay activities of nitrate and EE2, some induced xenobiotic stress and were endocrine active. Using omics data and measured chemical concentrations, we applied three different computational strategies to link chemical exposure to acute xenobiotic effects in fish. The objective was to identify whether the applied approaches were suitable to identify chemical-specific exposure effects from mixtures of lowly concentrated chemical compounds. Moreover, we were interested in determining biologically meaningful and reliable attributions of exposure-related effects to chemical drivers.

As a stand-alone approach, AR was not able to identify biologically meaningful outcomes and did not allow disentangling chemical exposure in respect to the given data and the applied settings. Still, AR might be useful for environmental toxicology to link chemical exposure to biological effects but to answer other research questions and consider not only one-exposure-to-one-gene association rules. Furthermore, it could be beneficial to consider data with more samples and more chemical compounds to facilitate the mining process.

DEA has been applied already successfully to distinct biological effects of drivers in mixtures. This study confirmed the importance of DEA to determine exposure-related biological effects but also highlighted its limitations when considering not independent and lowly concentrated chemical exposures.
In single compound exposure scenarios, the dependencies of exposure patterns and the subtle transcriptional responses limited the DEA outcome and affected the biological and toxicological interpretability negatively. In a compound group exposure scenario, the present study overcame the limitations of mild concentrations and not independent contaminants with exposure-correlated mixtures. Compound groups modelled reality better than single compound assumptions, especially when aiming at disentangling the biological response regarding chemical exposure.
The functional enrichment approach and the consideration of external toxicological knowledge from CTD and STITCH helped make sense of determined chemical-gene interactions on a higher biological level and inferred a biological meaning to exposure-associated tran-

scriptional effects. Transcriptional effects were, in part, significantly associated with known chemical-pathway associations in CTD when applying DEA.

The network inference approach WGCNA also resulted in biologically meaningful outcomes and supported the *in-vitro* bioassay results with effects of xenobiotic stress and endocrine disruption. Moreover, a method integrative functional enrichment, with DEA and WGCNA outcomes, associated *in-vitro* activities with subsets of the potential driving compounds. Consequently, the functional enrichment and method integration helped determine and disentangle the xenobiotic effects of exposure groups. The applied strategies allowed a more systematic risk assessment when concerning exposure scenarios and method integration.

This study investigated complex environmental data considering chemical exposure in freshwater and measured transcriptional effects. Although chemical analytics resulted in a somewhat small amount of detected compounds, the data became complex through the exposure co-dependencies. Furthermore, the transcriptomic effect data are high-content data comprising measurements of thousands of transcripts and, thus, complex. Thus, we investigated acute and mildly toxic exposures of the freshwater environment and somewhat unspecific xenobiotic responses. Nevertheless, this investigation allowed associating measured chemicals to *in-vitro* bioassay effects and determine a fitting response, including hints towards an endocrine disruption in fathead minnow liver tissue. Additionally, the presented limitations and the workflows to evaluate biological meaning and reliability are advantageous methodological knowledge for future studies, which may aim to disentangle chemical exposure effects with omics data.

Chapter 4

Deep learning prediction of chemical-biomolecule interactions

This chapter presents a novel approach combining text-mining and deep learning to generate novel hypotheses of stressor induced effects on the molecular level from the toxicological knowledge hidden in the literature and databases. We employed text-based data from SemMedDB and curated toxicogenomic knowledge from the comparative toxicogenomic database (CTD). The aim was to classify the relationship type as either stimulating or inhibiting, considering a particular chemical linked to a specific biomolecule for each relation. We implemented the task with a word embedding neural network and a subsequent feed-forward network. Furthermore, the Unified medical language system was employed to augment input data using ontology hierarchies.

We developed a deep learning approach that derived hidden knowledge in text-based resources from toxicological literature. The model trained with SemMedDB data reached an accuracy of up to 70% for independent test data. Furthermore, data augmentation and implementation with recurrent neural networks were beneficial for training with CTD data and resulted in an accuracy of 94%. Finally, we used data integrative application cases to evaluate the prediction models based on biomedical literature and curated toxicogenomic knowledge. However, the SemMedDB model was not able to reliably confirm the chemical-gene interactions in CTD data and vice versa.

Still, the predictive tool allowed identifying hidden knowledge from PubMed literature resources and transcriptomics-based exposure studies concerning the input data, respectively. The study was entirely data-driven and involved state-of-the-art computational methods of artificial intelligence.

4.1 Motivation

Thousands of chemical compounds and their degradation products are released into the environment on different traces due to their use in agriculture, industry, and households. Such anthropogenic exposures may lead to adverse effects on organisms in the environment. Therefore, a chemical perturbation on the biochemical/molecular level may provoke an adverse outcome (AO) on the organism or the population level. It is necessary to determine the exposure-related molecular interactions to understand the reasons for an AO and investigate it in a systems biology context. The previous study stressed such an investigation concerning detected complex mixtures and omics-based exposure studies. Next to empirical measurements, the entire current toxicological knowledge may help link chemical exposure to biological effects comprehensively.

The adverse outcome pathway (AOP) describes the toxicological knowledge in a sequential and modularised manner as a chain of key events (KE), leading from a molecular initiating event (MIE) up to an AO across the different *levels of biological organisation* (LOBO) [Ankley et al. 2010]. Still, further relevant toxicological knowledge would help describe and quantitatively prove KEs or even an AOP potentially. For example, AOP networks have led to new recombinations of existent KEs by merging or concatenating recent AOPs [Knapen et al. 2018, Villeneuve et al. 2018, Pollesch et al. 2019] and, thus, not yet investigated hypotheses have been computationally retrieved. The toxicological knowledge out of scope of recent AOP knowledge curation contains further potential unconsidered information and has been extended through data integration approaches [e.g. Nymark et al. 2018, Martens et al. 2021, Aguayo-Orozco et al. 2019, Jornod et al. 2020]. For example, Aguayo-Orozco et al. [2019] has linked activity measures of the ToxCast program [Dix et al. 2007] with the AOPwiki and summarised their efforts on the sAOP-webpage — a network representation of the AOP-knowledge with stressor-interactions for 6000+ chemical compounds and 200+ AOPs. Further external data resources have successfully retrieved new toxicological knowledge [Bell et al. 2016, Nymark et al. 2018, Wang et al. 2019, Watford et al. 2018]. For example, the *Comparative toxicogenomic database* (CTD) [Mattingly et al. 2003, Davis et al. 2021] has been frequently used [e.g. Oki and Edwards 2016, Rugard et al. 2020, Perkins et al. 2017] in environmental toxicology (ET). Furthermore, a plethora of literature is publicly available, and literature mining approaches have also emerged in ET [e.g. Carvaillo et al. 2019, Zgheib et al. 2021, Jornod et al. 2020; 2021]. Still, literature-based computational approaches have rarely been employed to comprehensively link chemical compounds to biomolecular effects, considering

entire knowledge bases. Thus, although chemical interactions have been well-investigated in toxicology, novel knowledge integration strategies are crucial to gain further information from current knowledge to a greater extent.

The National Library of Medicine and the National Institute of Health have provided essential tools for human health-centred toxicological research, which resulted in a broad biomedical data infrastructure, e.g. with MEDLINE [Ahlers et al. 2007], SemMedDB [Kilicoglu et al. 2012] and the *Unified Medical Language System* (UMLS) [Humphreys et al. 1998, Bodenreider 2004]. The UMLS comprises three main biomedical knowledge entities, which combine vocabularies and standards to enable interoperability between computer systems. Different data integration approaches or tools have employed the UMLS already [e.g. Kilicoglu et al. 2012, Martens et al. 2021]. Also, a comparative study has shown the ontological strengths of the UMLS by applying embedding approaches [Mao and Fung 2020]. However, the UMLS is a biomedical ontology and thus biased towards the human health domain. Researchers have already expanded the UMLS ontology to further domains such as pharmacogenomics or medical informatics [e.g. Ahlers et al. 2007, Rosemblat et al. 2013b], but not ET.
SemRep is a semantic analysis tool based on the UMLS resources [Rindflesch and Fiszman 2003, Rindflesch et al. 2005]. It has frequently been used to extract semantic relations in a biomedical context. Kilicoglu et al. [2012] have generated a large-scale knowledge resource called SemMedDB based on SemRep. The SemMedDB contains semantic predications retrieved from the titles and abstracts in PubMed. The predications have been ontologically unified with the UMLS terminology. SemMedDB has been an essential resource for literature-based discovery in a biomedical context [Kilicoglu et al. 2020, Kastrin et al. 2018, Cameron et al. 2013, Gao et al. 2021]. Still, SemMedDB may also be an important data source for the related scientific domain of environmental toxicology. Recent studies have applied the SemMedDB data, e.g. to develop a graph of interacting semantic predications [Hristovski et al. 2015, Cong et al. 2019] or identify causal drug-side-effect-relations [Mower et al. 2017]. In this context, (semi-)supervised machine learning approaches have been used to predict drug-disease relations [Rastegar-Mojarad et al. 2016, Bakal et al. 2018] or drug-side effect relationships [Mower et al. 2018]. For example, Bakal *et al.* [Bakal et al. 2018] have predicted causative relations between drugs and diseases. They have reached F1-scores of at least 90% for their test data. The pharmacovigilance prediction of Mower et al. [2018] has also used word embedding and composite feature vectors and has resulted in F1-scores of 90% and 84% for applied evaluations. Thus, the relational knowledge from SemMedDB has been suitable for predictive machine learning tasks, also with considerations of knowledge representation via word embedding. Deep learning technologies have been frequently used

in current literature- and knowledge-based discovery, in particular for biomedical tasks, with recurrent neural networks — e.g. the long-short-term-memory (LSTM) [e.g. Lee et al. 2019, Lai et al. 2021] — convolutional neural networks [e.g. Peng et al. 2018], or attention-based transformers — e.g. *Bidirectional Encoder Representations from Transformers* (BERT) [e.g. Peng et al. 2019, Michalopoulos et al. 2021]. However, literature-based knowledge has rarely been employed in ET research to link chemical effects to molecular effects [e.g. Zgheib et al. 2021, Jornod et al. 2020].
To increase the robustness of trained models, data integration has been applied for the pre-processing of input data [e.g. Lai et al. 2021, Choi and Lee 2019] or the retro-fitting of the prediction model [e.g. Zhang et al. 2019b]. Data integration can be considered data augmentation when input size increases by adding slightly modified samples or synthetic data. As a recent example, Lai et al. [2021] have used SemMedDB data augmented with synthetic negative samples to train a subsequent model with word embedding and LSTM. As a result, the models accurately predicted key event like hypotheses. However, negative sampling has a limitation for knowledge-based input data, as not yet investigated relations are considered false but might be proven true in the future. Still, such studies have highlighted the applicability of SemMedDB in deep learning prediction tasks also concerning toxicology-relevant research fields as AOP development.

This study applied a knowledge-based discovery approach for toxicological purposes. The study concentrated on chemical-biomolecule interactions — as an essential link of chemical exposure with molecular effects — and might direct towards MIE prediction in the future. We were interested in whether deep learning models with semantic representation could learn from knowledge-based input and whether the prediction of relations helped retrieve toxicologically meaningful outcomes.
The aim of the investigation was (1) the development of deep learning models that utilise public available text-parsed biomedical or curated toxicological knowledge to predict relationship types of chemical-biomolecule pairs, and (2) the evaluation of these models based on their ability to predict current knowledge from PubMed and a toxicogenomic source.
Architectures of word embedding models with and without LSTM neural networks were trained, tested and evaluated with chemical-biomolecule relations from SemMedDB and CTD. Furthermore, we examined the effect of data augmentation — exploiting the UMLS terminology — for literature-based input on the selected deep learning models.

4.1.1 Workflow

The presented workflow in figure 4.1 presents the general workflow * of this study for the example of SemMedDB input data. The literature-based knowledge from SemMedDB [semmedVER43_R, through June-23-2020 Kilicoglu et al. 2012] was downloaded (see section 2.2.1) and prepared as UMLS-annotated chemical-biomolecule relations with two gene regulatory relationship types (see section 2.2.2). The considered UMLS terminology is a standard ontology in biomedicine. It helped assign toxicological terms to levels of biological organisation and filter chemical-biomolecule interactions from SemMedDB. The relation data were split in training and validation set and augmented vertically and horizontally (see section 2.2.2). Deep learning models were trained to predict the relationship of a chemical-biomolecule pair based on current toxicological knowledge. The models included layers of word embedding, long-short-term memory, time-distributed layers and dense layers. Two model architectures were determined and adapted via hyperparameter tuning (see section 2.2.3). The architectures were tested in a 5-fold cross-validation approach with the three types of input, respectively. The CTD was variously applied for data augmentation, linking chemical-gene interactions to higher biological levels and evaluating prediction results (see section 2.2.4).

Additionally, we (1) investigated chemical-biomolecule interactions from SemMedDB with four instead of two relationship types and (2) with UMLS-annotated chemical-gene interactions given in CTD. In both cases, unseen test data were considered for model evaluation. Moreover, the CTD trained models were evaluated with SemMedDB chemical-biomolecule relations.

The work was primarily performed in python [version:3.6 Van Rossum and Drake 2009] using the deep learning packages `Keras` [version:2.4.3 Chollet et al. 2015], `TensorFlow` [version:2.5.0 Abadi et al. 2016], `scikit-learn` [version:0.0 Pedregosa et al. 2011], and `KerasTuner` [version:1.0 .3 O'Malley et al. 2019]. In addition, some figures were prepared in the R statistical programming language [version 3.6 R Core Team 2020].

* Data and code are available here: https://nc.ufz.de/s/emqxbigeWYPSnKp (*'Data Chapter4 (KEpredict)'* and *'Code Chapter4 (KEpredict)'*) with the following password *PhD_SKraemer*

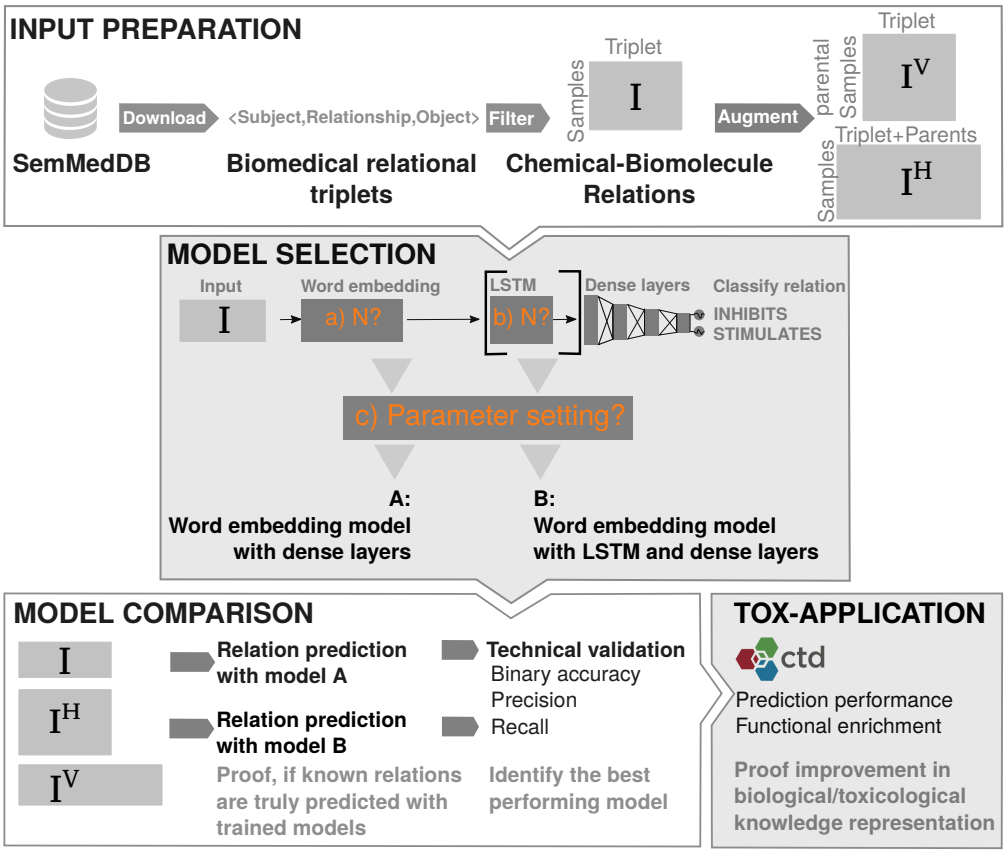

Figure 4.1. ***Workflow of the deep learning approach to predict chemical-biomolecule relationship types.*** *After the download of literature-based knowledge from SemMedDB [semmedVER43_R, through June-23-2020 Kilicoglu et al. 2012], data were reduced to chemical-biomolecule relations. The relation data were split in training and validations set and additionally augmented in vertical and horizontal direction. By comparing the word embedding size or the numbers of LSTM-neurons, two architectures of machine learning models were determined and adapted via hyperparameter tuning. The architectures were tested in a 5-fold cross-validation approach with the three types of input respectively. The model performance was evaluated with unseen data from SemMedDB and the comparative toxicogenomic database (CTD).*

4.2 Results

4.2.1 Input preparation

Chemical-biomolecule relations. In total, the predication data set from SemMedDB consisted of 112 796 186 triplets of biomedical knowledge (see figure 4.2). The unique set of 20 918 831 triplets was considered, and subjects and objects were listed with the UMLS concept unique identifiers (CUI). After removing negated relations, where the predicate had the suffix 'NEG_', 19 486 620 triplets remained. By filtering CUIs for levels of biological organisation (LOBO), 1 659 209 chemical-biomolecule relations were determined. This subset comprised 46 831 chemical concepts and 40 683 biomolecule concepts, where 9734 concepts had both annotations as a chemical and a biomolecule concept. The predicate of the triplet was chosen as the classification target. Therefore, to consider a balanced data set and keep the variability small, the predicates were filtered to *STIMULATES* ($n = 215\,934$) and *INHIBITS* ($n = 250\,468$). With these predicates, the chosen chemical-biomolecule relations could be semantically interpreted as (bio-)molecular interactions. Some subject-object-pairs were represented in the data set with both relationship types and were removed. The finally filtered and UMLS-annotated input data set consisted 373 904 chemical-biomolecule predications and is designated as I in this thesis (see figure 4.2).

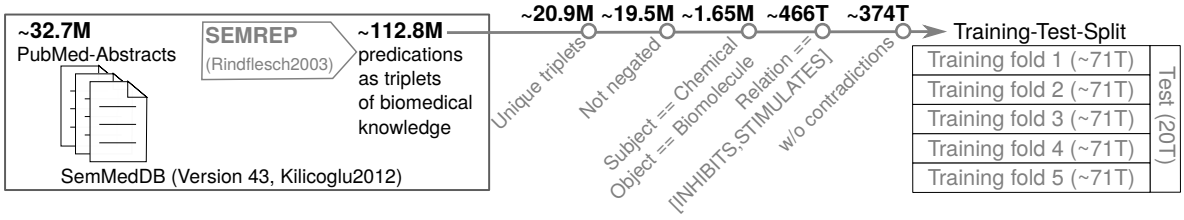

Figure 4.2. ***Preprocessing of chemical-biomolecule interactions from SemMedDB*** *SemMedDB contains pre-parsed and UMLS-annotated predications retrieved from PubMed abstracts and titles with the UMLS-tool SemRep . The downloaded predication data set was stepwise preprocessed to non-contradicting, non-negated and unique chemical-biomolecule relations with STIMULATES and INHIBITS as relationship type. The data set was split in a training (with five equally sized folds) and a test set.*

In summary, the prepared selection of predications comprised unique, non-negated, and non-contradicting chemical-biomolecule-predications with relations of stimulating or inhibiting interactions between chemical compounds and biomolecules. A test set I_E with 10 000 relations for each relationship type was randomly selected from I. It was ensured that each

chemical and each gene concept was represented at least once in the remaining data set – the training data I_T – which was split into five equally sized subsets ($I_{Ti},\ i \in [1,5]$) with the same ratios of relationship types (see figure 4.2). If a 5-fold cross-validation approach was applied, then for the i-th training fold, I_{Ti} was used as validation and $\bigcup_{j=1}^{5} I_{Tj}, j \in [1,5] \& j \neq i$ as training set. Else, I_{T1} was considered as validation set and $\bigcup_{i=2}^{5} I_{Ti}$ for training.

Data augmentation. One vertically augmented (I^V) and one horizontally augmented (I^H) version of I was generated (see section 2.2.2). Whereas I^V remained identical in the relational sequence length (n=2), I^H was prolonged (n=6). Due to the multiple hierarchically structured sources of UMLS Metathesaurus, both augmented inputs were expanded in their number of relations.

For a **vertical augmentation** of one chemical-biomolecule relation in I, either chemical or biomolecule was replaced by a direct parental concept given in the Metathesaurus. Due to the various sources in the Metathesaurus, each subject or object semantic concept might have multiple parental terms. Thus, one chemical-biomolecule-relation in I might result in multiple relations with parental terms. After vertical augmentation, contradicting relations might occur as I might already contain some parental concepts. These contradicting relations were removed. I was expanded by 774 014 chemical-biomolecule relations with parental terms. I^V consisted of 991 688 relations, and 42 759 were considered as test subset I_V^V. Consequently, I^V is nearly three times the size of I.

All the applied deep learning models contained a word embedding layer. Thus, the relations in I were read as a sequence of two related terms. However, a length of two might be not enough to learn semantic relationships and recent approaches using a word embedding consider text sequences, like sentences, directly or in at least longer padded sequences. The **horizontal augmentation** elongated the input sequences of I to lengths of six by adding (grand-)parental terms to chemical and biomolecule term (see figure 2.2). As already mentioned in the paragraph before, semantic terms might have multiple parental terms. The elongated data of I^H contained 3 605 625 not contradicting and unique sequences. As all parental terms were considered, the horizontal augmentation also led to an indirect vertical augmentation (see figure 2.2). I^H had nearly ten times more relations than I. The reasons for a more extensive vertical augmentation were the consideration of all recombinations of a two-sided elongation with parental terms with potential multiple parental terminologies and considered two levels of parental terms. Furthermore, some semantic concepts did not have a parental term in the UMLS semantic network, and respective parental terms were set to NA. Such sequences were also considered with lengths of six in I^H and were available input for the models, where the sequences were tail-padded with values equal to zero.

Coverage of input in recent toxicogenomic knowledge. The comparative toxicogenomic database (CTD) has been frequently used to extract chemical-gene or chemical-disease relations. This resource comprised exposure-related knowledge across different LOBOs and was applied to determine the coverage of SemMedDB known chemical-gene interactions in toxicogenomic knowledge (see figure 4.3).

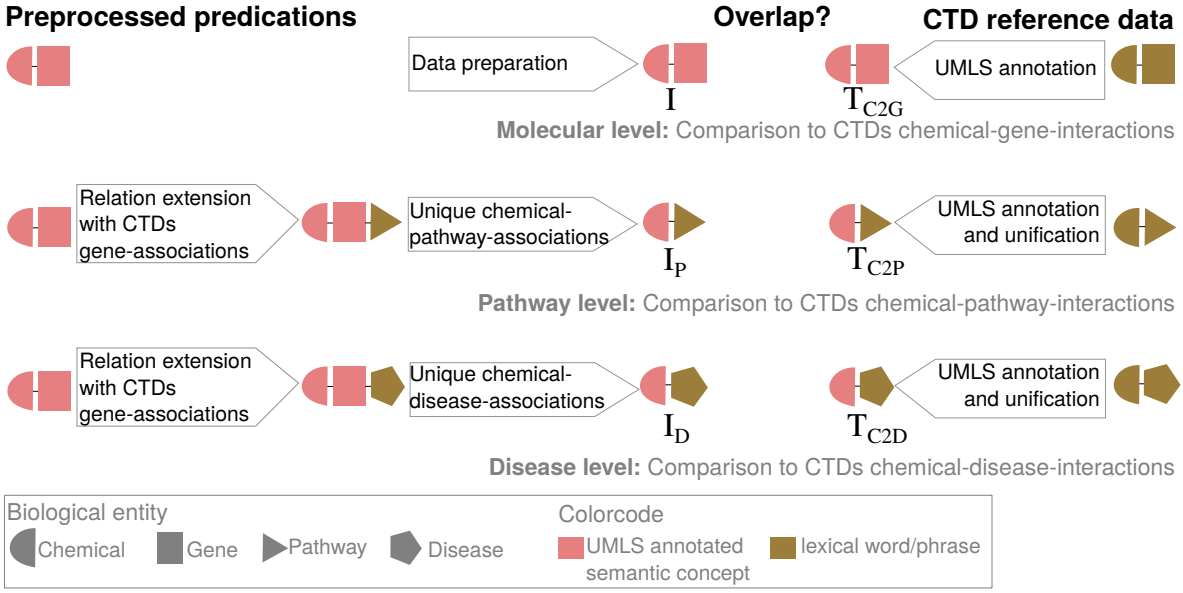

Figure 4.3. ***Coverage of SemMedDB in the comparative toxicogenomic database (CTD) was determined on gene, pathway and disease level.*** *The overlap of I to CTD's chemical-gene interactions (T_{C2G}) was determined. Furthermore, I was expanded with CTD gene-pathway- (T_{G2P}) and gene-disease associations (T_{G2D}) and compared to CTDs chemical-pathway- (T_{C2P}) and chemical-disease associations (T_{C2D}), respectively (see methods 2.2.4).*

Initially, chemical-gene interactions were downloaded and preprocessed for their use as prediction input (see table 2.2). In CTD, 134 relationship types had been considered for chemical-gene interactions. These were reduced to the UMLS relationship types *STIMULATES* and *INHIBITS*. Furthermore, the chemical names and genes in CTD were annotated to the UMLS terminology of CUIs. In total, the filtered set T_{C2G} contained 8 365 638 relations. The preprocessed input data considered genes and all biomolecules listed in the literature and included in the UMLS ($n = 19189$). On the other hand, CTD contained 37 253 unique UMLS-annotated genes and thus more objects than I. The overlap between I and T_{C2G} was determined as a reference of the coverage of recent toxicogenomic knowledge within input data I. Only those relations were considered, where the Subject-CUI and the Object-CUI were represented in I and T_{C2G}. Thus, 57 488 and 48 192 relations in I and 314 462 and 402 747 interactions in

T_{C2G} were considerable for the respective relationship types *INHIBITS* and *STIMULATES* comprising 6305 chemical subjects and 6016 biomolecular objects.
A matrix of all subjects times all objects represented the space of all possible combinations of chemical-gene-interactions in a data set and the given ratio of occurring relations the level of sparsity. Considering the space of subjects and objects, which were present in both data sets, we determined a sparsity level of 1.89% for T_{C2G} and 0.28% for I *. The level of sparsity in the matrix is considered here as a measure of information density, as it represents the amount of information in a given space. Thus, the information density of T_{C2G} is approximately ten times greater than for I.

Next to the comparison to T_{C2G}, the overlap to I on the pathway and disease level were determined to the toxicogenomic knowledge represented in T_{C2P} and T_{C2D}, containing 2 750 080 and 7 872 780 interactions, respectively (see figure 4.3). Therefore, I was expanded with the CTD gene associations either to pathways (I_P) or diseases (I_D). On these biological levels, the relationship types were not considered in CTD. In total, I_P comprised 3 771 710 unique chemical-pathway interactions and I_D 46 200 665 unique chemical-disease interactions.

On the molecular level, 8.5% and 6.35% input relations overlapped for the relationship types *STIMULATES* and *INHIBITS*, respectively (see figure 4.4 A). Thus, the coverage of toxicogenomic knowledge in I was relatively small.
Furthermore, 12.8% of I_P and 4.0% of I_D were covered in CTD (see figure 4.4 B and C).
In summary, the prepared input I covered the known toxicogenomic chemical-biomolecule interactions in CTD only by a small amount. Nevertheless, the increased coverage on the pathway level and the large set of millions of overlapping chemical-disease associations highlighted the importance of considering lower biological resolution levels. Consequently, we had to be cautious when considering biomedical information for a toxicological prediction task, but could expect to some extent that the toxicological information was also represented when learning a model.

4.2.2 Model selection

The predication triplets in I contained the model input — the subject-object-pair —, and the labelled target — the relationship type (*INHIBITS* or *STIMULATES*). Model architectures with an initial word embedding layer were trained with the input I to predict a relationship

* We also considered the sparsity level for the respectively full set of subjects and objects. In case of I, 373 904 relations with 25 761 subject concepts and 19 189 object concepts resulted in a sparsity of 0.076%. In case of T_{C2G}, 8 365 638 relations with 25 495 subject concepts and 37 253 object concepts resulted in a sparsity of 0.881%.

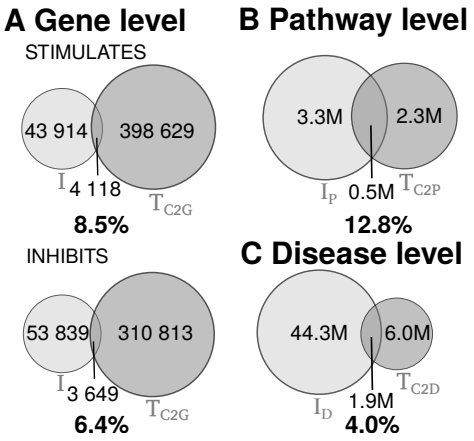

Figure 4.4. ***Determined coverage of SemMedDB in CTD on molecular, pathway and disease level.*** ***A)*** *UMLS-annotated chemical-biomolecule relations in I vs chemical-gene interactions from CTD (T_{C2G}). Each chemical and biomolecular concept in I is represented at least once in the UMLS annotated T_{C2G}.* ***B)*** *UMLS-annotated chemical-pathway relations from I_P vs chemical-pathway interactions in CTD (T_{C2P}).* ***C)*** *UMLS-annotated chemical-disease relations from I_P vs chemical-disease interactions in CTD (T_{C2D}).*

type for chemical-biomolecule pairs. Based on a selection of model architectures and hyperparameter tunings, the best performing was determined based on the minimal loss measured with binary cross-entropy and binary accuracy.

Word embedding model with dense layers. Initially, a deep learning architecture with an initial word embedding layer and a subsequent feed-forward network was implemented to determine the toxicogenomic relationships between chemical-biomolecule pairs. In a trained word embedding layer, each word was represented by a vector of the same length, and words with similar semantic meaning had a small Euclidean distance to each other. However, based on the complexity of the input semantics and richness in information, the word embedding size N should be chosen carefully. Five models with different word embedding sizes N were trained, evaluated and tested to assess the influence of N with the same data, I_T, I_E and I_V.

The UMLS-annotated chemical-biomolecule pairs in I were transformed to sequences of integer pairs *. The sequential **model architecture** consisted of an initial word embedding

* Each word of the vocabulary corresponded to another unique integer. Thus, in a one-hot-encoded vector matrix of the vocabulary, each integer i represented the row, where the i-th value was 1. The input representation in form of integers instead of one-hot-encoded vectors saved memory during training and application of a model.

layer and following dense layers with decreasing number of neurons. The sequences were first projected to a sequence of vectors of length N with the word embedding. The vectors reflected the semantic relationship of words within the trained vocabulary given the context of chemical-biomolecule relationships in the training data. A layer to flatten the sequence of vectors of length N to one vector of length $2 \times N$ was applied to adapt the hidden input for the following subsequent feed-forward network. In fully-connected neural networks, the vector lengths were stepwise reduced by a factor of 2 to 2.5 down to a size of 25. Each fully-connected neural network had a final ReLU-activation except the last using the sigmoid activation function to predict the probability p for the relationship type *INHIBITS*. If $p \geq 0.5$, the subject-object-pair was predicted as inhibited else as stimulated.

Five **models with different word embedding sizes** of $N = [100, 500, 1000, 2500, 5000]$ were trained. The models were compared based on their loss and accuracy performance validation curves and the ratio of correct predictions within the test set (see figure 4.5). The maximal number of training epochs was set to 1000 for a batch size of 1000. The left plot in figure 4.5 A shows the validation loss curve (binary cross-entropy) in a range from 0 to 3. In all five models, the loss decreased steadily. Thus, the model learned to predict targets more correctly. The loss converged around 0.6 for all models. The validation accuracy during training (right plot in figure 4.5 A) increased in all five models to 0.70.

The confusion matrices in figure 4.5 B presents the prediction outcomes for unseen data I_E. Out of 20 000 relations, the five models predicted 13 768 to 13 937 relationships correctly, and thus, reached accuracies of approximately 0.70 (see table 4.1). The models predicted relationships with *INHIBITS* more accurate, potentially induced by the slight off-balance of relationship types in I_T. Consequently, the precision in predicting *INHIBITS* was higher in all models, whereas recall was higher for *STIMULATES* (see table 4.1).
No trained model seemed clearly better than another based on all performance measures and learning curves over epochs. Thus, the most straightforward model architecture with the least number of parameters was chosen, which was, in this case, the model with the smallest word embedding vector size. Consequently, the output vector sizes of the word embedding layer were set to 100 in the following.

An **additional training** was performed for $n = 5000$ a learning rate of $\alpha = 1e - 6$ to test whether large word embedding sizes might benefit from training with a lower learning rate and, thus, a longer training duration. All five models stopped training after 40 to 108 epochs early, as the decrease in loss was less than 0.01 in a period of 20 epochs. The number of training steps increased, the smaller the vector size was. The choice of the learning rate did influence the training duration in terms of epochs as expected (see supplemental section S3.3).

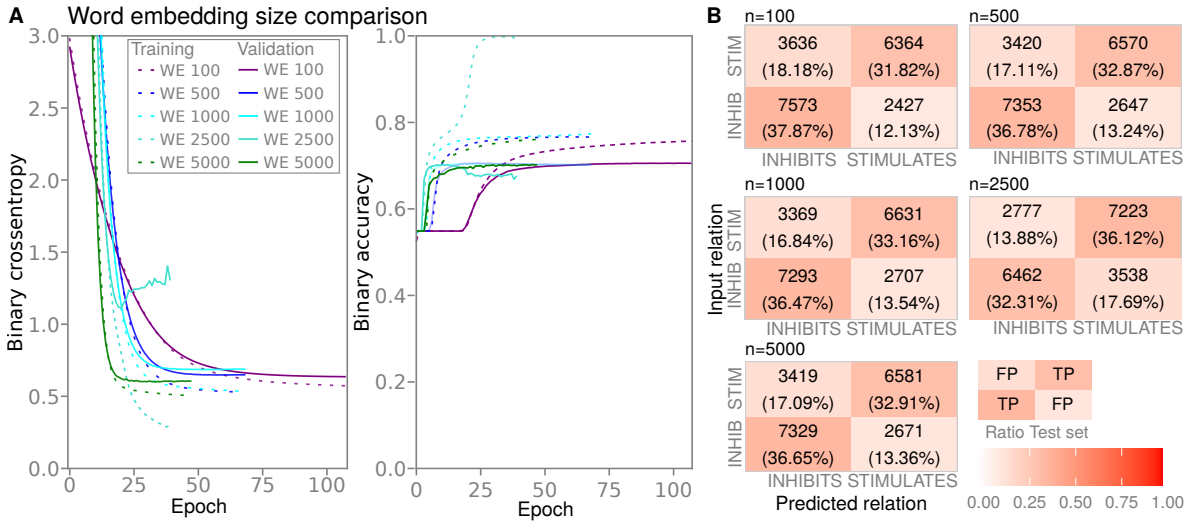

Figure 4.5. ***Comparison of the model performances with different word embedding vector sizes.*** *Deep learning models were trained with UMLS-annotated chemical-biomolecule-pairs retrieved from SemMedDB to predict their toxicogenomic relationship. The models have an initial word embedding layer followed by multiple dense layers, which decrease in size. The word embedding vector sizes were set to* 100 *(purple),* 500 *(blue),* 1000 *(orange),* 2500 *(red), and* 500 *(brown).* ***A:*** *The* `keras` *model performance across epochs was determined with validation set* I_{T1} *(n≈71T) and binary accuracy(dashed line) and binary cross-entropy (solid line) were tracked during training. The models trained for different numbers of epochs as early stopping occurred, when loss decreased less than* 0.01 *in a period of twenty epochs. The loss converged around* 0.6 *for all models. All accuracies reached similar values of* 0.70. ***B:*** *Confusion matrices for prediction results to the test set* I_E *(n=20T) for all five trained models. The true positive rate valued in all cases approximately* 0.7 *and supports the trainings performances in A.*

Table 4.1. ***Evaluation of deep learning models with different sizes of word embedding vectors.*** *`Keras` models with word embedding and subsequent feed forward layers were trained with chemical-biomolecule relations. Five word embedding model sizes were tested. Based on the relations in I_E (n=20 000) – equally balanced with INHIBITS and STIMULATES relationships – the performance measures of binary accuracy (Acc), precision (Prec), recall (Rec) and the F1-score (F1) were calculated. Model with $n = 100$ performed the best. The minimal loss in hyperparameter tuning (see chosen settings table S3-1) resulted in a slightly better performance considering test data I_E.*

Model	Target	Acc	Prec	Rec	F1
n=100	INHIBITS	0.70	0.76	0.64	0.71
	STIMULATES		0.64	0.72	0.68
n=500	INHIBITS	0.70	0.74	0.68	0.71
	STIMULATES		0.65	0.72	0.68
n=1000	INHIBITS	0.70	0.73	0.69	0.71
	STIMULATES		0.67	0.71	0.69
n = 2500	INHIBITS	0.69	0.72	0.68	0.70
	STIMULATES		0.66	0.70	0.68
n=5000	INHIBITS	0.69	0.74	0.68	0.71
	STIMULATES		0.65	0.71	0.68
Best model	INHIBITS	**0.70**	**0.76**	**0.68**	**0.71**
	STIMULATES		**0.64**	**0.72**	**0.68**

However, the model performance was neither improved during training, nor in evaluation with test data. The accuracy (0.685) for the unseen test data was smaller than for the models with higher learning rates (see table 4.1).

To further optimize the parameter setting, a **hyperparameter tuning** was performed by applying the `kerastuner`-function `RandomSearch()`. The overall model architecture remained identical for the case of word embedding vector size equal to 100. Five hundred sixty parameter recombinations varied in the activation functions, the number of dense layers, the severity of dropout and the degree of L2-regularization in the word embedding layer(see supplemental table S3-1).
All parameters might have a crucial influence on the model, but the number of recombinations was relatively high. Thus, only one hundred randomly selected parameter recombinations were examined. The recombination with minimal loss (see last column table S3-1) reached a binary cross-entropy of 0.5820 and a test accuracy of 0.70, which was a marginal improvement to the initial training. The performance measures were also improved considering test data I_E (see the last row of table 4.1 and supplemental figure S3-1).

Word embedding model with LSTM and dense layers. In text-based discovery approaches based on deep learning, recurrent neural networks have been applied frequently. As a potential improvement in model architecture, LSTM was added to the determined word embedding model. A recurrent neural network might improve the performance in the prediction task, and an additional LSTM layer was chosen.

The **model architecture** was very similar to the previous one, and the word embedding vector length was set to 100. A time distributed dense layer was added between word embedding and flattening an LSTM layer. The hidden output from the word embedding layer was given as input to another neuron of the LSTM layer. Thus, the hidden neuronal output of a word vector influenced the hidden output of the next word vector in the sequence in LSTM. The output of the LSTM layer was a sequence with equal length to its input and was the input for the following time-distributed dense layer. This layer transformed each sequence entity with the same fully-connected neural network, and the number of vector units was not reduced. The following flatten layer transformed the sequence of vectors with N units to a one-dimensional vector with $len(sequence) \times N$ units. The number of following dense layers varied in considered models with LSTM and depended on the length of the flattened vector. However, the settings for dense layers and the subsequent activation remained identical to the architecture described above.

Five **models with varying numbers of LSTM-neurons** ($N = [2, 10, 25, 50, 100]$) were

trained with I. The number of output neurons in the following time-distributed layer was also set to N. Dependent on the size, the respective models contained 2, 2, 3, 4 and 5 dense layers after flattening.

Independent of the number of LSTM units N, the trained models resulted in similar performance outcomes (see figure 4.6). The models trained for 86 to 152 epochs, whereby the number increased, the smaller N was chosen. All five models resulted in a final validation loss between 0.578 and 0.606 (see figure 4.6 A).

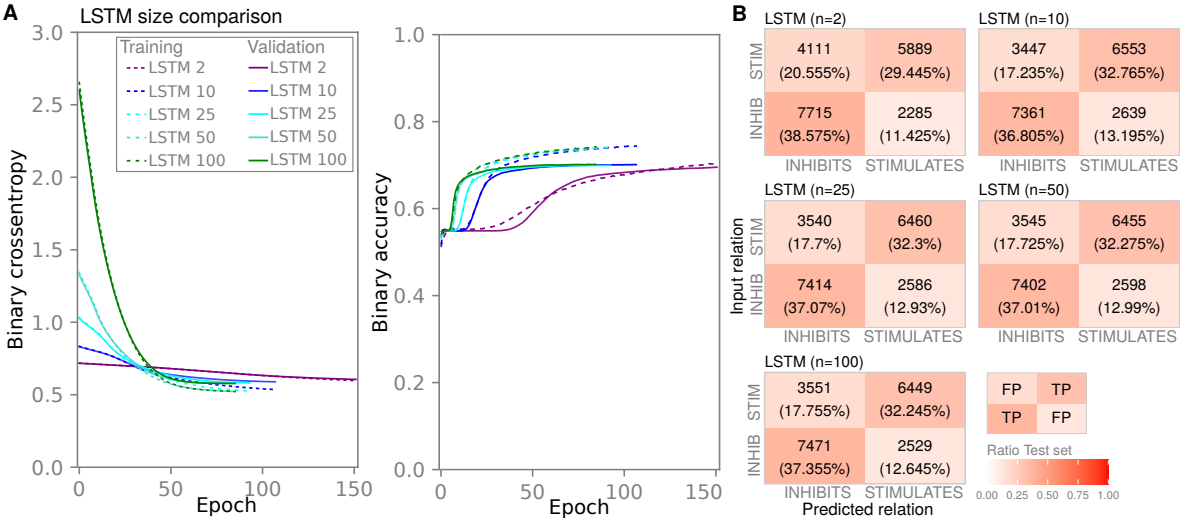

Figure 4.6. ***Comparison of the training performances for different number of neurons in LSTM layer.*** *Deep learning models were trained with UMLS-annotated chemical-biomolecule-pairs retrieved from SemMedDB to predict their toxicogenomic relationship. The sequential model consisted of an initial word embedding layer (len(vector) =* 100*), a long-short-term-memory (LSTM), a time-distributed dense neural network, a flatten layer and multiple dense layers, which decrease in size. The model was trained for a binary classification task. In the five models, the number of neurons for LSTM and the time-distributed dense layer varied (*2 *(purple),* 10 *(blue),*25*(turquoise),*50 *(lightgreen), and* 100 *(mossgreen) neurons). Early stopping occurred in training, when loss decreased less than* 0.01 *in a period of twenty epochs.* ***A:*** *The* `keras` *model performance across epochs was determined with validation set* I_{T1} *(n≈71T). The binary accuracy (right) and binary cross-entropy (left) were tracked during training. The loss converged around* 0.6 *for all models. All accuracies reached similar values of* 0.70. ***B:*** *Confusion matrices for prediction results to the test set* I_E *(n=20T) for all five trained models. The true-positive rate supports the trainings performances in A.*

The model evaluation with I_E (see figure 4.6 B and 4.2) also showed similar ranges of correctly

predicted relations across both relationship types. The predictions with *INHIBITS* were more accurate than with *STIMULATES*. The loss was slightly minimized compared to the best word embedding model (see figure S3-1), and the accuracy was nearly equal, ranging from 0.69 to 0.70. Thus, all five trained models resulted in similar performances as the best word embedding model.

To consider a more or less similar architecture with and without LSTM, we decided to choose $N = 100$. Consequently, the number of dense layers and the length of vectors after flattening became identical in both architectures.

Table 4.2. ***Evaluation of deep learning models with varying numbers of LSTM neurons.*** *A* `keras` *deep learning prediction model for chemical-biomolecule relationships was developed based on word embedding and long-short-term memory (LSTM) and subsequent feed forward layers. Various numbers of neurons in the LSTM layer and the following time-distributed dense layer were tested. Based on the relations in* I_E *(n=*20 000*), which were equally balanced with INHIBITS and STIMULATES relationships, the performance measures of binary accuracy (Acc), precision (Prec), recall (Rec) and the F1-score (F1) were calculated. The model with* $n = 100$ *performed the best. The minimal loss in hyperparameter tuning (see chosen settings table S3-2) resulted in slightly lower performances than initial setting considering test data* I_E*.*

Model	Target	Acc	Prec	Rec	F1
n=2	INHIBITS	0.68	0.77	0.65	0.71
	STIMULATES		0.59	0.72	0.65
n=10	INHIBITS	0.70	0.74	0.68	0.71
	STIMULATES		0.66	0.71	0.68
n=25	INHIBITS	0.69	0.74	0.68	0.71
	STIMULATES		0.65	0.71	0.68
n =50	INHIBITS	0.69	0.74	0.68	0.71
	STIMULATES		0.65	0.71	0.68
n=100	INHIBITS	0.70	0.75	0.68	0.71
	STIMULATES		0.64	0.72	0.68
Best model	INHIBITS	**0.69**	**0.74**	**0.68**	**0.71**
	STIMULATES		**0.64**	**0.71**	**0.68**

Again, a **hyperparameter tuning** was performed. Already tuned parameters of the word embedding layer and the dense layers were not changed, except the dropout. The dropout

parameters in LSTM and after the flattening were included in tuning but with identical values. Further parameters relevant to the LSTM layer were tested in a hyperparameter tuning approach (see supplemental table S3-2).

Like the word embedding layer, the activity regulariser was set to $L2 = 0.001$, whereas the kernel L2-regulariser was tuned with three different factors. The bidirectional LSTM-layer examined a sequence from both directions and thrown out its concatenated vector with $2 \times N$ units. Unrolling might increase the memory usage. However, it allowed considering an LSTM architecture without feedback-loops but learning from the sequence with *memory*. Very short sequences were considered, and thus, unrolling was applicable without exploding the memory usage. The last column in table S3-2 showed the recombination of parameters with the lowest loss in hyperparameter tuning.

A run for the selected model architecture was trained for 78 epochs and reached a final validation loss of 0.598 (see figure S3-2 A). The confusion matrix shows that 13 846 relations in I_E were correctly predicted (see supplemental figure S3-2 B), which values an accuracy of 0.69. Thus, the performance of the selected LSTM model seems not superior in performance in comparison to the selected word embedding model, when I_{T1} was considered as the test set and $\bigcup_{i=2}^{5} I_{Ti}$ as the training set.

4.2.3 Model comparison

This study trained deep-learning models with an initial word embedding layer with chemical-biomolecule pairs to predict the relationship type. The two selected model architectures — one without (A) and one with LSTM layer (B) — were compared in a 5-fold cross-validation approach to evaluate whether input and its split influenced model performance. The models were trained with the chosen architectures in a 5-fold cross-validation approach with the inputs I,I^V and I^H, respectively (see annotation models table 4.3). The six trained models were compared in their ability to predict known relations, which were not seen in training. The model with minimal validation binary cross-entropy across all folds was chosen for each combination of architecture and input. Per fold, the i-th set of $I_T^{(V|T)} = \bigcup_{i=1}^{5} I_{Ti}^{(V|T)}$ was considered as a validation set.

The performance curves in figure 4.7 A present the learning loss and accuracy for model architecture A regarding the validation data. In all three cases, the validation curves were very similar across folds, and none of them was visually distinctive to the other folds. Thus, in the case of models $A^\star$, A^V and A^H, the training-validation split did not influence the learning behaviour.

Table 4.3. ***Used symbols for model comparison.*** *Two model architectures were implemented and tuned in section 4.2.2 (A and B). The models were trained afterwards in a 5-fold cross-validation (see annotation rows 3 and 4) either with original SemMedDB input, the vertically augmented or the horizontally augmented version (see row Input, T: Training, V: Validation, E: Testing).*

	Original	vertical Augmentation	horizontal Augmentation
Input	$I_{T,V,E}$	$I^V_{T,V,E}$	$I^H_{T,V,E}$
A: WE+dense	$A^\star$	A^V	A^H
B: WE+LSTM+dense	$B^\star$	B^V	B^H

Comparing the differences in respect to input data, the trained models varied in their duration of training. Whereas I^H was ten times and I^V three times greater than I , the epochs was approximately 2 and 2.5 times less in the respective training. The loss curves for training with I^V_T and I^H_T converged rather quickly and the implemented early stopping forced to quit training after, on average, 42 and 35 epochs. For training with I_T, the mean duration was 91 epochs. For each training, the validation loss curves decreased steadily in the early training epochs and reached a plateau with minimal values of binary cross-entropy at 0.641 ± 0.003, 0.690 ± 0.003 and 0.701 ± 0.007. The mean maximum binary accuracy across folds valued 0.704 ± 0.002, 0.650 ± 0.003, and 0.666 ± 0.004. Considering both, loss and accuracy, the model $A^\star$ performed best concerning the validation results.

Figure 4.7 B presents the loss and accuracy curves for model architecture B regarding their performance in the validation data across epochs. Similar to model architecture A, the differences across folds for respective training with I_T, I^V_T or I^H_T resulted in marginal differences for loss and accuracy. The training durations were 95, 43, 32 epochs on average, and all validation loss curves decreased over time in similar behaviour as with model architecture A. Thus, all trained word embedding models made use of the implemented early stopping as training for 300 epochs would be exhausting when a plateau in the loss was already reached early. The validation loss was slightly decreased when using model architecture B. Again, especially in training with augmented data, the loss plateau was reached after a few epochs. Per input set, I_T, I^V_T and I^H_T, the differences in the minimal validation loss (0.628 ± 0.003, 0.674 ± 0.002, and 0.683 ± 0.004) were marginal across folds. The accuracy was the highest for training with I (0.701 ± 0.003). The augmented input resulted in slightly lower accuracy but was also similar across folds (vertically augmented: 0.646 ± 0.004; horizontally augmented:

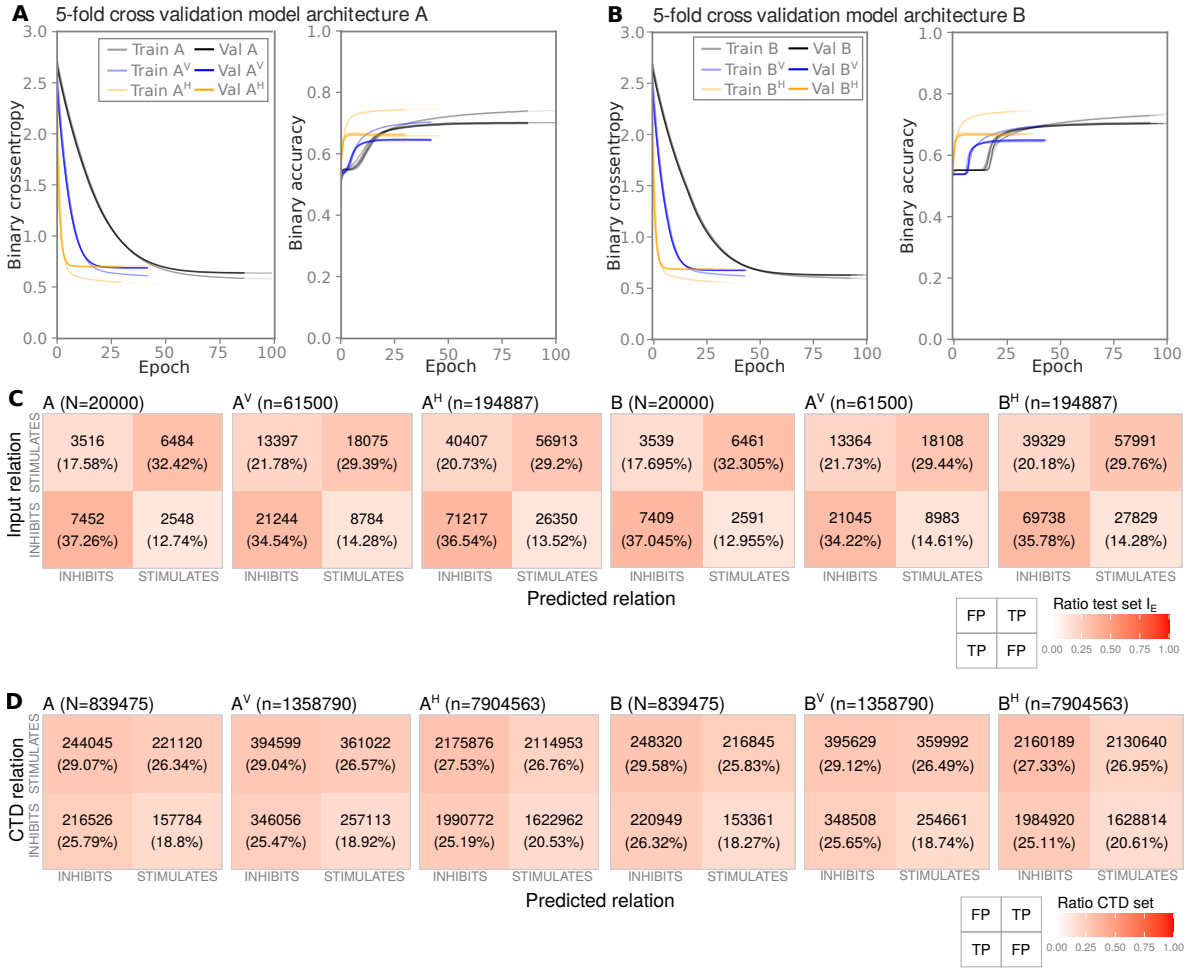

Figure 4.7. ***Model performance and validation.*** ***A)*** *Validation curves (dashed: Binary accuracy, solid: Binary cross-entropy) of 5-fold cross-validation of the word embedding model trained either with original input (black), vertically augmented (blue) or horizontally augmented (orange) versions.* ***B)*** *Validation curves (dashed: Binary accuracy, solid: Binary cross-entropy) of 5-fold cross-validation of the word embedding model with LSTM trained eiter with original input (black), vertically augmented (blue) or horizontally augmented (orange) versions.* ***C)*** *Confusion matrices of models with lowest loss for all six recombinations of model architecture and input data. Unseen chemical-biomolecule relations of input data (I_E, I_E^V or I_E^H) were used for evaluation of the model performance.* ***D)*** *Confusion matrices of models with lowest loss for all six recombinations of model architecture and input data. Knowledge from the comparative toxicogenomic database (CTD) was prepared as UMLS-annotated set of chemical-gene relations (T_{C2G}, T_{C2G}^V or T_{C2G}^H) were used as application to evaluate the model performance in toxicogenomic context.*

0.665 ± 0.003). Thus, the training without augmented input performed best considering the validation results for loss and accuracy.

Furthermore, the six selected models were evaluated with predictions for the respective test set I_E, I_E^V or I_E^H and the performance measures of binary accuracy, precision, recall, and F1-score. Per 5-fold cross-validation, the fold with the lowest binary cross-entropy was chosen for the comparison across model architectures. The relationship type was predicted for the samples in I_E, I_E^V or I_E^H (see figure 4.7 C). The true-positive rate ranged from 0.6366 to 0.6968 across all six models. Furthermore, the *INHIBITS* relationship was predicted more accurate in all six models, potentially as the ratio of training relations was slightly unbalanced towards relations with *INHIBITS*.

The observations of the performance curves already showed that the models trained with I outperformed those with augmented inputs, which was also the case in the evaluation with the in training unseen chemical-biomolecule relations (see table 4.4). Based on the validation and evaluation performances, the best performing model was $A^\star$.

Table 4.4. ***Test performances of best folds of 5-fold cross-validated models.*** *Based on the relations in the test sets (I_E, I_E^V and I_E^H) the performance measures were calculated. Next to binary accuracy (Acc), the other measures (precision (Prec), recall (Rec) and the F1-score (F1)) were considered per relationship type. The word embedding model $A^\star$ trained with I performed the best (bold text).*

Relationship	**A**	Acc	Prec	Rec	F1	**B**	Acc	Prec	Rec	F1
INHIBITS	$A^\star$	**0.70**	**0.75**	**0.68**	**0.71**	$B^\star$	0.69	0.74	0.68	0.71
STIMULATES			**0.65**	**0.72**	**0.68**			0.65	0.71	0.68
INHIBITS	A^V	0.64	0.71	0.61	0.66	B^V	0.64	0.70	0.61	0.65
STIMULATES			0.57	0.67	0.62			0.58	0.67	0.62
INHIBITS	A^H	0.66	0.73	0.64	0.68	B^H	0.66	0.71	0.64	0.67
STIMULATES			0.58	0.68	0.63			0.60	0.68	0.63

4.2.4 Toxicogenomic application

Additionally, the selected models were applied to predict relationships for the known toxicogenomic reference sets from CTD (T_{C2G}, T_G^V, T_G^H). The beforehand trained models were evaluated on how well they predicted toxicologically meaningful results (see table 4.5).

Figure 4.7 D presents the confusion matrices for the toxicogenomic application case. In

Table 4.5. ***SemMedDB trained models evaluated with chemical-gene interactions from CTD.*** *Based on the relationships in the CTD reference sets ($T_{G,E}$, $T^{V}_{G,E}$ and $T^{H}_{G,E}$), the performance measures of binary accuracy (Acc), precision (Prec), recall (Rec) and the F1-score (F1) were calculated. The word embedding model B trained with I performed the best (bold text), albeit the performance measures valued in similar ranges.*

Relationship	**A**	Acc	Prec	Rec	F1	**B**	Acc	Prec	Rec	F1
INHIBITS	$A^{\star}$	0.52	0.58	0.47	0.52	$B^{\star}$	**0 .52**	**0.59**	**0.47**	**0.52**
STIMULATES			0.48	0.58	0.52			**0.48**	**0.59**	**0.52**
INHIBITS	A^{V}	0.52	0.57	0.47	0.51	B^{V}	0.52	0.58	0.47	0.52
STIMULATES			0.48	0.58	0.53			0.48	0.59	0.53
INHIBITS	A^{H}	0.52	0.55	0.48	0.51	B^{H}	0.52	0.55	0.48	0.51
STIMULATES			0.49	0.57	0.53			0.50	0.57	0.53

contrast to the evaluation before, the ratio of correctly predicted knowledge was reduced clearly in all six cases of models — the accuracies were at 0.52. Thus, the models did not perform better than a prediction of relationships by random guesses. Consequently, the selected models could not reliably retrieve biologically meaningful results from an other toxicogenomic resource. Furthermore, the training with augmented input did not improve the accuracy. As neither sequence elongation nor increased in the number of samples improved the performance or coverage of biologically meaningful results, training with input data I might be more relevant for further investigations.

Comparing the performances for predicting toxicogenomic knowledge, $B^{\star}$ was marginally better in predicting toxicogenomic knowledge than $A^{\star}$ (see figure 4.5). On the contrary, the evaluation with I_E resulted in a better performance for $A^{\star}$ (see table 4.4). Regarding the somewhat similar evaluation results, we chose the more straightforward model architecture A for a chemical wise toxicogenomic application across different LOBOs.

Functional enrichment of predictions with genesets from CTD. The coverage of current knowledge could be investigated on the gene, pathway and disease level (see figure S3-4). Thus, exposure effects were considered on three different levels of biological organisation (molecules, cells and individuals). When considering a prediction model that reliably retrieves biologically meaningful toxicogenomic chemical-biomolecule relationships, the functional enrichment approach would benefit hypothesis generation in the AOP development. However, this was not the case with an accuracy of 0.52 for applied CTD data, even for the best per-

forming model. The enrichment workflow and results are shown in supplemental section S3.2 to help understand the conceptual framework but not generate a toxicogenomic hypothesis in this study's scope.

4.2.5 Horizontal augmentation without tail-padding

In the previous comparison, one possible adaption of the input was its elongation with parental and grandparental semantic concepts to a length of six instead of two. However, not every subject and object concept had a parental term. Therefore, sequences were tail-padded with zero-masked tokens to prevent their shortening. Thus, all sequences had an equal length of six. The position of words might also imply a specific semantic meaning, but the tail-padding destroyed the order of the input sequence —an essential anchor in the machine learning training process. Therefore, training with horizontally augmented input was performed with zero-masked tokens but without tail-padding conserving the semantically identical order across all samples and, thus, investigating the influence of the sequence order.

The input preparation remained identical as before for horizontally augmented inputs, except that sequences were not tail-padded where parental or grandparental terms were missing. Consequently, when samples were integer encoded, the zero-values were not at the end but on its semantically specified position (<Subject, Subject Parent, Subject Grandparent, Object, Object Parent, Object Grandparent>). A 5-fold cross-validation training was performed with model architecture A and B, respectively. The supplemental figure S3-6 shows the training curves with both model architectures and the confusion matrices considering the unseen test data set.

Similar to the previous case (see figure 4.7 A and B (yellow graphs)), the training duration was relatively short, with 29 to 35 epochs. The five folds had somewhat similar training and validation curves, and thus splitting of input data seemed not to affect the training performance. In addition, the validation loss decreased rather quickly within the first five epochs and converged to a value at minimally 0.69. The binary accuracy curve for validation data increased within the first five epochs and converged to 0.65.
The evaluation with unseen test data resulted in true-positive rates of 65% for both model architectures. In addition, the values of relationship type-specific performance measures were similar across models (see table S3-5) compared to the previous experiments with tail-padded sequences (see table 4.4 A^H and B^H). Consequently, preserving the semantically meaningful order of the horizontally augmented input did not significantly affect the model performance.

4.2.6 Four-class problem formulation

The previous examinations employed the input data set I with relationship types *STIMULATES* or *INHIBITS*. The following presents the model training with an extended input I_4 considering four relationship types. In total. twenty-five predicates were considerable in the partly preprocessed SemMedDB input set of chemical-biomolecule relations (see supplemental table S3-6).

Next to the relationship types mentioned above, *AUGMENTS* and *DISRUPTS* were included. According to Kilicoglu et al. [2012], these relationship types were related to pharmacogenomics and, thus, might also be suitable to represent toxicogenomic chemical-gene interactions. Consequently, it was investigated whether an increase with additional predications improved the model performance. The input I_4 consisted of 632 864 unique chemical-biomolecule pairs, which all were assigned to one predicate only. After removing contradictions, the predicate ratios were 38.5% for *INHIBITS*, 27.7% for *STIMULATES*, 16.1% for *AUGMENTS*, and 17.7% for *DISRUPTS* * (see total numbers in supplemental table S3-6). A tenth of I_4 was considered as evaluating test set. The remaining was split into five equally sized subsets for 5-fold cross-validation training. In all subsets, the ratios of predicates remained identical to the overall set. Two types of learning tasks were applied with the input data containing four relationship types.

In the first experiment, the prediction task was considered as a categorical classification task considering four predicate classes. The selected model architectures A and B were applied identical, except that the last dense layer contained four neurons and a softmax activation layer.

In a second experiment, the relationship types were assigned to *NEGATIVE* regulations (*INHIBITS* and *DISRUPTS*) and *POSITIVE* regulations (*STIMULATES* and *AUGMENTS*). Again, an input with two contradicting predicates was employed in a binary classification task, as it was initially implemented with the selected model architectures.
The results of respective 5-fold cross-validation are shown in supplemental figure S3-7. In summary, the expanded input I_4 improved the overall performance neither in a categorical classification nor a binary classification.

Independent of the classification task and the model architecture, the five folds of a cross-

* The ratios of predicates for the previous consideration were 53.7% *INHIBITS* to 46.3% *STIMULATES* and were considered as mildly unbalanced. Adding two further predicates increased the imbalance, but was still expected as acceptable to test, whether the expansion of relationships might help improve training and evaluation performance of the model.

validation training behaved similarly and were visually not distinct (see supplemental figure S3-7 A-D). The validation loss decreased relatively fast in the first twenty epochs and converged to approximately 0.67 after 60 to 120 epochs. The validation accuracies reached maximal values of 0.68.
True-positive rates of 0.683, 0.664, 0.683 and 0.680 were measured for the categorical case (with model architecture A and B) and binary case (A and B) considering the test set $I_{4,E}$ with 63 287 relations (see supplemental figure S3-7). Consequently, the previous experiments resulted in marginally better performances when considering input I and two predicates.
However, the categorical consideration revealed an impressive characteristic in confusion matrices. Although overall classification performed nearly identical, the model could distinguish pharmacogenomic relationship types (*AUGMENTS* and *DISRUPTS*) from substance interactions (*STIMULATES* and *INHIBITS*), with a true-positive rate of 99.4% with both model architectures. Thus, the models recognised the general semantic meanings of the relationship types and respective suitable chemical and biomolecular semantic concepts. The applied binary classification task presented that the models could distinguish negative regulations from positive relations by 0.68% accuracy. Albeit not that accurate as for the interaction types, still the model was able to predict positive or negative directions of chemical-biomolecule interactions.

Most likely, the input data itself allowed the very accurate relationship discrimination of pharmacogenomic activities and substance interactions. Whereas *subject* concepts overlapped by 30% across predicate types, *STIMULATES* and *INHIBITS* shared less than 2% of *object* concepts with *DISRUPTS* and *AUGMENTS* (see supplemental figure S3-8) but 37.78% and 13.12% within the predicate groups. Consequently, the models learned to differ between relationship types potentially through the *object* concept.

4.2.7 Training with CTD data

The trained models differed marginally in their performances when considering the SemMedDB input data. Furthermore, the models were accurate to maximally 70% for unseen SemMedDB relations. In the case of input I, neither vertical nor horizontal augmentation improved the model performances. The chemical-biomolecule interactions from CTD were used as training input to examine how the model architectures and data augmentation might affect another input data set.

The data set T_{C2G} was prepared similarly to I for training in 5-fold cross-validation. T_{C2G} was also vertically (T^V_{C2G}) and horizontally augmented (T^H_{C2G}) (similar to augmentation of I described in section 2.2.2). A test set for evaluation was split off from all three inputs. The

remainder was taken into account for training and validation and was divided into five equal subsets $\bigcup_{i=1}^{5} T_{C2G,i}$.

For all three kinds of inputs, a 5-fold cross-validated training was performed with model architecture A and B. Per 5-fold cross-validation, the performance curves of loss (binary cross-entropy) and binary accuracy did not differ visually (see figure 4.8). Thus, the five folds of training performed equal and had similar training durations in all six cases. Consequently, the training-validation split did not affect the performance of models trained with CTD input.

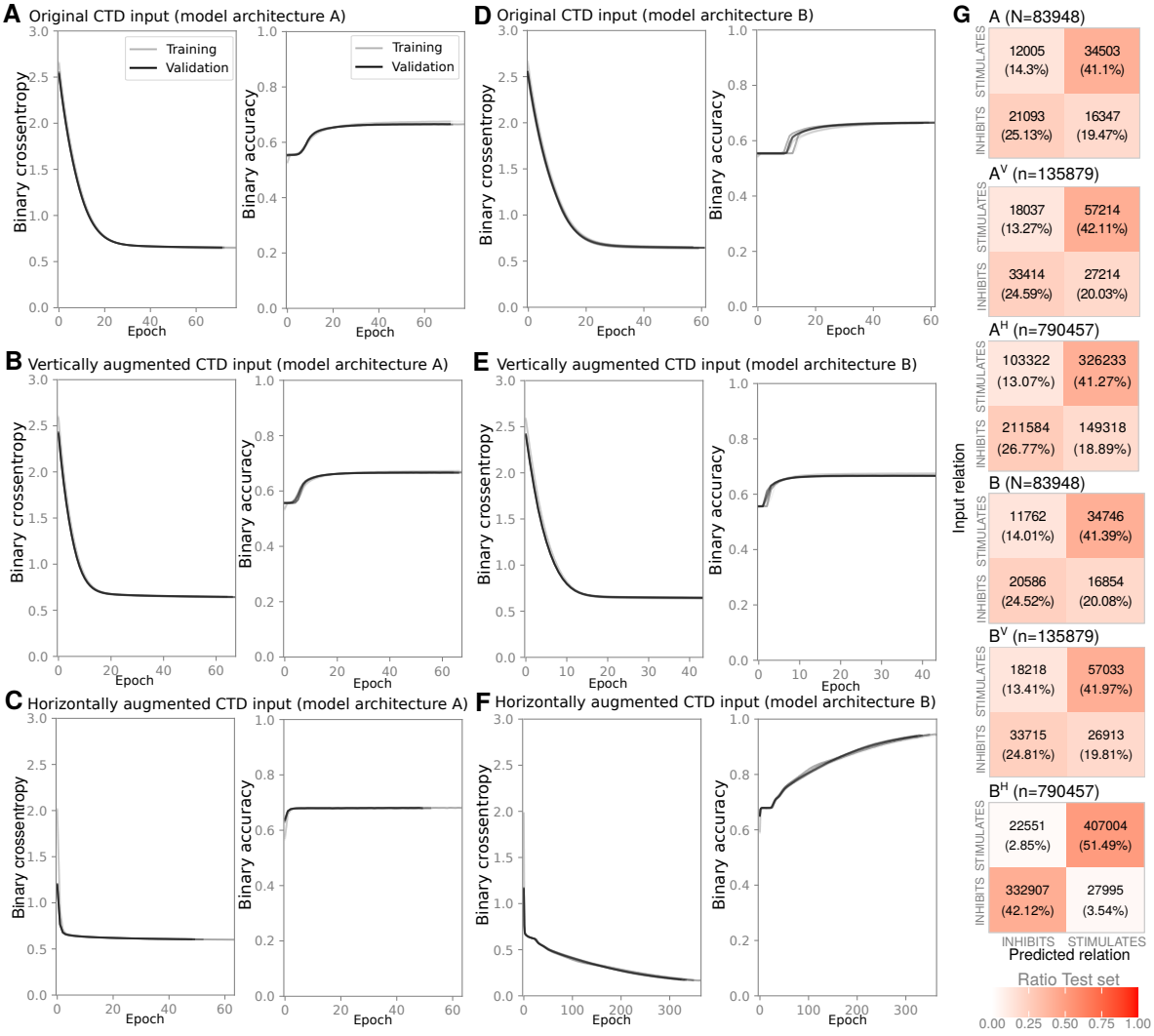

Figure 4.8. ***Performance of models trained with CTD input.***

The training with T_{C2G} and T_{C2G}^{V} had similar validation curves and evaluation performances as the models trained with SemMedDB data. The loss curves decreased relatively fast within twenty epochs and converged to a value of approximately 0.65. The binary accuracy increased within the first twenty epochs and converged at values of around 0.665.

For trained models with original and vertically augmented inputs (see table 4.6 and figure 4.8 G), the test accuracies ranged between 0.66 and 0.67. However, the F1-scores were higher for predicting *STIMULATES* with CTD input and, on the contrary, higher for *INHIBITS* with SemMedDB input (see table 4.4). The most likely reason was the difference in the predicate balance in the two kinds of input. Albeit detecting only slight imbalances between *INHIBITS* and *STIMULATES* for both inputs, *STIMULATES* was overrepresented in T_{C2G} (I: 3 671 698, S: 4 413 646) but underrepresented in I (I: 204 219, S: 169 685), which probably resulted in converse prediction measures per relationship type. Furthermore, due to the larger imbalance in T_{C2G}, the difference in the F1-score between relationship types was higher in training with CTD input.

Table 4.6. ***Comparison of selected model architectures with different training inputs from CTD.*** *Based on the relationships in the CTD reference sets ($T_{C2G,E}$, $T^{V}_{C2G,E}$ and $T^{H}_{C2G,E}$), the performance measures of binary accuracy (Acc), precision (Prec), recall (Rec) and the F1-score (F1) were calculated for models that were trained with CTD input T_{C2G}, T^{V}_{C2G}, T^{H}_{C2G}. The word embedding model B^{H} performed much better (bold text) than remaining models.*

Relationship	**A**	Acc	Prec	Rec	F1	**B**	Acc	Prec	Rec	F1
INHIBITS	$A^{\star}_{CTD}$	0.66	0.56	0.64	0.60	$B^{\star}_{CTD}$	0.66	0.55	0.64	0.59
STIMULATES			0.74	0.68	0.71			0.75	0.67	0.71
INHIBITS	A^{V}_{CTD}	0.67	0.55	0.65	0.60	B^{V}_{CTD}	0.67	0.56	0.65	0.60
STIMULATES			0.76	0.68	0.72			0.76	0.68	0.72
INHIBITS	A^{H}_{CTD}	0.68	0.59	0.67	0.63	B^{H}_{CTD}	**0.94**	**0 .92**	**0.94**	**0.93**
STIMULATES			0.76	0.69	0.72			**0.94**	**0.94**	**0.94**

However, the most crucial difference was the outcome of the training with model architecture B and horizontally augmented CTD input T^{H}_{C2G} (see table 4.8 F). The loss curve dropped relatively fast in the first ten epochs and slowly decreased over more than 300 epochs from binary cross-entropy values of 0.65 down to 0.17. Thus, the validation loss differed clearly from all other validation curves considering SemMedDB and CTD input and reached lower loss values. Consequently, the horizontal augmentation and the use of an LSTM layer helped minimize the loss. Thus, the binary accuracies of validation data were much higher than those of other trained models.

For the unseen chemical-gene relations in $T^{H}_{C2G,E}$, the true-positive rate was valued at 93.61% (see figure 4.8 G). The performance measures of model B^{V}_{CTD} outperformed all others when

evaluating with unseen CTD data (see table 4.5 and compared to table 4.4). No apparent difference in performance measures was observable between the two relationship types in the horizontally augmented case. In consequence, horizontal augmentation and applying a deep learning model with LSTM layer allowed to train a highly accurate model which also convinced through high specificity (precision values) and high sensitivity (recall values).

An evaluation of the CTD trained model B^H_{CTD} was also performed with SemMedDB input I^H. Similar to the toxicogenomic application, the evaluation crossing inputs for relational data resulted in relatively small performance values (see table 4.7). With an accuracy of 0.53, again, the model performance was slightly better than random guessing.

The preprocessed inputs I and T_{C2G} (and their augmented versions) considered chemical-biomolecule interactions, but noticeable differences had to be pointed out. As mentioned above, both data were slightly imbalanced but in converse directions for the two considered predicates leading potentially to bad performances with data from other resources. However, the primary difference between both inputs was the set of biomolecules. Whereas SemMedDB input contains all types of biomolecules, CTD was based only on (UMLS-transformed) gene names. Thus, the context was more specified to one type of biomolecule. It was very likely that such differences provoked poor performance for relation samples from other references.

Table 4.7. ***Application case of SemMedDB data to CTD trained model with horizontal augmentation.*** *The selected model architecture B was applied to train with horizontally augmented input T^H_{C2G} in a 5-fold-cross-validation. With UMLS-annotated chemical-biomolecule relations from an independent data set I^H, the performance was measured calculating accuracy, precision, recall and F1-score.*

	Relationship	Acc	Prec	Rec	F1
B^H_{CTD}	INHIBITS	**0.53**	**0.50**	**0.58**	**0.54**
	STIMULATES		**0.57**	**0.49**	**0.52**

4.3 Discussion

This study employed a relevant data set covering literature knowledge of exposure-related biomolecular interactions from SemMedDB. We applied this data set for the deep learning model prediction of chemical-biomolecule relationships. The best performing model resulted in a 70% test accuracy. However, the trained model predicted the current knowledge badly in a toxicogenomic application. We determined lower accuracy (52%) when predicting chemical-gene interactions from CTD in this SemMedDB trained model — marginally better than a random guess.
Additionally, we employed chemical-gene interactions from CTD to train similar deep learning models. In such a case, the advantages of a model architecture with the LSTM layer and horizontal input augmentation were exploited. This CTD trained model reached 94% accuracy. However, it was not applicable to predict SemMedDB relations accurately (54%).

Consequently, the generated deep learning models were not consistent beyond resources used for training. Thus, the approach was not yet ready for toxicological knowledge discovery regarding the link between chemical exposure and molecular biological effects. Still, the CTD application showed that artificial intelligence models based on knowledge representations predicted accurately within a resource's scope. The study's achievements might be a start towards a hypothesis generation tool for AOP development. We presented conceptual ideas and new data integration strategies for toxicological knowledge.

4.3.1 Transferring biomedical knowledge towards toxicology.

Callahan et al. [2017] highlighted different tasks within knowledge-based biomedical data science. For example, biomedical researchers generated knowledge representation [e.g. Hristovski et al. 2015, Cong et al. 2019] and have applied such to predict relations with the help of embedding models and NLP applications [e.g. Bakal et al. 2018, Kastrin et al. 2018].
This study considered a toxicology-related subset of the NLP-based knowledge representation of SemMedDB in a deep learning prediction model initialised with a word embedding layer. The generated input data set I was the reduced subset of SemMedDB, which represents UMLS-annotated chemical-biomolecule interactions, and thus, a toxicological knowledge representation on a molecular level. The literature-based discovery tasks of *named entity recognition* and *relation extraction* were performed already in the applied input. In respect to the hierarchical LBD workflow, according to [Zhao et al. 2021] (see figure 1.6), SemMedDB would allow aiming for hypothesis generation tasks or pathway generation tasks directly.

Interoperability of data through UMLS. The SemMedDB predications had a UMLS-annotated triplet structure. Subjects and objects were represented as semantic concepts, whereas the predicates were based partly on the relationship tree from the UMLS semantic network. The reduction to chemical-biomolecule relations led to a selection of the SemMedDB subset of interest. Also, other researchers reduced in an expert-driven manner SemMedDB [e.g. Rastegar-Mojarad et al. 2016, Bakal et al. 2018]. The conceptual novelty of this study was the grouping of semantic concepts in different LOBOs. Consequently, the adaption of the UMLS-standardised annotation made the biomedical knowledge available for a systems biological perspective and potential applications in AOP development. Toxicologists could easily apply the annotation of lexical names to UMLS concepts and thus to a LOBO group with the given UMLS resources. Furthermore, the available assignment of LOBOs to UMLS semantic types were contextual and semantic groupings of concepts, which helped prepare a FAIR (findable, accessible, interoperable and reproducible) input.

Semantic concepts with multiple assignments. The toxicological scope was relatively broad considering the variety of biomolecules, e.g. genes, enzymes, protein complexes, or metabolites. Moreover, some external chemicals considered in SemMedDB, such as hormones for birth control, could also be biologically synthesised in cellular systems. In the UMLS Metathesaurus, concepts with multiple assignments to semantic types were present. Consequently, some semantic concepts in I were chemicals and biomolecules as well. For example, the semantic concept 'estradiol' (C0014912) belonged to semantic types of 'organic chemical' (T109), 'pharmacological substance' (T121) and 'hormone' (T125). However, the first two belonged to the chemical LOBO, and the latter belonged to the biomolecular LOBO. Thus, 'estradiol' might occur in I as a chemical subject as well as a biomolecular object. Such concepts could limit learning semantic representations in word embedding, especially, when considering directed relations. For example, chemical concepts had only chemical parental terms, and biomolecular concepts had only biomolecular parental terms in data augmentation to reduce the negative effects of multiple concept assignments within this study. Moreover, the augmented data were free from indirect duplicated relations *.
In summary, we must be cautious when interpreting the biological meaning as semantic concepts might have multiple semantic types and multiple LOBOs.

Contradictions in SemMedDB. The triplets in SemMedDB had a plethora of predicates. In the context of chemical-biomolecule relationships, the number of relations per predicate was highly unbalanced (see table S3-6). The chosen predicates should also comprise the in-

* A subject-object relation occurred again, when considering the subject parental term or object parental term.

teractions between an external chemical and a biomolecule. Therefore, a balanced subset was selected to train prediction models in a toxicological context with frequent predicates *STIMULATES* and *INHIBITS*. Kilicoglu et al. [2012] mentioned that the chosen predicates were related to substance interactions in SemRep. Furthermore, the contradictory semantic meaning was beneficial for the predictive classification task and thus biological interpretation of the predicates.
Cong et al. [2019] already described and discussed the occurrence of contradictory relations in SemMedDB. The database projected the nearly complete current PubMed knowledge across, e.g. different species, study designs, scientific domains, types of exposures and types of adverse outcomes. Thus, the contradictions were induced, in parts, through the comprehensiveness of SemMedDB. Contradictions would negatively influence the training performance, especially when considering a binary classification task. There were some chemical-biomolecule relations in I, which occurred with both relationship types. Therefore, we removed the ambiguities of information to gain a model, which could perform the learning task.

The **hierarchical structure of semantic concepts in the UMLS Metathesaurus** was helpful to augment the number of samples and the length of relational sequences in the selected set of chemical-biomolecule relations from SemMedDB. In the UMLS, broader semantic concepts were superior to narrower ones and considered as parental terms. However, the different resources in the UMLS might comprise multiple broader concepts to a narrower one. For example, the resource MESH might associate multiple broader concepts with one narrower concept due to its mesh-like hierarchical architecture. For example, Diclofenac — a non-steroidal anti-inflammatory agent — was represented by the semantic concept of *Diclofenac* (C0012091) in UMLS. The respective parental term *Analgesics, Anti-Inflammatory* (C0002773) was superior, but twelve further parental terms were given in the Metathesaurus, and four were based on knowledge retrieved from MESH.
The horizontal augmentation resulted in sequences of different lengths and sample-wise information richness. As the number of parents (and grandparents) was partly greater than one, one original subject-object pair might result in multiple horizontally augmented subject-object relations. Thus, when considering UMLS knowledge, the horizontal augmentation also induced a vertical augmentation.
The ambiguities in the UMLS mentioned above affects both kinds of data augmentation. The semantic concepts have either none, one or multiple parental terms. Thus, the vertical augmentation did not result in equal amounts of expanded samples across all original relations. As a result, the input data had potentially an increased study bias induced by UMLS. Thus, rarely occurring and not so well-studied semantic concepts might become underrepresented in

augmented data. However, rare semantic concepts which shared a parental term with many concepts might become easier associated with a relationship type. For example, if subject semantic concepts with an identical parent had the same object semantic concept, a rarely represented subject might become easier predicted regarding the relation similarities of the subject terms with the same parental term.

In this study, we augmented the input data with the hierarchical knowledge of the UMLS Metathesaurus. The knowledge representation was expected to become denser when adding further semantic, not-causal knowledge. However, the input became noisier as the various UMLS sources had different structures, various depths and were partly ambiguous when not considered in their original scientific context.

Zhang et al. [2019a] highlighted the importance of curation and preprocessing noisy knowledge base data. Dependent on the research field, it could be suitable to reduce the extensions to domain-specific resources and, thus, contextualise UMLS or SemMedDB differently. The combined use of multiple biomedical knowledge bases in the UMLS resources generated reliable biomedical results, e.g. retrieving accurate medication information [Bejan and Denny 2014, Wei et al. 2013]. Within the present study, the preprocessing and curation of SemMedDB data allowed generating a prediction model for chemical-biomolecule relation with an accuracy of 70%. However, the model was not applicable to reliably predict toxicogenomic knowledge stored in another database. Furthermore, the data augmentation did not improve predictability when transferring biomedical knowledge towards toxicogenomic knowledge given the considered data.

To circumvent the shown limitation of SemMedDB, a toxicology-specific research question might reduce considerable input data. In recent AOP development, text mining approaches were successfully applied [e.g. Rugard et al. 2020, Jornod et al. 2020], focusing on specific chemical exposures or adverse outcomes. However, these approaches also relied on manual and human interpretable curation and were applied without any deep learning approach. The here presented work aimed to train a deep learning model, which learns in an automated way of a broader current knowledge in biomedicine and toxicology. This study was based on an entirely data-driven approach which could be meaningful for such AOP-relevant text mining tasks in the future.

In the present study, we removed the contradictory relations in SemMedDB entirely. A focus on chemical-biomolecule relations helped contemplate toxicological instead of biomedicine knowledge. The data augmentation increased number of samples and the length of sequence samples in I, which might increase the information richness by reflecting the hierarchical

and not-causal relations for chemicals and biomolecules. A pre-choice of domain-specific hierarchies from the UMLS could improve the predictability of toxicological relations. For example, compound-specific relations could be considered for data augmentation, e.g. the UMLS resource *drugbank*. This resource selection would set more focus on pharmacology-related semantic hierarchies.

4.3.2 Deep learning with biomedical knowledge representation

A straightforward model for training with SemMedDB.

There were various possibilities to apply a **knowledge representation**. For example, the representation learning with the *TransE*-model or its derivatives retrieved knowledge from multi-relational databases, also for chemical-gene interactions as shown by Choi and Lee [2019]. Furthermore, text-based knowledge [e.g. Rotmensch et al. 2017, Zhang et al. 2019b, Lee et al. 2019] was considered and was mined and contextualised with the help of vector representations catching the semantic meaning of a text corpus. We chose a straightforward, but state-of-the-art deep learning architecture. The finally selected word embedding model architecture A trained with input data I reached a maximal accuracy of 70%. Our deep learning model architecture with a word embedding layer captured the semantic meaning of chemical-biomolecule relations from the text-based and biomedical-centred data set of SemMedDB.
The comparison of different embedding vector sizes N led to more or less identical test performances. Expanding N was connected to a larger number of fully-connected layers. Moreover, the size of fully-connected layers was decreased by factors in the range of 2 to 2.5. Consequently, greater values of N induced a larger number of word embedding weight parameters and more weight parameters in fully-connected layers. The maximal word embedding size had a slightly better minimal loss (see figure 4.5). However, increasing the model complexity decreased the test accuracy slightly. In this investigation, we chose an initial word embedding layer with the smallest number of tunable parameters leading to $n = 100$. The word embedding itself and following layers had already many parameters. Considering representations in larger vectors seemed irrelevant when smaller ones covered the semantic information equally.

A recurrent neural network, like **LSTM**, allows to learn from sequential data and considers the knowledge of the previous items in a sequence. In this investigation, the recurrent neural network architecture of LSTM did not improve the performance of the SemMedDB prediction model. Initially, there were some reasons to expect an improved performance — also for short sequences. By applying LSTM, a recurrent neural network layer was added to the model, which was not considered in the simpler model architecture A. Thus, an improvement was

not directly expected due to the ability to overcome the vanishing gradient problem but by learning from sequential data of a recurrent neural network in general. Furthermore, with adding an LSTM layer, also a time-distributed dense layer was added. The LSTM output was a sequence, as we considered a many-to-many LSTM architecture. The time-distributed layer applied a fully-connected network with the exact same weights on each vector of the sequence, which was somewhat a normalisation across the LSTM output sequence. Thus, another layer with additional parameters influenced the training performance. Expanding the sequential model architecture with a recurrent neural network and a time-distributed network increased the model complexity. Potentially, it could help recognise learning task-specific patterns. However, the model architecture B with LSTM did not improve the validation and test performance considering data of SemMedDB.

In the cases with **horizontally augmented input**, longer sequences were considered. It was expected that this might influence the model performance in at least the architecture with LSTM. Although not every semantic concept had a parental term, considering equally sized sequences was possible by masking empty spots in a sequence with a zero value. The zero-masking and padding to the same length can be performed differently. One possibility was to keep the sequence order * with zero values in the centre of the sequence and, thus, potentially had longer distances between meaningful entities. Another possibility was tail-padding which re-arranged zero masked words to the tail of the sequence to overcome this limitation. Assuming that each position was associated with a specific semantic role in the horizontally augmented sequences, the training of a model might be more challenged with tail-padded sequences, e.g. when sometimes a subject parent and sometimes an object was on position two during training. In the presented results, both input versions were examined as horizontally augmented data. Considering the respective evaluation data, the performance of the trained model A^H and $A^H_{ordered}$ were slightly weaker than for $A^\star$, which might be associated with the more complex input due to sequence elongation.
For the model architecture B with LSTM layer, B^H or $B^H_{ordered}$ showed worse test performances in comparison to $B^\star$. Thus, considering a more complex LSTM architecture with horizontally augmented sequences was not beneficial.

In the present investigation, a word embedding model with subsequent feed-forward neural networks performed the best when considering SemMedDB input. Although different architectures, tunings and input settings were considered, the test accuracies were limited at a maximum of 70%. This study was the first of its kind, which developed a deep learning model with knowledge representation to predict chemical-biomolecule interactions based

*$< Subject, Subject\ parent, Subject\ grandparent, Object, Object\ parent, Object\ grandparent >$

on relationships, which were retrieved from texts. Two different model architectures were considered, but a plethora of others and even more advanced deep learning approaches for knowledge representation might be applicable. The already investigated potentials of convolutional neural networks [e.g. Alawad et al. 2019, Peng et al. 2018, Zhao et al. 2019] or a concatenation of multiple model architectures [Zhao et al. 2019] or the use of model ensembles [e.g. Peng et al. 2018] might be essential alternatives when investigating biomedical knowledge representations. The list of potential alternatives of deep learning models had not been fully elaborated. However, a further discussion would exhaust the scope of this thesis.

LSTM model for training with horizontally augmented CTD input

The implemented deep learning workflow predicted the relationship type of a relational sequence. In consequence, it was not crucial whether the input originates from text-based or curated knowledge. The chemical-gene interactions from CTD were also applicable as input. It resulted in similar test performances as determined for SemMedDB when considering not augmented input data. As a result, models $A^{\star}_{CTD}$ and $B^{\star}_{CTD}$ were trained with originally formatted input and were 66% accurate for unseen CTD relations.
However, the most promising outcome was determined when training a model architecture with LSTM and horizontally augmented input. A significant increase in accuracy, precision, and recall (for both relationship types) was detected. Thus, for not literature-based but empirically measured and human-curated data: (i) A very accurate model (94%) was determined. (ii) Input augmentation and data integration have been shown beneficial. (iii) The use of a recurrent neural network for longer sequences became highlighted.

The CTD trained model considered only subjects and objects represented in the selected SemMedDB subset *. One observation was the relatively small ratio of overlapping subject and object concepts. We compared the recombinations with subject and object concepts occurring in CTD and SemMedDB. Most semantic concepts in the SemMedDB were missing in the CTD training input (Subjects: 18393 out of 25761, Objects: 13182 out of 19189). We calculated the information density as sparsity of the subject-object-occurrence matrices for both not-augmented inputs. Albeit both data sets covered only a small ratio of the possible recombinations of chemical-biomolecule relations, the density for CTD input (2.19%) was

* The integer encoding and word embedding matrices were adapted to fit every SemMedDB semantic concept in the CTD model. The trained model applied initialised word vectors without training adaption for samples with semantic concepts not represented in CTD. Thus, these word embedding vectors did not learn any semantic representation, and a semantically meaningful prediction is not likely. Consequently, the unification with UMLS annotation helps only in parts and trained models are not applicable for data of other toxicological knowledge bases when the overlap of semantic concepts is small.

approximately ten times greater than for SemMedDB input (0.275%) *. Still, the models trained with CTD performed best with horizontal augmented input and a model architecture with LSTM. The results led to the recommendation to horizontally augment the input for model training and apply the word embedding model architecture with LSTM. It should be clarified why augmentation and LSTM helped for training with CTD input but not for SemMedDB. The greater information density of CTD input might be one potential reason. Input data have to overcome a specific sparsity value in their subject-object occurrence matrix to benefit from data augmentation with the UMLS knowledge.
As the subject-occurrence matrix in CTD was denser than for SemMedDB, more relations were available per concept. Thus, the chance of co-correlated genes in CTD was greater than for SemMedDB. Thus, we expected a better performance to some extent. If one chemical-gene interaction was removed from training data and used for test data, closely correlated genes were still available for one specific chemical compound during training. Thus, the test relation was more likely to be predicted correctly. However, the information density for relationships separately was not considered yet. In consequence, we must also investigate whether the ratio of positive and negative correlations might influence the training outcome.
In the future, we will prove the effectiveness of a greater information density on SemMedDB. Therefore, we will add not contained subject-object-relations considering sibling information † from the UMLS file `MRREL.RRF` (see section 2.2.1). The expanded input will be used in a similar model comparison as shown with I and T_{C2G}. If the training of a model with LSTM layer will perform better with horizontally augmented input, we will generate proof that a minimal information density would be needed to gain from the strengths of horizontal augmentation and recurrent neural networks.

4.3.3 Data integration

Integrating data from different sources allows generating a knowledge graph through joining, merging or concatenating. Such knowledge-based data integrations were used in computational toxicology, as shown by a various application cases [e.g. Mower et al. 2018, Pittman et al. 2018, Nair et al. 2020, Taboureau et al. 2020, Aguayo-Orozco et al. 2019]. For example, the CTD was frequently considered for data integration with exposure-associated effects [e.g. Gu et al. 2019, Davis et al. 2018, Oki and Edwards 2016]. The present study applied CTD also as a toxicogenomic reference. Chemicals and genes in CTD were annotated to the UMLS ter-

* The information density was also ten times greater, when considering also concepts not represented in the other database (SemMedDB: 0.076%, CTD: 0.785%)

† `MRREL.RRF` contains more than 900 predicates for two UMLS-concepts, some of them present sibling relations of concepts, e.g. SAME_AS, ALIAS_OF or COMMON_NAME_OF.

minology. By reducing relationship types in the CTD database to the considered ones from SemMedDB, it was possible to retrieve similarly structured chemical-biomolecule-triplets, which were toxicologically relevant and had empirical support. Consequently, knowledge from CTD was helpful as training input itself, as mentioned before, and applicable as an evaluation data set for the SemMedDB trained prediction models.

The CTD also contained **chemical interactions on different LOBOs**. Gene sets of biological pathways included knowledge on the cell or tissue level, whereas disease gene sets were relevant on the organism level. The expansion of chemical-biomolecule relations to sets of chemical-pathway and chemical-disease relations allowed to check the overall coverage of exposure-related toxicological knowledge within the chosen subset of SemMedDB. The coverage was relatively small on the molecular level, which was expected as CTD considered genes and not all types of biomolecules as opposed to text that include, e.g., enzymes, proteins and metabolites. Furthermore, the binarisation of relations in preparation might affect the resulting CTD set T_{C2G} through additional ambiguous relations or information loss. Thus, we already expected a limited coverage of the chemical-biomolecule interactions. However, a higher exposure-related toxicological coverage was determined on the pathway level. Moreover, the absolute number of overlapping relationships on the pathway and disease level showed a much larger subset of SemMedDB information (see figure 4.4). Thus, the chosen input data were representative of at least some toxicogenomic knowledge. In consequence, the data integration of SemMedDB knowledge with CTD input might be relevant for future applications aiming to predict toxicological interactions on higher biological levels.
The reader and potential applicant should know that merging UMLS terminology and CTD annotation came with some drawbacks *. This study did not stress further analysis regarding species coverage, study bias or further potentially occurring stratifying or performance disruptive effects, but it should be considered in future investigations.

In the context of **toxicological information retrieval from biomedical knowledge**, Choi and Lee [2019] applied a comparison of various knowledge representation approaches based on input from MalaCards [Rapaport et al. 2013], CTD and BioGRID [Oughtred et al. 2019]. They determined the best performances for the knowledge representation with a TransE-model when rank-based predicting gene-gene, gene-disease, chemical-gene, chemical-disease

* The UMLS is a biomedical ontological system and focuses on human health. Albeit ecotoxicogenomic considerations were contained in chemical biomolecular level, the CTD gene annotations were human annotated. Although, this helps in terms of interoperability of both data sets, it induces also a curation bias. Furthermore, the overlapping gene annotations (and chemical annotations) did hardly overlap and only a UMLS-centred subset of CTD knowledge was considered. This might increase an already existing study bias in the database.

and disease-symptom relations. The study also presented the model's superiority to the implemented inference approach in CTD. When considering one subject, the TransE model calculated inference scores for each rank-specific object and such gene inference rankings were reasonable for functional enrichment.
In the present study, models trained with UMLS-annotated CTD chemical-gene interactions predicted relationship types with a 94% test accuracy. The learning task was the main conceptual difference from Choi and Lee [2019] to our approach. Whereas Choi and Lee [2019] predicted the subject or object of a triplet, we predicted the relationship type.
Both studies applied the CTD for the toxicogenomic evaluation of the trained models and showed practical strategies to determine toxicologically meaningful exposure-related links in future applications. The present study determined the relationship based on the prediction probability without considering other triplets with the identical subject (or object). Nevertheless, we conceptually elaborated a functional enrichment approach strategy on how to apply CTD knowledge and the UMLS terminology (see supplemental section S3.2). The geneset generation was, in general, possible on different LOBOs. Consequently, sequences of toxicological knowledge across different LOBOs could be generated with the help of relation predictions and functional enrichment.

In the context of AOP development and ET, integrating data from **alternative data sources**, like STITCH [e.g. Taboureau et al. 2013; 2020, Perkins et al. 2017, Schroeder et al. 2016], ToxCast [e.g. Jeong and Choi 2020, Oki and Edwards 2016, Doktorova et al. 2020, Aguayo-Orozco et al. 2019] or the AOPwiki [Martens et al. 2021, Pittman et al. 2018, Aguayo-Orozco et al. 2019] had to be mentioned as additional possibilities for data integration and evaluation for future studies. The recent toxicology-related data integration approaches relied on network inference or merging data but less on machine learning or even deep learning approaches. The present study showed novel data integration strategies considering SemMedDB, UMLS and CTD.

Pre-trained word embeddings for biomedical purposes. Computer scientists employed biomedical knowledge representations in deep learning tasks [e.g. Bojanowski et al. 2017, Devlin et al. 2019, Zhang et al. 2019b, Jimeno Yepes 2017]. Publicly available trained word embedding models and knowledge representations were generated [e.g. Zhang et al. 2019b, Michalopoulos et al. 2021, Alsentzer et al. 2019]. In the recent approaches, the attention-based deep learning models [Vaswani et al. 2017] gained trust [e.g. Zhang et al. 2019b, Lee et al. 2019, Gu et al. 2019]. For example, the BERT model was developed to contextualise word embeddings [Devlin et al. 2019, Chen 2021]. Various training of BERT models focused on knowledge from the biomedical literature [e.g . Michalopoulos et al. 2021,

Alsentzer et al. 2019, Lee et al. 2020]. For example, the UMLSBERT allowed a contextualisation for clinical domain knowledge *. The authors highlighted the superior model performance of the UMLSBERT in comparison to other biomedical BERT models. This model could be applied and task-specifically re-trained. However, it could be disadvantageous that UMLSBERT was considered within clinical research and not in ET. As the present study aimed to shift towards toxicological applications, the UMLSBERT might be misleading due to its context. Moreover, UMLSBERT was trained to identify similar vectors for lexical words belonging to one semantic concept. In the present study, semantic concepts were considered only. Thus, the primary benefit of the re-trained BERT model would not be practical for our prediction purpose. Consequently, a re-training of the UMLSBERT might have its limitations for predicting chemical-biomolecule relationships as it was already re-contextualised.
Zhang et al. [2019b] re-trained a PubMed-based word embedding model with random paths retrieved from MESH to contextualise biomedically the embedding. This knowledge representation — BioWordVec — also recognised subword information with the help of the fastText algorithm [Bojanowski et al. 2017]. The subword embedding was especially helpful for biological and clinical terms with shared suffixes containing essential semantic knowledge. Consequently, out-of-vocabulary words became also vector represented. BioWordVec could embed UMLS semantic concepts into vectors for our applied toxicological task, used as word embedding initialisation. The approach and the model could be helpful also for our prediction task. For example, we could use the hierarchical information or the non-synonymous relationships between concepts from UMLS (see section 2.2.1) similar to the contextualisation through random paths retrieved from MESH. Such contextualisation might be beneficial to improve models trained with SemMedDB.

Negative sampling. In the present study, chemical-biomolecule interactions presented the input sequences for a deep learning prediction model. Also, a change in the prediction task might allow successful learning. In collaboration with Chih Lai and his colleagues, we developed a similar deep learning model using SemMedDB data as input [Lai et al. 2021]. However, we employed the triplet <SUBJECT,PREDICATE,OBJECT> instead of <SUBJECT,OBJECT> pairs for training. Furthermore, we applied no LOBO assignment but pre-selected a broader group of semantic concepts to learn. Consequently, the learning task considered relations beyond exposure-related molecular interactions. However, we trained a similar model architecture with word embedding, LSTM and a subsequent feed-forward neural network. The

* Michalopoulos et al. [2021] re-trained the biomedical BERT model with UMLS terminology resources. The word embedding model was trained to recognise similar meanings for lexical words with the same UMLS semantic concept and capture the meaning of UMLS semantic types.

model predicted whether a random UMLS-annotated triplet should be known in the literature or not.
The input samples were triplets instead of pairs, but the sample information was still somewhat small. To increase the task-specific information of the input, negative samples — randomly chosen triplets that were unknown in SemMedDB — were added to the input. Negative sampling was necessary to have samples representative for both prediction outcomes. Similar to data augmentation, it increased the number of samples. As a result, the prediction model trained with negative samples was highly accurate [Lai et al. 2021].
However, such a learning task comes also with a conceptual limitation. Negative samples were unknown interactions but were expected to be falsified knowledge. Thus, they could also be verified in future and should be considered as a positive sample instead. Consequently, a trained model was highly dependent on the current knowledge status represented in SemMedDB.
The presented approach within this dissertation considered a horizontal and vertical augmentation with hierarchical UMLS knowledge instead of negative sampling. Still, for future investigations, the alternative of negative sampling might be taken into account when adapting the deep learning strategy to another learning task.

4.4 Conclusion

The current toxicological knowledge might link chemical exposure to biological effects on a comprehensive level. We were interested in whether deep learning models with semantic representation could learn from current knowledge and whether the prediction of relations helped retrieve toxicologically meaningful outcomes. In this study, we applied a deep learning task to predict the toxicological relation between a chemical and a biomolecule from literature and a toxicological database.

Published literature knowledge from PubMed was available via SemMedDB and was filtered to chemical-biomolecule interactions. The problem was formulated as a binary classification task, where chemical-biomolecule relations were described as either inhibited or stimulated. We used a biomedical knowledge representation for a toxicological learning task. The selected model predicted 70% of unseen text-based chemical-biomolecule relationships correctly. Thus, the trained deep learning model with a hidden knowledge representation architecture could learn from an input of literature-based relations.
An additional recurrent neural network improved the model performance when considering the empirically-based knowledge from CTD instead of literature-based input from SemMedDB. In such a case, an elongation of the input sequences helped. Finally, we trained the model with CTD relations and reached an accuracy of 94% in predicting unseen chemical-gene interactions from the same source. Again, deep learning with knowledge representation could learn from toxicologically relevant relations.

The deep learning strategy with knowledge representation worked in general. The models were able to predict relationship types in their database scope. However, both models failed to predict known relations of the other data set reliably. In both cases, evaluated prediction models performed slightly better than random guessing. Thus, it was not possible to retrieve toxicologically or biologically meaningful predictions of relationship types regarding multiple data sources. Furthermore, only a part of the SemMedDB data set was covered in the toxicogenomic knowledge of the CTD on gene, pathway and disease level. Consequently, the chosen data sets were potentially limited to accurately predict the knowledge in the scope of the other database. We assume that the stated toxicological coverage was insufficient to retrieve a reliable tool across the knowledge bases.

Nevertheless, we determined a suitable data set and prepared it for toxicological knowledge-based discovery covering the current biomedical literature knowledge of biomolecule interactions with chemical compounds. The chosen SemMedDB data represented already an output of the biomedical NLP-task of named entity recognition and normalisation and was already

used in medical contexts for further NLP or prediction tasks.
With the help of the UMLS, we employed a unified and harmonised biomedical language. The knowledge-based discovery with UMLS was potentially not restricted to biomedicine only. Related domains, e.g. ET, might also benefit from the publicly available tools from the National Library of Medicine. The UMLS terminology and its available tools were shown as a powerful resource for achieving interoperable knowledge and databases. This might help for knowledge-based discovery in ET and AOP research potentially.
The here presented study might not yet enable toxicologists to use deep learning for a chemical-biomolecule interaction prediction which was biologically meaningful and empirically reliably supported across multiple data sources. However, the annotation with the UMLS terminology could be interesting for future computational toxicology studies. This study encourages considering a predictive computational approach with a harmonised and not entirely ET unfamiliar language — the UMLS.

From a future perspective, the merged knowledge should be applied to train a deep learning model. We expect that the merged data scope might allow training a model predicting unseen knowledge from the merged database better than by random guess. Consequently, the model might also be more open for knowledge of other independent toxicological knowledge bases.
Integrating further toxicology knowledge would allow predicting relations on a molecular level and higher levels of biological organisation. Thus, the recent toxicological knowledge representation based on text and empirical data could be merged systematically and efficiently. This might generate new hypotheses of toxicological effects across different entities of biological systems. We should consider these new possibilities as potential strategies for AOP development to generate key event hypotheses, fill knowledge gaps, and discover not yet considered AOPs.

Thousands of chemicals are in the environment, and all have an adverse effect on organisms potentially. Although the entire spectrum of chemicals in the environment has not yet been considered in the recently published knowledge, the information mass is immense and hardly digestible by human-made curations. Therefore, machine learning and text-based discovery approaches might reveal the hidden knowledge and toxicological information about chemicals and their links to biological systems. Thus, a combination of natural language processing, machine learning approaches and data integration has to be considered to link chemical exposure to biological effects on a comprehensive level. This study presents one initial achievement for such strategies in ET research considering the complex data structures of current knowledge.

Chapter 5

Conclusion and Future perspectives

Data for environmental toxicology can be complex and originate from chemical analytical, bioanalytical or omics-based measurements, literature and databases. Factors like the immense size or contradicting relations challenge the investigation of exposure-related biological effects with such data. Besides, their data integration have to deal with, e.g. different vocabularies, different research contexts or a small information coverage. The thesis' objective was to computationally link chemical exposure to biological effects employing such complex data.

In chapter 3, we employed data of an omics-based exposure study considering mixture effects in freshwater. We applied three approaches and different exposure scenarios to disentangle environmental mixture exposure effects and reliably attribute biological outcomes to chemical drivers on gene and pathway levels. The correlation-based compound groups helped understand some xenobiotic effects applying differential gene expression and network inference.

Published knowledge represents comprehensively complex information from environmental toxicology. In chapter 4, we employed semantic predications from a current text-based biomedical knowledge base and curated knowledge from a toxicological database. We implemented a word embedding neural network with a subsequent feed-forward network that predicted the toxicogenomic relationships of chemical-biomolecule interactions. Data augmentation and recurrent neural networks were beneficial for training with curated toxicological knowledge.

The developed approaches allowed assessing the hazard of chemicals more systematically, e.g. with correlation-based compound groups, and support the prioritisation of chemicals for testing, e.g. with prediction models. This section sets the thesis' achievements in the context of environmental toxicology to understand exposure-related molecular effects through method comparison, deep learning and data integration.

5.1 Conclusion

In this dissertation, we took examples of complex environmental data into consideration to link chemical exposure to biological effects applying computational methods based on linear modelling, network inference, and machine learning. We employed omics-based measurements in exposure studies and data from published and curated knowledge bases. Both studies gave important and joint insights regarding the linkage of chemical exposure to molecular effects.

5.1.1 Investigating complex mixtures in the environment

When determining exposure-related effects on an omics-based molecular level, environmental toxicologists frequently considered multiple approaches to filtering down results, refining the specificity or sensitivity, and integrating knowledge on the pathway or disease level. However, a strategy was missing to compare stand-alone computational methods, when applied to the same task, particularly for assessing ecological hazards and biomonitoring. Therefore, we investigated the extent to which computational approaches were suitable to link complex chemical mixture exposure to biological effects. To be more precise, we linked complex chemical exposure in freshwater sites to transcriptional effects in fathead minnow liver tissue. The aim was to investigate three stand-alone computational approaches in their suitability to determine exposure-related effects on molecular and pathway levels, highlighting a biologically meaningful and reliable attribution to adverse effects. Given the data of the preliminary exposure study, highly correlated and subtle chemical exposure patterns led to weak transcriptional effects. However, the application of DEA and WGCNA, stand-alone and in combination, was practical to verify the *in-vitro* measured xenobiotic stress and endocrine disruption. We determined potentially endocrine effect-driving subsets of the measured chemical exposure.

The method combination resulted in biologically relevant results. However, these results were distinct from those of single experiments. Both, applying approaches stand-alone and in combination, were sound to gain a comprehensive understanding of exposure-related effects on a molecular level. Each approach had its assumptions and limitations. Thus, significant outcomes were interesting in light of the approach's assumptions but might define and weigh characteristics of interestingness differently. In this thesis, comparing overlapping and singularly detected results identified limitations when considering transcriptomic data measured in biological systems after complex mixture exposures. Such limitations were (1) the predefined set of investigated chemicals of concern, (2) the not-independent exposure patterns of compounds due to a relatively small set of selected sites and (3) the low concentrations of

detected compounds affecting fish somewhat unspecific.

Nevertheless, considering different exposure scenarios allowed disentangling environmental complex mixtures, in parts, to subgroups of potential chemical drivers. The study presented differences in examined transcriptional effects for overall, stream-wise, single chemical and chemical group exposures. Especially in the context of complex environmental mixtures with lowly concentrated chemicals, the presented approach and elaborations shed light from multiple perspectives that may be all relevant in exposure-related studies. In this study, the novel perspective of correlation-based compound groups allowed determining potential drivers of xenobiotics effects.

In conclusion, DEA and WGCNA were suitable for linking endocrine disruption and xenobiotic stress responses to a subset of co-correlated compounds, although the investigated complex mixtures consisted of lowly concentrated chemicals only. The presented strategies and established methods helped understand chemical exposures in the selected stream water sites comprehensively and allowed assessing the risk of chemicals more systematically. In that respect, the study in chapter 3 explicitly contributed to the investigation of *proxies of the eco-exposome* by delivering achievable computational strategies for assessing the environmental hazard for aqueous species.

5.1.2 Complex knowledge from literature and curated databases predict chemical-biomolecule interactions

Although stressor-agnostic, toxicologists used the AOP framework to investigate links between chemical stressors and molecular effects or adverse outcomes on higher biological levels. Exposure-associated effects on the molecular level helped understand the early steps of toxic mechanisms and correlate them to adverse outcomes. Therefore, data integration and knowledge representation approaches were fruitful to infer new information from databases and text. We were interested in whether information-rich knowledge from literature and databases is suitable for learning toxicologically meaningful exposure-related interactions.

This thesis trained deep learning models employing toxicology-related knowledge of semantic predications from SemMedDB and curated knowledge from CTD. We evaluated the models' suitability to predict toxicologically meaningful chemical-biomolecule interactions. In particular, we predicted chemical-biomolecule relations considering natural language processed data and subsequent deep learning as novel approach in the context of ET.

Toxicological databases and also literature have various terminologies. The UMLS contains various vocabularies from biomedicine and allows unifying lexical names in the biomedical

context through assignment to semantic concepts and semantic groups. For example, it allows unifying various chemical names and pharmaceutical products for similar compounds. However, the UMLS is biased towards biomedicine and human health, and many concepts are not relevant for toxicology. We presented a straightforward approach to applying UMLS in a toxicological context by adding a LOBO assignment to the UMLS terminology. Therefore, we considered biologically relevant terminologies across all LOBOs. Consequently, the selected UMLS terminology lost at least some of its biomedical bias, including a smaller ratio of clinical centred terms. Before, such a framework to assign semantic concepts to toxicological entities was not available.
The UMLS resources had another valuable purpose. Integrating the knowledge from the UMLS Metathesaurus to the input of SemMedDB or CTD allowed the hierarchically structured consideration of more generalised biomedical concepts. The parental terms allowed expanding the data set in the number of samples and elongating relational sequences. For CTD, horizontal augmentation was the key to improving the deep learning model performance when training a model with a recurrent neural network. We determined that the CTD set had a ten times higher information density than the SemMedDB data set. Thus, the input might need a sufficient information density to benefit from parental terms, which was likely not the case for the prepared input from SemMedDB.

Conclusively, the current and information-rich knowledge from SemMedDB and CTD were separately suitable to predict unseen exposure-related interactions from the same data source. We made a novel prediction tool for ET available that learned UMLS-annotated and -integrated chemical-gene interactions. The applied computational strategy expanded the spectrum of predictive toxicology and evaluation approaches for exposure-related and omics-based data.

5.1.3 Linking chemical exposure to biological effects by integrating CTD

In both studies of this thesis, we applied the toxicologically relevant knowledge base CTD for data integration. CTD allowed us to look at exposure-related data across different LOBOs and to validate the biological meaning of outcomes on gene and pathway levels. In chapter 3, CTD and STITCH proved the applicability of three stand-alone methodologies to determine biologically meaningful and reliable results. The overlap on the gene level and the significances of overlap on pathway level helped compare the computational approaches. In chapter 4, we highlighted the versatility of CTD. Initially, we considered chemical-gene interactions from CTD as evaluation data set to assess the biological reliability of the predictive model. An up-scaling to pathway and disease level, similar to the approaches in the first study, helped

assess the coverage of toxicogenomic relevant knowledge in used literature-based knowledge. We assumed to handle less noise in data when up-scaling on a lower resolution level, e.g. from gene to pathway level. Potentially up-scaled results might be generalised, trivial or non-informative, and have to be considered with caution. Nevertheless, for both studies, the lower resolution level for outputs was beneficial in evaluating the computational approaches in their biological meaning.

The exposure-related biological effects on the gene level from CTD were beneficial to train a deep learning model. This CTD application could be considered a bridge between the two applied studies of this dissertation. The database comprised knowledge retrieved from empirically measured omics-based investigations. Thus, CTD contains information which are determined with computational strategies as compared and combined in chapter 3.
The text-based input from SemMedDB comprises knowledge from abstracts and titles only. Thus, the SemMedDB knowledge did not cover the complexity of knowledge in published scientific papers. Moreover, the knowledge from an abstract might not entirely represented in SemMedDB considering the limitations of SemRep [Cong et al. 2019] and the bias of the UMLS terminology. Databases like CTD are information-rich, especially in a meta-analytical scope. For the study in chapter 4, it was essential that CTD contained human-curated information from publicly available data but not necessarily from published papers. We potentially considered a deeper investigation level of published knowledge applying CTD instead of SemMedDB. Both resources were useful for deep learning prediction tools.

Conclusively, the computational investigation of empirical measurements for one specific site, one specific model organism and one specific environmental system (chapter 3) benefited from curated toxicological knowledge. Moreover, an investigation of knowledge representative data (chapter 4) profited from integrating such external toxicological information from CTD.

5.2 Future perspectives

The investigations in this thesis computationally linked chemical compounds to biomolecules employing complex environmental data sets. We considered two challenges within the ET research.

In chapter 3, we assessed the environmental status of small streams and adverse effects on one studied fish species. We focused on three methods to determine biologically meaningful results with reliable attributions when investigating complex mixtures of lowly concentrated chemicals. Future tasks for methodological comparisons for exposure-related and omics-based studies have arisen from the findings in chapter 3. Recent studies also highlighted other

approaches, e.g. toxicogenomic profiling with self-organising maps, which were not considered for the comparison in this thesis. Moreover, the study in chapter 3 did not emphasise the full potential of method integration combining the results of DEA, AR or WGCNA and other external references. We considered straightforward strategies. However, studies showed that more sophisticated ones [e.g. Schroeder et al. 2017, Sutherland et al. 2018] or meta-analysis approaches [e.g. Krämer et al. 2020, Ewald et al. 2020] were meaningful to examine exposure-related effects and, thus, might be relevant equally relevant in respect to our research question.

The study in chapter 4 generated knowledge-driven hypotheses of chemical-biomolecule interactions that might also be a fruitful contribution to AOP development. In chapter 4, we developed models to predict exposure-related molecular effects based on given toxicological knowledge, represented through complex data from literature and analysis of exposure-related studies. We identified limitations of the considered data sets, which were pressing to future studies. For example, integrating the knowledge from CTD, UMLS and SemMedDB might retrieve a comprehensive prediction tool in ET. This integration would help determine new hypotheses (chapter 4) and assess chemical-related effects in the environment (chapter 3).

Based on the thesis' achievements, we determined future tasks and challenges as presented in more detail in the following.

Knowledge representations as AOP networks. The network principle has been frequently used to project current knowledge and infer new links. UMLS is also helpful for network inference in ET research. Assigning and unifying words to LOBOs allows generating a graph representation of knowledge from databases or parsed text. Such a graph would have structurally much in common with an AOP network. Whereas the nodes would be biological entities or events on a specific LOBO, represented by, e.g. UMLS semantic concepts, the given relations would define the direct edges. These relations could be considered empirical evidence originating from an expert-based publication or a statistically significant outcome curated in a toxicological database. By adding the information of LOBOs as node characteristics, we could identify chains across different levels of biological organisation. The presented approach in chapter 4 and trained models are helpful to such a graph representation for weighting, adding or doubting edges between two semantic concepts. Albeit a lot of work and steps in between are necessary, such a strategy potentially allows data integration across multiple data- and textbases to retrieve a knowledge graph representation with a unified vocabulary of UMLS semantic concepts. Although such a network generation and inference approach were not in the scope of this thesis, the highlighted characteristics of UMLS will allow a linkage of databases and literature for AOP development approaches in the future.

UMLS for environmental toxicology. The UMLS is a semantic terminology built for biomedical research. The UMLS terminology has been already extended to other research fields [Rosemblat et al. 2013b] like medical informatics [Zhao et al. 2021], public health [Rosemblat et al. 2013a] or pharmacogenomics [Ahlers et al. 2007]. In general, this ontological extension is also achievable for the ET context. However, such a domain-switch of an ontology will be a resource-intensive task. An interdisciplinary team of environmental toxicologists, chemists and biologists will be needed with further expert knowledge from linguistics and computational sciences. This task will comprise various time-consuming tasks, e.g., deciding whether current concepts are necessary, which concepts are missing, how to re-frame the semantic network, or whether re-definitions of relationship types are necessary. Furthermore, it will be necessary to reduce the selection of data and add further toxicology specific databases as the UMLS semantic network comprises biomedical and biological sources.
However, a UMLS extension to the ET context could allow different deliverables. For example, we could generate an ET-centered SemMedDB of PubMed-Abstracts and titles. Consequently, applying the strategy from chapter 4 would result in an ET-centered and literature-based deep learning prediction model. The adapted semantic terminology would be helpful for a predictive AOP hypothesis generation — for chemical-biomolecule interactions and interactions with other LOBOs of the AOP framework.

Identify knowledge from literature. SemMedDB is a high-content and comprehensive data source. The NLP tool SemRep has retrieved the SemMedDB input from PubMed literature abstracts. It is regularly updated with the most recent PubMed citations, but do not comprise text-based information beyond available PubMed abstracts and titles. However, it will potentially help parse entire Publications with SemRep. In terms of effort, only a selection of publications should be considered. A time efficient NLP tool to find a subset of relevant publications, particularly in AOP development, is the `AOP-helpFinder` webserver [Jornod et al. 2021], which associates stressors and biological events with the help of an artificial intelligence screening of PubMed. In ET research and AOP development, selection factors will be, e.g., compounds of interest, taxonomic groups, study designs, or investigated biological systems.
Furthermore, other texts as peer-reviewed publications will be helpful. For example, in biomedical literature-based discovery, researchers have considered electronic medical records. Equivalents in ET research are, e.g. reports of biomonitoring, descriptions of AOPs and KEs in the AOPwiki or descriptions of biological pathways and networks from KEGG or Reactome. The application of SemRep will help integrate further literature-based knowledge to SemMedDB easily and will already be unified with UMLS semantic concepts. Thus, knowledge from

toxicology-relevant literature beyond SemMedDB will become interoperable with toxicological databases.

Combined input of SemMedDB and CTD to train a deep learning model. The UMLS offers, next to hierarchical information of parents, a data set for sibling information (see section 2.2.1). This allows determining further relation pairs, which are not considered in SemMedDB or CTD, respectively. When considering only semantic concepts already represented in the inputs, we will increase the information density (e.g. measured as sparsity in the occurrence matrix) with every additional sample. As observed in chapter 4, the CTD input had a ten times higher information density, which might be a potential reason, that the application of a horizontal augmentation and LSTM layer was beneficial. By considering the sibling information, the SemMedDB will have an increased information density and thus could also benefit of horizontal augmentation and a recurrent neural network layer.
An alternative way of data augmentation will be increasing the number of samples with the SemMedDB input by relations from CTD. Pre-trained word embeddings and retro-fitting have been shown advantageous in the biomedical context (see section 1.2.5). In conclusion of the outcomes in chapter 4, we will retrain the word embedding matrix in the separate SemMedDB-based model with the fully initialised, partly trained word embedding matrix from the CTD model. Such a word embedding will represent the semantic meaning of relations from CTD and SemMedDB. This approach could improve the evaluation of the SemMedDB model with unseen data of the CTD input. However, further sources — e.g. STITCH, ToxCast or EcoToxDB * — should be taken into account for evaluation of models trained with the combined toxicological knowledge from CTD and SemMedDB.
Consequently, various follow-up steps to the here shown achievements in chapter 4 are possible to aim for a comprehensive tool to generate hypotheses for molecular key events. Such further investigations could be helpful for the entire AOP development. Moreover, such upcoming achievements will have their use for exposure-related omics-based studies and environmental hazard assessment as highlighted in the following paragraphs.

Intelligent chemical selection. Switching the purpose of the developed predictive tool from chapter 4 towards environmental hazard assessment will also help future studies interested in using the presented computational strategies from chapter 3. The prediction of chemical-biomolecule-interactions under the purpose of an intelligent selection of chemicals helps also exposure studies of environmental sites. For example, a predictive tool, as presented in chapter 4 could prove recent lists of chemicals of concern, e.g. by predicting the

* https://cfpub.epa.gov/ecotox/

chemicals which are likely to be associated to genes of a known adverse outcome or biological pathway. Consequently, such a predictive tool would support the prioritisation of chemicals for testing, which would be useful for hazard assessment and biomonitoring.

Evaluate knowledge from transcriptomic exposure studies. As shown in chapter 3, exposure studies that apply omics-based approaches, such as microarray gene expression analysis, generate exposure-related associations of molecules or even higher biological effects. This thesis showed that variations in biological and statistically significant effects could be determined across assumed exposure scenarios and applied methods. Methods have their limitations potentially leading to some false-positive results. To validate the outcomes, a reliable external reference is necessary. In current ET research, such validation and evaluation systems have commonly been based on toxicological databases such as CTD. The presented predictive strategy from chapter 4 would have its means for evaluating exposure-related biological effects, e.g., those from chapter 3. Especially, a prediction tool trained on multiple references, e.g. SemMedDB and CTD, would present a comprehensive perspective on current toxicological knowledge. Models fulfilling some prerequisites, e.g. high test accuracies for knowledge within training data source and across other databases, would have the future potential as evaluation system for omics-based exposure studies.

In conclusion, this thesis compared computational approaches and developed predictive models to link chemical exposure to biological effects on molecular level with complex environmental data. With the help of knowledge and data integration, the achievements of this dissertation and its future perspectives allow benefiting of the complexity of data. The thesis' achievements might support the prioritisation of chemicals for testing and an intelligent selection of chemicals for monitoring in future exposure studies. This will help discover new knowledge and verify empirically measured information for an ET context comprising biomonitoring and hazard assessment purposes, but also a better conceptual understanding in e.g. the AOP development. In consequence, this thesis contributed not only to a better understanding of exposure-related effects in anthropogenically perturbed environments, but also to computationally investigate *Proxies of the Eco-exposome*.

Chapter S1

Supplement Chapter 1

S1.1 Example of an estrogen bioassay

In figure S1-1, the estrogen activity becomes easily detectable through luminescence. The luminescent intensity in a bioassay-test is equivalent to the joint toxic potency of all compounds in a mixture with the same mode of action, assuming a concentration addition. In this example, the activation of the estrogen receptor affects the biological endpoint 'reproduction'. A toxic or bioanalytical equivalent concentration is defined to compare bioassay measures of sites. For estrogen activity, the equivalent activity levels of natural or synthetic estrogens are referred like 17-β-estradiol (E2) or 17-α-Ethinyl-estradiol (EE2). Based on this equivalent value to a well-investigated estrogen, the severity of the adverse reproductive outcome gets estimated without testing a whole organism. The ecological risk is determined by comparing the equivalent estrogen concentration to a standardised and previously defined threshold.

S1.2 Types of mode of action

After the intake of a compound in a biological entity, it may interact with a protein as a binding ligand. Such biomolecules may be receptors, enzymes or other target proteins. Under normal conditions, the receptors respond to specific endogenous signalling ligands, e.g. hormones or neurotransmitters, and lead to a cell regulating response by, e.g. interfering with ion channels G-protein coupled receptors or nuclear receptors. However, exogenous compounds may have a chemical with a similar active group as the endogenous ligand leading to a concurrency for the receptor binding sites with the natural ligand. Consequently, xenobiotic ligands may activate the receptor protein as an agonist or inactivate the receptor as an antagonist. Thus, the overall receptor activity level in a biological entity may be up- or downregulated by the xenobiotic influences. For example, the interactions with nuclear receptors affecting hormone regulation may induce *endocrine disruption*, or the affection of neurotransmitter interactions infer with ion channels and may induce *neurotoxicity*.

Next to the receptors, other proteins may interact with xenobiotic compounds and lead to *cytotoxicity* when negatively regulated in their protein activity. Covalent or non-covalent bindings provoke such interactions leading to irreversible or reversible protein inactivation. In consequence, metabolic processes, local and peripheral transport of ligands, or cytoskeleton stability are affected. A *teratogen* chemical perturbation disturbs a developmental stage of an organism as like an embryo or produces a malformation.

The intake of xenobiotic compounds may also induce a xenobiotic defence. Thus, the *xenobiotic metabolism* transforms the compound into an endogenous metabolite, excreted through

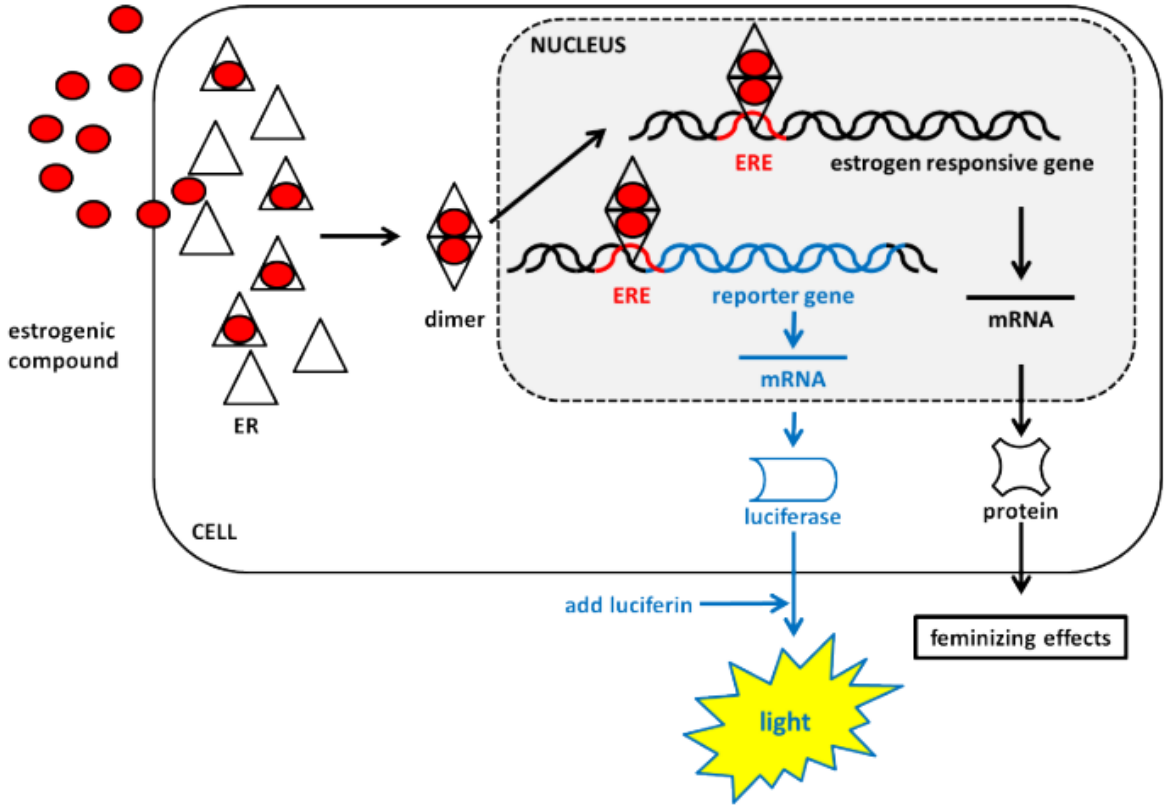

Figure S1-1. ***Example of a cell-based bioassay.*** *An exogenous compound binds to an estrogen receptor and induces a dimerisation of estrogen receptors. The dimer migrates in to the nucleus and binds to the DNA and induces the transcription of a gene important for the hormone signaling pathway. The bioassay-cell is genetically manipulated and transcribes next to the estrogen-related gene, the reporter gene luciferase. Added luciferin binds to the produced luciferase-protein and emits light. The luminiscent activity in an bioassay-test is equivalent to the mixtures potency to bind to the estrogen receptor. Figure taken from [Kraak, Michiel 2021].*

faeces or urine [van Straalen, Nico M. 2021a].
The main location of xenobiotic metabolism is the liver tissue. The activating enzymes cytochrome P450 and NADPH cytochrome P450 reductase are mainly produced in hepatocytes. After enzymatic activation, hydrophilic groups are introduced into the xenobiotic compound. This reaction allows further conjugations through transferase enzymes, which transforms the xenobiotic to an endogenous metabolite, ideally with an increased property of water solubility. The biotransformed xenobiotic will be excreted either by transport via bile and gut to faeces or via kidney to urine.

However, the xenobiotic metabolism may also cause toxicity. The activation of a xenobiotic through cytochrome P450 may make the xenobiotic extremely reactive. For example, xenobiotic polyaromatic hydrocarbons build DNA adducts after activation and induce mutations. Thus, next to protein interactions, exogenous compounds may also bind to other biomolecules such as DNA, or RNA leading to *genotoxic* or *carcinogenic* effects. Further xenobiotics, e.g. polychlorinated biphenyls, do not quickly degrade, albeit the enzymes for biotransformation are induced in high potency. In general, a xenobiotic compound may cause ROS formation in various ways, e.g. by redox reactions, or indirectly, e.g. by interaction with ROS-scavenging antioxidants. Consequently, the induced cytochrome P450 increasingly forms oxygen metabolites, called reactive oxygen species (ROS). The compound itself does not cause oxidative stress, but the ROS binds to endogenous biomolecules like proteins or DNA and leads to cellular damage. Oxidative stress and biotransformation of xenobiotics are cellular responses due to stress. Thus, biotransformation and oxidative stress are co-dependent processes on a molecular level [van Straalen, Nico M. 2021a].

Integrating chemicals into cellular and mitochondrial membranes may disturb the phospholipid bilayer's integrity and functioning, leading to *membrane damage and narcosis* [van Straalen, Nico M. 2021a]. Chemicals may be partitioned in the lipid bilayer, not dependent on the chemical compound. Thus, each xenobiotic compound exerts this MOA, which is considered the baseline of toxic effect. Nevertheless, each chemical compound has another lipid-water-partitioning coefficient and thus differ in their potency of *baseline toxicity*. Next to narcosis, changes in the electrolyte gradient of the membrane can disrupt the membrane integrity through augmented or reduced ion transport.

Furthermore, chemicals may affect immune cells and induce various forms of *immunotoxicity*. The immune system is complex consisting of different cell types. A network of cellular and component interactions protects an organism from infections and pathogens, such as xenobiotic compounds. A xenobiotic perturbation may induce immunoregulation via the innate or the adaptive immune system. For example, a xenobiotic increased ROS production may

activate an innate immune response leading to the production and the release of cytokines in macrophages [Dong et al. 1998].

Synthetically designed pharmaceuticals and biocides are produced to affect specifically to cell types, species-specific proteins or biochemical pathways. Thus, a plethora of chemicals has *further target-specific* MOAs practical for the human population to maintain health and produce food and products for industry and personal care. However, these MOA may also affect non-target organisms when released into the environment and may also have a cumulative adverse effect on public health due to the food web.

The concept of MOA is frequently used to describe toxic effects in environmental toxicology and helps classify compounds and exposures by their molecular effects. In general, a chemical compound may perturb organisms across multiple MOAs. Therefore, researchers aim to identify the concentration levels in environmental monitoring and risk assessment, where MOAs may get active in biological entities. Mixture considerations are essential and frequently investigated when considering cell- or species-specific effects. In general, the MOAs, defined by specific interactions with biomolecules, have lower effect levels but are more restricted to specific groups of chemical exposures. Therefore, Laboratory experiments considering single compounds or artificial mixtures of chosen compounds may reflect specific and unspecific effects on the chosen model organism. However, the investigation of environmental samples expands the complexity of the diverse compounds enormously. Significantly, the mixtures of lowly concentrated or even undetectable - albeit present - compounds influence the xenobiotic perturbation as it affects mixtures of specific MOAs and impacts narcotic effects.

S1.3 The dogma of molecular biology.

In 1953, Francis Crick and James Watson have described the double-stranded deoxyribonucleic acid molecule (DNA) [Watson and Crick 1953] *. The blueprint of one or multiple functions † is conserved in the DNA sequence which defines a gene. The DNA has been already understood as a carrier system of genetic information, but also other biomolecules are relevant to releasing a gene's molecular function. In this respect, a second essential publication of Francis Crick in 1970 has helped understand the biomolecular flow of genetic information

* Shared with Maurice Wilkins, both researchers were honored with the Nobel price in Physiology or Medicine 1962. However, it has to be highlighted, that the assisting chemist Rosalind Franklin made also an important contribution. She was the first researcher, who imaged the DNA molecule via X-ray diffraction, which allowed the deciphering of the DNA's double-helix structure.

† In the case of alternative splicing, identical transcribed mRNA sequences may be modified to different pre-translational mRNA transcripts, where each may generate different proteins with different functions.

and led to the description of the central dogma of molecular biology [Crick 1970](see figure S1-2). The two main processes of *transcription* and *translation* describe the transfer of genetic information from DNA sequences via ribonucleic acid (RNA) to a protein.

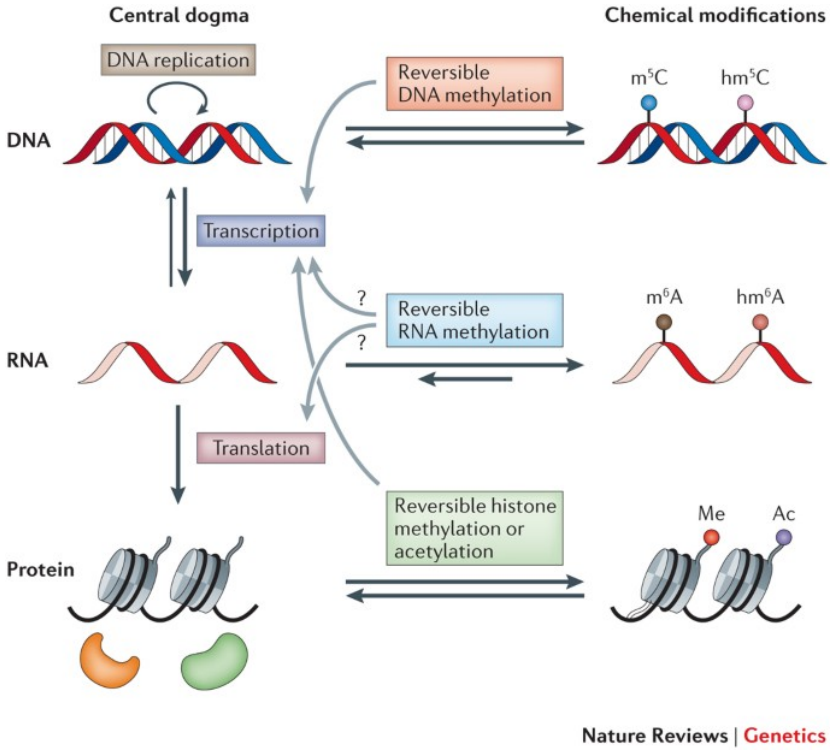

Figure S1-2. ***The central dogma of molecular biology*** *The biological processes of transcription and translation transform genetic information from DNA via RNA to proteins. In all three cases, chemical modifications may stabilise or activate the molecules in the environment of a cell. Whereas genetic information on RNA level is reversible to the DNA level, the protein level is irreversible. Taken from [Fu et al. 2014].*

The transcription describes the molecular process of generating a primary transcript of a DNA sequence section in a complementary RNA sequence. In general, two types of RNA molecules may be generated: messenger or non-coding RNA. The messenger RNA (mRNA) is the intermediate product used as a stencil in the molecular process of translation to produce an amino acid sequence forming the protein. Therefore, mRNA migrates from the nucleus to the ribosome. The non-coding RNA (ncRNA) does not contain a protein-coding sequence but forms molecules, which are essential for gene expression regulation, e.g. small nuclear RNA and microRNA, or for translation, e.g. transfer RNA (tRNA) and ribosomal RNA (rRNA). The rRNA is an essential building block for a ribosome - the cell compartment where the translation occurs. The tRNA molecules bind and transport amino acids. During translation,

the presented anticodon sequences of tRNAs bind complementary to the mRNA sequence in the ribosome. The tRNA-docked amino acids are chained up sequentially via peptide bonds, and the built protein is released into the cell's plasma. As also highlighted in figure S1-2, DNA, RNA and proteins are chemically modified, e.g. the polyadenylation of the end of a transcribed mRNA or the splicing of post-transcribed mRNA in eukaryote cells. The latter process may lead to different pre-translational mRNA transcripts, albeit read out from the same DNA sequence during transcription and thus to different protein products.
Independent of translating either mRNA or ncRNA from a DNA sequence, biomolecules are produced, which fulfil some function in some biological entity. Genetic feedback loops may positively or negatively regulate the gene translation based on the stimulus *, which results in an up- or downregulation in transcription and thus more or fewer copies of specific mRNA transcripts. Such extracellular stimuli may also be endogenous or exogenous xenobiotics, like xenobiotic uptake by the organism, and may induce a direct or indirect gene regulation or lead to an induced or a repressed cellular response cascade.

S1.4 Transcriptomics

One may choose between the hybridisation- or sequence-based transcriptomic approaches. The *hybridisation-based* techniques are based on microarrays - small glass plates with fixed labelled cDNAs. The microarray design shall represent the complete gene complement of the organism [van Straalen, Nico M. 2021b]. For analysis, a mixture of different RNA molecules is extracted and enriched from samples of lyzed cells. After induced reverse transcription, the prepared samples of cDNAs are pooled, e.g. in treatment and control groups. Then, they are hybridised to the manufactured or customised set of probes on a microarray. The mRNA molecules bind sequence-specific to probes of labelled cDNA. The microarray analysis quantifies the amount of a specific set of mRNA transcript copies on microarrays - small glass plates with fixed labelled cDNAs. Per exposure condition, the normalised intensity of gene expression per probe is detected, and a gene's response is expressed relative to the measured intensity of transcripts to a control condition. Further preprocessing steps are used to correct, for example, significant gene expressions from the background noise. Then, computational and bioinformatics approaches are used to model gene expression dependent on exposure to environmental chemicals.

* A cellular function may be not only managed by changes in translation. Such stimuli may also regulate post-transcriptional modifications and translational processes, and lead to more or less gene products or metabolites.

In comparison to microarray-based approaches, RNA-seq is considered to be a more systematic analysis of gene expression patterns [Qian et al. 2014]. The total RNA from (pools) of biological entities are isolated first, and reverse transcription is induced. The sequence sizes are selected, and cDNA gets labelled with barcode labels to facilitate the sequencing. Sequencing cDNA pools allow the transcriptome assembly based on a reference genome. If no reference genome for a species is available, the analysis requires a greater sequencing depth and usually ends in many incomplete transcripts (*de-novo* sequencing). The output is corrected to equalise effects of, e.g. total RNA yield, library size, and gene length. Per exposure condition, the normalised number of transcripts per gene is counted, and a gene's response is expressed relative to the number of transcripts to a control condition. Considering count instead of light intensity data, data processing and statistical analysis slightly differ. The advantages of RNA-seq are its ability to quantify a broad coverage of RNA transcripts, including the unknown variants (e.g. splice variants), and its better applicability for experiments in non-model organisms. However, this may also be the main limitation, as computational analysis may become more cumbersome [Martins et al. 2019, Qian et al. 2014]. Although hybridisation-based is taken over by the rapid and high-throughput sequence-based approaches, the microarrays are still frequently used in environmental toxicology due to (ostensible) cost-efficiency, their standardised computational and bioinformatics analysis, and prioritisation of the assessed genes [Martins et al. 2019].

Chapter S2

Supplement Chapter 3

Table S2-1. *List of selected compounds for chemical analysis in stream water samples. (Continues on next page)*

Chemical Name	CAS	PubChem CID	Chemical Name	CAS	PubChem CID
4-Methyl Benzotriazole (4TTZ)	29878-31-7	122499	Lincomycin	154-21-2	3000540
5,6-Dimethyl Benzotriazole	4184-79-6	77849	Lomefloxacin	98079-51-7	3948
5-Methyl Benzotriazole	136-85-6	8705	Miconazole	22916-47-8	4189
Benzotriazole	95-14-7	7220	Norfloxacin	70458-96-7	4539
2-aminobenzothiazole	136-95-8	8706	Norgestimate	35189-28-7	6540478
2-Hydroxy Benzolthiazole	934-34-9	13625	Ofloxacin	82419-36-1	4583
Benzothiazole	95-16-9	7222	Ormetoprim	6981-18-6	23418
4-Nonylphenols	84852-15-3		Oxacillin	66-79-5	6196
4-Nonylphenol monoethoxylates	84852-15-3		Oxolinic Acid	14698-29-4	4628
4-Nonylphenol diethoxylates	84852-15-3		Penicillin G	61-33-6	5904
Octylphenol	1806-26-4	15730	Penicillin V	87-08-1	6869
BisphenolA	80-05-7	6623	Roxithromycin	80214-83-1	133611834
Triclosan	3380-34-5	5564	Sarafloxacin	98105-99-8	56208
Acetaminophen	103-90-2	1983	Sulfachloropyridazine	80-32-0	6634
Azithromycin	83905-01-5	447043	Sulfadiazine	68-35-9	5215
Caffeine	58-08-2	2519	Sulfadimethoxine	122-11-2	5323
Carbadox	6804-07-5	5353472	Sulfamerazine	127-79-7	5325
Carbamazepine	298-46-4	2554	Sulfamethazine	57-68-1	5327
Cefotaxime	63527-52-6	5742673	Sulfamethizole	144-82-1	5328
Ciprofloxacin	85721-33-1	2764	Sulfamethoxazole	723-46-6	5329
Clarithromycin	81103-11-9	84029	Sulfanilamide	63-74-1	5333
Clinafloxacin	105956-97-6	60063	Sulfathiazole	72-14-0	5340
Cloxacillin	61-72-3	6098	Thiabendazole	148-79-8	5430
Dehydronifedipine	67035-22-7	128753	Trimethoprim	738-70-5	5578
Diphenhydramine	58-73-1	3100	Tylosin	1401-69-0	134693945
Diltiazem	42399-41-7	39186	Virginiamycin M1	21411-53-0	46936184
Digoxin	20830-75-5	2724385	1,7-Dimethylxanthine	611-59-6	4687
Digoxigenin	1672-46-4	15478	Furosemide	54-31-9	3440
Enrofloxacin	93106-60-6	71188	Gemfibrozil	25812-30-0	3463
Erythromycin-H2O	114-07-8	12560	Glipizide	29094-61-9	3478
Flumequine	42835-25-6	3374	Glyburide	10238-21-8	3488
Fluoxetine	54910-89-3	3386	Hydrochlorothiazide	58-93-5	3639
10-hydroxy-amitriptyline (Oxalate)	1246833-15-7	131871090	2-Hydroxy-ibuprofen	51146-55-5	10443535

Chemical Name	CAS	PubChem CID	Chemical Name	CAS	PubChem CID
Ibuprofen	15687-27-1	3672	Meprobamate	57-53-4	4064
Naproxen	22204-53-1	156391	Methylprednisolone	83-43-2	6741
Triclocarban	101-20-2	7547	Metoprolol	51384-51-1	4171
Triclosan	3380-34-5	5564	Norfluoxetine	83891-03-6	4541
Warfarin	81-81-2	54678486	Norverapamil	67018-85-3	104972
Albuterol	18559-94-9	2083	Paroxetine	61869-08-7	43815
Amphetamine	300-62-9	3007	Prednisolone	50-24-8	5755
Atenolol	29122-68-7	2249	Prednisone	53-03-2	5865
Atorvastatin	134523-00-5	60823	Promethazine	60-87-7	4927
Cimetidine	51481-61-9	2756	Propoxyphene	469-62-5	10100
Clonidine	4205-90-7	2803	Propranolol	525-66-6	4946
Codeine	76-57-3	5284371	Sertraline	79617-96-2	68617
Cotinine	486-56-6	854019	Simvastatin	79902-63-9	54454
Enalapril	75847-73-3	5388962	Theophylline	58-55-9	2153
Hydrocodone	125-29-1	5284569	Trenbolone	10161-33-8	25015
Metformin	657-24-9	4091	Trenbolone acetate	10161-34-9	66359
Oxycodone	76-42-6	5284603	Valsartan	137862-53-4	60846
Ranitidine	66357-35-5	3001055	Verapamil	52-53-9	2520
Triamterene	396-01-0	5546	Diatrizoic acid	117-96-4	2140
Alprazolam	28981-97-7	2118	Iopamidol	60166-93-0	65492
Amitriptyline	50-48-6	2160	Citalopram	59729-33-8	2771
Amlodipine	88150-42-9	2162	Tamoxifen	10540-29-1	273356
Benzoylecgonine	519-09-5	448223	Cyclophosphamide	50-18-0	2907
Benztropine	86-13-5	1201549	Venlafaxine	93413-69-5	5656
Betamethasone	378-44-9	9782	Amsacrine	51264-14-3	2179
Cocaine	50-36-2	446220	Azathioprine	446-86-6	2265
DEET	134-62-3	4284	Busulfan	55-98-1	2478
Desmethyldiltiazem	84903-78-6	25834477	Clotrimazole	23593-75-1	2812
Diazepam	439-14-5	3016	Colchicine	64-86-8	6167
Fluocinonide	356-12-7	9642	Daunorubicin	20830-81-3	30323
Fluticasone propionate	80474-14-2	444036	Doxorubicin	23214-92-8	31703
Hydrocortisone	50-23-7	5754	Drospirenone	67392-87-4	68873
Etoposide	33419-42-0	36462	Moxifloxacin	151096-09-2	152946
Oxazepam	604-75-1	4616	Medroxyprogesterone Acetate	71-58-9	6279
Metronidazole	443-48-1	4173	Rosuvastatin	287714-41-4	446157
Teniposide	29767-20-2	452548	Zidovudine	30516-87-1	35370
Melphalan	148-82-3	460612			

Table S2-2. ***Site information of 10 selected streams in Minnesota.*** *Next to geographical Description, number of investigated microarray samples (#S) and number of detected chemical compounds (#C) are listed per stream site.*

	Site	Stream_ID	Ecosystem	Latitude	Longitude	#S	#C
1	Bevens Creek	15EM014	Mixed Wood Shield	44.714898	-93.699639	6	13
2	Tributary to Leaf River	15EM015	Mixed Wood Shield	46.424197	-94.897796	5	2
3	West Branch Baptism River	15EM017	Mixed Wood Shield	47.536336	-91.316632	7	2
4	Knife River	15EM027	Mixed Wood Shield	45.94708	-93.321934	6	6
5	Red Lake River	15EM032	Temperate Prairies	47.890496	-96.944903	6	4
6	Pike River	15EM037	Mixed Wood Shield	47.702987	-92.316834	6	1
7	South Fork Whitewater River	15EM038	Mixed Wood Shield	43.978064	-92.173732	7	17
8	Boiling Spring Creek	15EM046	Temperate Prairies	44.651513	-95.371109	7	15
9	Perch Creek	15EM067	Temperate Prairies	43.853553	-94.462972	6	5
10	Tributary to Zumbro River, North Fork Rice	15EM070	Temperate Prairies	44.233771	-93.097111	7	9
C	Control	Control	Control			7	0

Table S2-3. ***Data flags in chemical analytics.*** *This table has to be used to interpret the supplement data file* FlaggedChemicalInput.xlsx. *All measurements with data flag 'B' and 'U' were set to* 0*ng/L in this study. (Data flags had been defined for the preliminary study by Ferrey et al. [2017] and had been provided for this study by collaboration partner Dalma Martinović-Weigelt. For more details, see [Ferrey et al. 2017])*

Flag	Definition
B	Analyte found in the sample and the associated laboratory blank
U	Analyte not detected at reporting limit
K	Peak detected but did not meet quantification criteria, the result reported represents the estimated maximum possible concentration
N	Authentic recovery is not within method/contract control limits
TIC	Compound identity and concentration are estimated
V	Surrogate recovery is not within method/contract control limits
H	Analyte concentration is estimated
NQ	Data is not quantifiable
T	The result was recalculated against alternate labeled compound(s) or internal standard
MAX	Analyte concentration is an estimated maximum value
D	Dilution data

Bioanalytical assessment for endocrine activity. Next to chemical analyses, our collaborators performed an endocrine disruption assessment and provided further information. ToxCast predicted estrogenic targets were identified in two out of five streams with *in-vitro* estrogenic activity considered with EE2-equivalent concentrations (see supplemental figure S2-5). The exposure-pattern clustering identified Naproxen, Ciprofloxacin and 5,6-Dimethyl-Benzotriazole as estrogen activity related (EE2 in compound group 4), although not describing the total *in-vitro* assessed endocrine effect. All three compounds were detected in only one stream with estrogenic activity (EE2). Nitrates are considered as pollution- and landuse marker. Their exposure pattern represents also an endocrine activity result in this thesis. Nitrate belongs to compound group 1 and the co-correlated compounds potentially reflect an endocrine activity. For example, Bisphenol A and Erythromycin were detected in streams 7 and 10, which showed estrogen activity across all three test settings (see supplemental table S2-5). One stream showed an androgenic *in-vitro* activity (see supplemental table S2-5). The exposure patterns of Lincomycin, Sulfamerazine, Stream9 and CG8 are identical and may highlight androgenic-related exposure effects.

Table S2-4. ***List of 29 detected chemical compounds*** *with CAS-identifier, estimated range of fish toxicity, the structural formula, and additional information to chemical classification.*

Chemicals	CAS	ToxArea (Fish) mg/L	Formula	Class
5,6-Dimethyl-Benzotriazole	1354973-50-4	10	C8H9N3	Benzotriazole
Sulfamerazine	127-79-7	500	C11H12N4O2S	Antibacterial Agent
Lincomycin	154-21-2	1000	C18H34N2O6S	Antibacterial agent
Iopamidol	60166-93-0	5000	C17H22I3N3O8	Contrast Agent
Nitrate		1	NO3	NA
Metformin	657-24-9	500	C4H11N5	Biguanide
4-Methyl-Benzotriazole	29878-31-7	50	C7H7N3	Benzotriazole
Amitriptyline	50-48-6	1	C20H23N	Antidepressant
Ethinyl Estradiol	57-63-6	1	C20H24)2	Estrogen
Naproxen	22204-53-1	50	C14H14O3	Non-Steroidal anti-inflammatory drug
Cotinine	486-56-6	100	C10H12N2O	Alkaloid
Triclosan	3380-34-5	0.1	C12H7Cl3O2	Polychloro phenoxy phenol
5-Methyl-benzotriazole	136-85-6	10	C7H7N3	Benzotriazole
Caffeine	58-08-2	500	C8H10N4O2	Methylxanthine alkaloid
DEET	134-62-3	10	C4H13N3	Insect repellent
Diazepam	439-14-5	1	C16H13ClN2O	Anxiolytic Agent
Triamterene	396-01-0	10	C12H11N7	Potassium sparing Diuretic
Sertraline	79617-96-2	0.1	C17H17Cl2N	Antidepressiva
Benzothiazole	95-16-9	10	C7H5NS	Benzothiazole
2-Hydroxy-Benzothiazole	934-34-9	1	C7H5NOS	Benzothiazole
Trimethoprim	738-70-5	100	C14H18N4O3	Antibacterial Agent
Benzotriazole	95-14-7	10	C6H5N3	Benzotriazole
Bisphenol A	80-05-7	1	C15H1602	Xenoestrogen
2-Amino-Benzothiazole	136-95-8	1	C7H6N2S	Benzothiazole
Erythromycin-H20	114-07-8	50	C37H67NO13	Antibacterial Agent
Sulfamethoxazole	723-46-6	100	C19H11N3)3S	Antibacterial agent
Meprobamate	57-53-4	10	C9H18N2O4	Antidepressant
Carbamazepine	298-46-4	10	C15H12N2O	Antidepressant
Ciprofloxacin	85721-33-1	100	C17H18FN3O3	Antibacterial agent

Table S2-5. ***Estrogen and androgen receptor-related activities in water samples of ten selected streams.*** *Low levels of estrogen activity were detected in surface water samples (in vitro cell assays, below 0.25 ng EE2 equivalents/L) (see second row). Pathway enrichment results were associated with estrogen receptor signalling (fourth column). Enrichment results were negative for one site but vitellogenin upregulation was determined (*). (This overview of endocrine activities was measured in a preliminary study by Ferrey et al. [2017] and provided for this study by collaboration partner Dalma Martinović-Weigelt. For more details, see [Ferrey et al. 2017])*

Site	Estrogen (ToxCast)	Estrogen (*in-vitro*)	Estrogen (Microarray)	Androgen (*in-vitro*)
1				
2			*	
3		✓	✓	
4		✓	✓	
5				
6				
7	✓	✓	✓	
8		✓	✓	
9			✓	✓
10	✓	✓	✓	

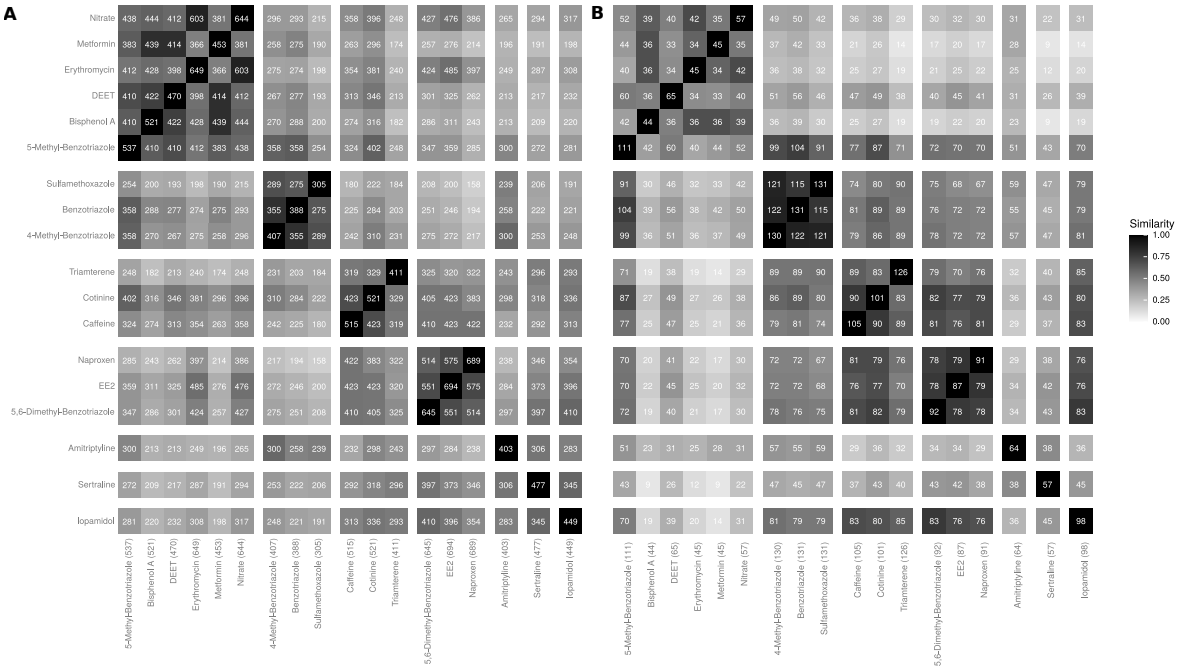

Figure S2-1. ***Pairwise Jaccard similarity and overlap of DEA results for single compound exposure scenarios.*** *Here, only single compound scenarios were compared when exposure patterns have at least two values unequal zero (multiple detected compounds). Thus eighteen compounds are represented in columns and rows. The chemicals are assigned to compound groups. In both cases, DEA results have, in any case, any overlap. The similarity degrees of DEA results support the clustering to compound groups, but shows some exceptions (e.g. 5-Methyl-Benzotriazole or Cotinine).* ***A)*** *The overlap of significantly differentially expressed genes.* ***B)*** *The overlap of enriched terms.*

Figure S2-2. ***Module-trait-correlations across all*** 37 ***considered exposure traits (X) and*** 21 ***modules (Y).*** *Black framed tiles represent significant biweight mid-correlations. The color code represents the direction and the strength of correlation. The exposure traits are grouped into four exposure scenario groups (general treatment, stream-wise, single compound and compound group). The module-trait-correlations were considered significant, if* $|bicor| \geq 0.3$ *and* $p_{bicor} \leq 0.05$.

Table S2-6. ***Overlap of association rule mined chemical-gene interactions to CTD and STITCH.*** *The AR results with single compound exposure scenario (N(tot)) overlapped to the chemical-wise gene sets in STITCH (N(S)) and CTD (N(C)). The results are cumulatively sorted by overlap to CTD (U(C)) and STITCH (U(S)). The endocrine compounds Bisphenol A and EE2 are covered the best considering the relative overlap (%()) to CTD.*

Scenario	Compound	N(tot)	N(C)	N(S)	U(C)	U(S)	%(C)	%(S)
Single	**Bisphenol A**	211	6678	291	115	4	**54.5**	1.9
	Carbamazepine	525	902	343	38	13	7.2	2.5
	EE2	91	3967	624	32	8	**35.2**	8.8
	Triclosan	399	631	202	25	15	6.3	3.8
	Sulfamethoxazole	266	7	204	0	38	0.0	14.3
	Cotinine	174	14	405	1	27	0.6	15.5
	Triamterene	180	4	153	0	23	0.0	12.8
	Caffeine	164	281	1008	7	8	4.3	4.9
	Metformin	164	256	1154	7	7	4.3	4.3
	Diazepam	399	27	480	2	12	0.5	3.0
	Ciprofloxacin	500	23	512	1	12	0.2	2.4
	Benzotriazole	266	1	151	0	0.0	13	4.9
	Naproxen	216	45	342	0	9	0.0	4.2
	Trimethoprim	422	4	252	0	9	0.0	2.1
	Benzothiazole	422	3	131	0	7	0.0	1.7
	Sertraline	159	64	0	5	0	3.1	0.0
	Amitriptyline	155	68	0	4	0	2.6	0.0
	DEET	164	83	111	0	2	0.0	1.2
	Erythromycin	211	28	370	1	0	0.5	0.0

Table S2-7. ***Overlap of association rule mined compound group-gene interactions to CTD and STITCH.*** *The AR results with compound group exposure scenario (N(tot)) overlapped to the chemical-wise gene sets in STITCH (N(S)) and CTD (N(C)). The results are cumulatively sorted by overlap to CTD (U(C)) and STITCH (U(S)). The endocrine compounds Bisphenol A and EE2 are covered the best considering the relative overlap (%()) to CTD.*

Scenario	Compound	N(tot)	N(C)	N(S)	U(C)	U(S)	%(C)	%(S)
CG	**Bisphenol A**	422	6678	291	249	4	**59.0**	0.9
	EE2	500	3967	624	175	25	**35.0**	5.0
	Carbamazepine	525	902	343	38	13	7.2	2.5
	Sulfamethoxazole	525	7	204	1	46	0.2	8.8
	Triclosan	399	631	202	25	15	6.3	3.8
	Caffeine	399	281	1008	17	23	4.3	5.8
	Metformin	422	256	1154	12	25	2.8	5.9
	Naproxen	500	45	342	1	17	0.2	3.4
	Benzotriazole	525	1	151	0	17	0.0	3.2
	Cotinine	399	14	405	2	15	0.5	3.8
	Diazepam	399	27	480	2	12	0.5	3.0
	Ciprofloxacin	500	23	512	1	12	0.2	2.4
	Trimethoprim	422	4	252	0	9	0.0	2.1
	Benzothiazole	422	3	131	0	7	0.0	1.7
	Erythromycin	422	28	370	2	3	0.5	0.7
	Amitriptyline	155	68	0	4	0	2.6	0.0
	DEET	422	83	111	2	2	0.5	0.5
	Triamterene	399	4	153	1	1	0.3	0.3

Table S2-8. ***Overlap of chemical-gene interactions to CTD and STITCH applying differential gene expression analysis.*** *The DEA results with single compound exposure scenario (N(tot)) overlapped to the chemical-wise gene sets in STITCH (N(S)) and CTD (N(C)). The results are cumulatively sorted by overlap to CTD (U(C)) and STITCH (U(S)). The endocrine compounds Bisphenol A and EE2 are covered the best considering the relative overlap (%()) to CTD.*

Scenario	Compound	N(tot)	N(C)	N(S)	U(C)	U(S)	%(C)	%(S)
Single	**Bisphenol A**	521	6678	291	318	10	**61.0**	1.9
	EE2	694	3967	624	248	21	**35.7**	3.0
	Caffeine	515	281	1008	21	17	4.1	3.3
	Metformin	453	256	1154	11	25	2.4	5.5
	Triclosan	423	631	202	23	3	5.4	0.7
	Carbamazepine	184	902	343	21	3	11.4	1.6
	Erythromycin	649	28	370	2	12	0.3	1.8
	Naproxen	689	45	342	3	10	0.4	1.5
	Trimethoprim	425	4	252	0	10	0.0	2.4
	Cotinine	521	14	405	0	8	0.0	1.5
	Sulfamethoxazole	305	7	204	0	7	0.0	2.3
	Diazepam	423	27	480	1	6	0.2	1.4
	Ciprofloxacin	495	23	512	0	5	0.0	1.0
	Sertraline	477	64	0	5	0	1.0	0.0
	DEET	470	83	111	2	3	0.4	0.6
	Benzotriazole	388	1	151	0	4	0.0	1.0
	Triamterene	411	4	153	1	2	0.2	0.5
	Benzothiazole	425	3	131	0	2	0.0	0.5
	Amitriptyline	403	68	0	1	0	0.2	0.0
	Iopamidol	449	10	18	0	1	0.0	0.2

Table S2-9. ***Overlap of compound group-gene interactions to CTD and STITCH applying differential gene expression analysis.*** *The DEA results with compound group exposure scenario (N(tot)) overlapped to the chemical-wise gene sets in STITCH (N(S)) and CTD (N(C)). The results are cumulatively sorted by overlap to CTD (U(C)) and STITCH (U(S)). The endocrine compounds Bisphenol A and EE2 are covered the best considering the relative overlap (%()) to CTD.*

Scenario	Compound	N(tot)	N(C)	N(S)	U(C)	U(S)	%(C)	%(S)
CG	**EE2**	640	3967	624	219	19	**34.2**	3.0
	Bisphenol A	239	6678	291	147	5	**61.5**	2.1
	Metformin	239	256	1154	5	13	2.1	5.4
	Ciprofloxacin	640	23	512	0	10	0.0	1.6
	Naproxen	640	45	342	2	5	0.3	0.8
	Triclosan	83	631	202	6	1	7.2	1.2
	Amitriptyline	800	68	0	4	0	0.5	0.0
	Caffeine	83	281	1008	2	2	2.4	2.4
	Trimethoprim	239	4	252	0	3	0.0	1.3
	Erythromycin	239	28	370	0	3	0.0	1.3
	Iopamidol	632	10	18	0	3	0.0	0.5
	DEET	239	83	111	0	2	0.0	0.8

Table S2-10. ***Overlap of chemical-gene interactions to CTD and STITCH applying weighted gene correlation network analysis.*** *The WGCNA results (N(tot)) with single compound (top) and compound group (bottom) exposure scenario resulted in compound wise gene sets. These overlapped to compound wise gene sets in STITCH (N(S)) and CTD (N(C)). The results are cumulatively sorted by overlap to CTD (U(C)) and STITCH (U(S)). The endocrine compounds Bisphenol A and EE2 are covered the best considering the relative overlap (%()) to CTD.*

Scenario	Compound	N(tot)	N(C)	N(S)	U(C)	U(S)	%(C)	%(S)
Single	**EE2**	2178	3967	624	680	50	**31.2**	2.3
	Naproxen	2178	45	342	7	34	0.3	1.6
	Bisphenol A	37	6678	291	21	2	**56.8**	5.4
	Metformin	37	256	1154	4	3	10.8	8.1
	Erythromycin	176	28	370	1	6	0.6	3.4
	Benzotriazole	37	1	151	0	3	0.0	8.1
CG	**EE2**	2178	3967	624	680	50	**31.2**	2.3
	Ciprofloxacin	2178	23	512	4	45	0.2	2.1
	Naproxen	2178	45	342	7	34	0.3	1.6
	Sertraline	1086	64	0	5	0	0.5	0.0
	Iopamidol	2504	10	18	1	1	0.0	0.0

Chapter S3

Supplement Chapter 4

S3.1 Hyperparameter tuning results

We examined the improvement to the initial model (see figure 4.5). Therefore, we performed a training run for the best-performing word embedding model (with embedding vector size $n = 100$) according to a `RandomSearch` with `kerasTuner` across one hundred hyperparameter recombinations (see table S3-1). The training and validation curves and the confusion matrix for the test set are shown in supplemental figure S3-1.

A similar examination for an LSTM model with $n = 100$ neurons (see figure 4.6) was done. We determined the parameter setting for LSTM model (with $n = 100$ neurons) by determining the best-performing model in a`RandomSearch` with `kerasTuner` across one hundred hyperparameter combinations (see table S3-2). The outcome of the training run with this parameter setting is shown in supplemental figure S3-2.

Table S3-1. ***Hyperparameter tuning for word embedding model.*** *Applying* `kerastuner` 100 *combinations of variable parameters were tested. The list of variable instances is shown in the second column. The model with the lowest validation loss was considered the best model, and its respective parameters are shown in the third column.*

Parameter	Variable instances	Best model
WE L2-regularization	[0,1e-6, 1e-8]	0
Activation function after WE	[True, False]	True
Activation function	[ReLU, tanh]	tanh
Dropout dense layers	[0,0.2,0.4,0.6,0.8]	0.4
Number dropouts	[0,1,2]	2
Number dense layers	[3,4,5]	5

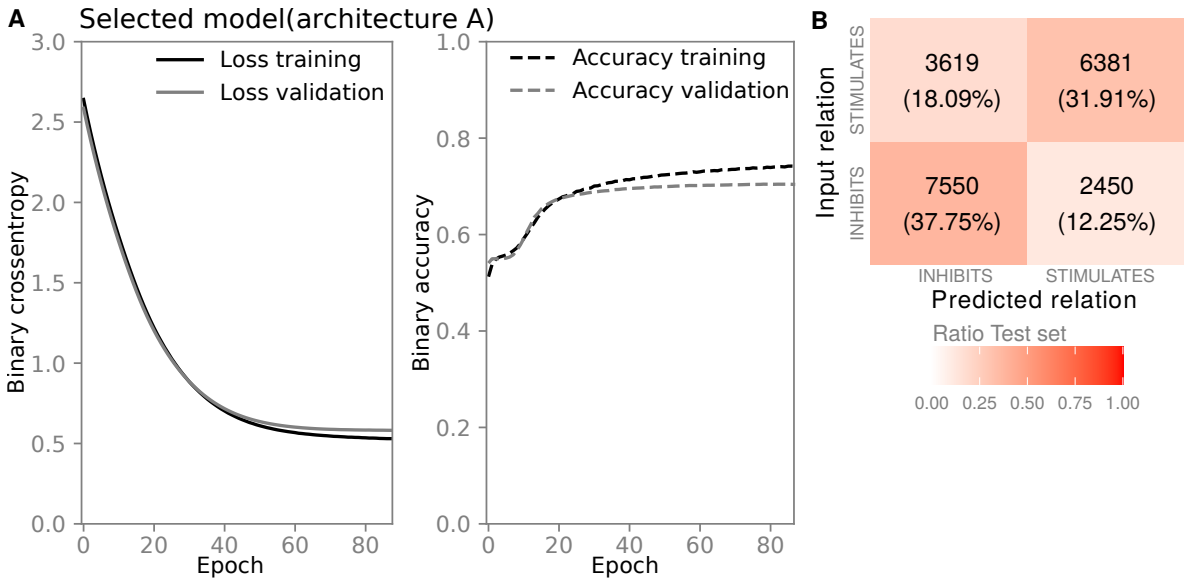

Figure S3-1. ***Model performance of the selected word embedding model after hyperparameter tuning.*** *The deep learning model was trained with UMLS-annotated chemical-biomolecule-pairs retrieved from SemMedDB to predict their toxicogenomic relationship. The model has an initial word embedding layer followed by five dense layers, which decrease in size. The training performance across epochs was determined for a validation set with approximately 71T relations (20% of training).* ***A:*** *The binary accuracy (dashed line) and binary cross-entropy (solid line) for training and validation. The validation loss decreased and finally converged to* 0.582*. The maximal validation accuracy was valued at* 0.70*.* ***B:*** *The confusion matrix presents the prediction results for the test set* I_E *(n=20T) and resulting in a true-positive rate of* 0.697*. This model is marginally better than the original word embedding model with vector size* $n = 100$ *(see figure 4.5).*

Table S3-2. ***Hyperparameter tuning for model with LSTM.*** *The number of units for the output vectors of LSTM and time distributed dense layer were set to* 100*. The* `RandomSearch` *function of* `kerastuner` *tested* 100 *randomly selected combinations of parameters. The best recombination is shown in the last column.*

Parameter	Variable instances	Best model
LSTM L2-regularization	[0, 1e-2, 1e-6]	0
Bidirectional LSTM	[True, False]	False
unroll LSTM	[True,False]	True
Dropout	[0,0.2,0.4,0.6,0.8]	0.4
len(Time distributed)	[100, 50]	100

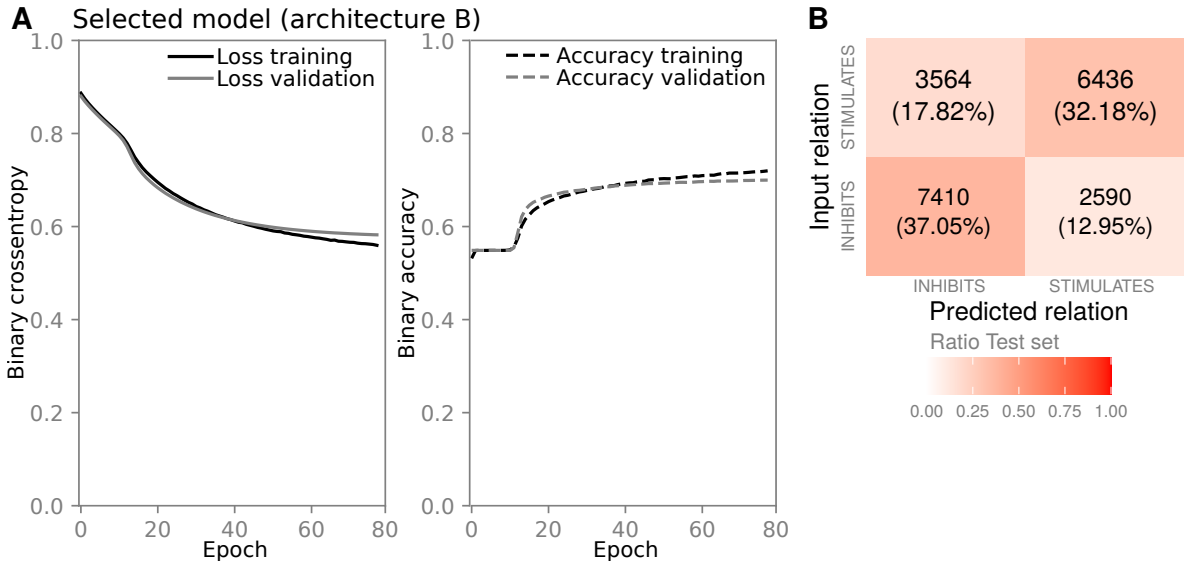

Figure S3-2. ***Training performance of the selected model with LSTM layer*** *The deep learning model was trained with UMLS-annotated chemical-biomolecule-pairs retrieved from SemMedDB to predict their toxicogenomic relationship. The model has an initial word embedding layer. The following LSTM layer and time distributed dense layer are set up with two neurons each, equal to the sequence length. Then, five dense layers, which decrease in size, are trained to classify the relationship type from the hidden sequence information.* ***A:*** *The binary accuracy (dashed line) and binary cross-entropy (solid line) for training and validation. We determined the training performance across epochs for a validation set with approximately* $71T$ *relations (*20% *of training). The binary accuracy (dashed line) and binary cross-entropy (solid line) were tracked for training and validation. The loss decreased over time and reached a saturated state at approximately* 0.583*. The maximal accuracy was valued at* 0.70*.* ***B:*** *The confusion matrix presents the prediction results for the test set* I_E *(n=20T) and resulting in a true-positive rate of* 0.692*. This model is marginally better than the original word embedding model with vector size* $n = 100$ *(see figure 4.5).*

Hyperparamer tuning of model with LSTM. LSTM contains many parameters that are potentially helpful when considering such complex input data (based on a vocabulary of more than 40 000 semantic concepts). However, the performance improvement was marginal after adapting the model architecture B based on hyperparameter tuning when considering loss and accuracy. Although we considered different parameter settings for the application of LSTM and investigated the influence of the LSTM neuron output length, this study is far from an extensive parameter analysis. Still, the decisions for chosen parameters should be

described shortly.
The dropout helps to improve training processes and to overcome overfitting in training. It was seen as a standard parameter to test. What might be surprising for the reader is the range of dropout values, also considering dropouts for more than half of the input vector (see table S3-2). However, as in the early stages of the model training learning curves had a relatively fast loss drop in less than ten epochs and reached a saturation plateau relatively fast, the chances of overfitting were considered as high with the given input. Consequently, the dropout was investigated in a broad range, also considering potentially aggressive values.

As already mentioned, there are various deep learning knowledge representation alternatives possible. Alone for the case of recurrent neural networks, various alternative architectures may be considerable. For example, bidirectional recurrent neural networks were frequently used to capture biomedical knowledge from text [Zhao et al. 2019, Lyu et al. 2017, Peng et al. 2018]. Consequently, this possibility was also considered within the hyperparameter tuning of LSTM in the present investigation, albeit not applied in the final model architecture B. A bidirectional LSTM might be expected to improve for the short sequence length of two marginally. Applying a bidirectional layer might influence the horizontally augmented data more substantially. However, it was decided to determine the same architecture when considering all different inputs. The effect of bidirectionality was not tested for augmented input as this option was neglected in hyperparameter tuning beforehand. Due to the potential early rejection of bidirectionality, bidirectional LSTMs for elongated sequences should be considered also in future investigations.

S3.2 Functional enrichment with predicted chemical-gene interactions and CTD reference pathway genesets

We need to filter chemical-gene interactions to perform a functional enrichment based on their prediction probability. Therefore, a filtering threshold f was determined first. We determined the true-positive rate (TPR) for the known chemical-gene interactions in T_{C2G} for different values of f to determine a threshold for f (see table S3-3). We expected that false-positive results tend to have a probability distribution closer to 0.5 than the distribution of true-positive results.

As the violin plot for the probabilities of predictions for T_{C2G} in figure S3-3 shows, the probability distributions of the false-positive predictions were visually somewhat similar to those of true-positives. However, the left red violin plot — the true-positive *INHIBITS*

predictions — tapers less than the right red violin plot towards lower prediction values. We observed a slightly higher ratio of lower prediction probabilities (below 0.2) for true-positive *INHIBITS* results. Equivalently, true-positive *STIMULATES* predictions had higher ratios for the higher prediction values (above 0.8) (see blue violin plots).

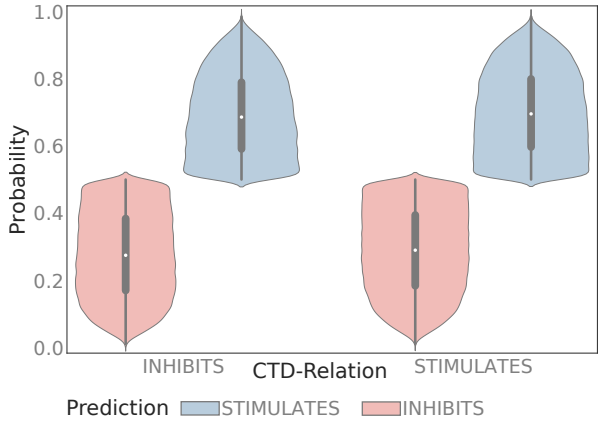

Figure S3-3. ***Relationship grouped violin plots of prediction probability for chemical-gene interactions from*** T_{C2G}***.***

The value of TPR presents the ratio of the true-positive predictions to the total amount of samples. Whereas the TPR for *STIMULATES* decreased for enlarged values of f, the TPR for *INHIBITS* increased. Although both TPR-trends diverged, the overall TPR increased with increasing f (see table S3-3), albeit the values were still close to 0.5 and not convincingly better than a random guess. Approximately, false-positive results were equally often with higher probabilities as true-positive results. However, a filter threshold of $f = 0.9$ was chosen to apply an example of a functional enrichment to chemical semantic concepts.

We performed a **chemical wise evaluation** for the ten most frequent chemical semantic concepts in T_{C2G} applying an overrepresentation analysis (ORA). This study regarded the pathway level only. However figure S3-4 shows that toxicological evaluation with CTD is technically possible on chemical and disease levels.

We compared the distribution of prediction probabilities based on prediction classes to T_{C2G} and chose a filter threshold $f = 0.9$. As the prediction was a binary classification, we selected all predictions for relationship type *INHIBITS* with a probability $p \geq f$ and *STIMULATES* with $p \leq (1 - f)$ as *potentially biologically meaningful relations*. A selected chemical-biomolecule relation $C \rightarrow^{Relation} B$ could be interpreted as a toxicologically regulated biomolecule B after exposure to chemical C. We predicted the relationship type across all biomolecular semantic concepts represented in T_{C2G} and I for each of the ten chemical semantic concepts. The chemical wise predictions were grouped in *STIMULATED*, *INHIB-*

Table S3-3. ***Influence of probability threshold f on true-positive rate (TPR).*** *Different prediction probability thresholds (f) were considered, and the true-positive rate (TPR) was calculated for each relationship type (TPR_I and TPR_S) and over the entire set T_{C2G} (TPR_{All}). Whereas the TPR_I increased with higher thresholds f, the TPR_S decreased. Nevertheless, the TPR_{All} increased with increasing f.*

f	TPR_I	TPR_S	TPR_{All}
0.5	0.578	0.475	0.527
0.6	0.598	0.468	0.533
0.7	0.623	0.459	0.541
0.8	0.662	0.458	0.551
0.9	0.738	0.390	0.564

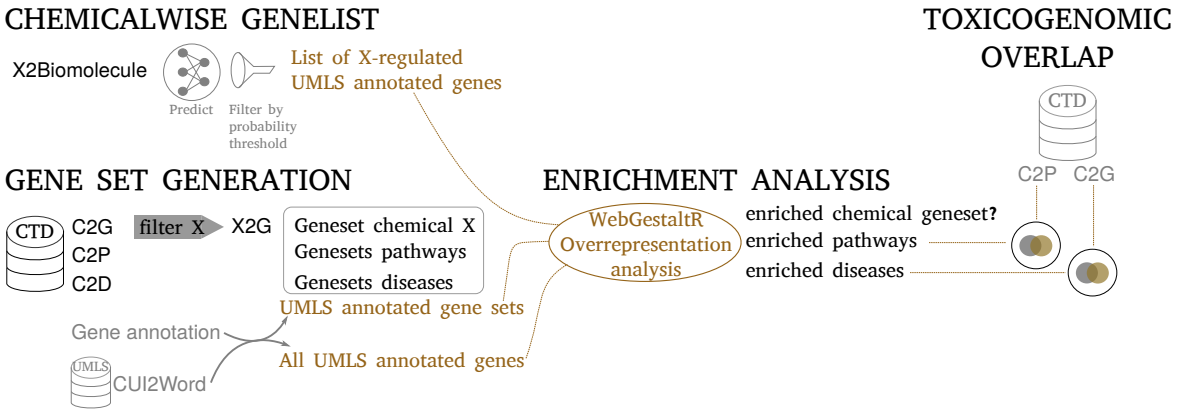

Figure S3-4. ***Toxicological evaluation applying functional enrichment on CTD reference sets.*** *For one chemical CUI X, the relationship type can be predicted to each gene CUI the model is trained with. When applying a filter based on the prediction probabilities, we retrieve a set of predicted genes that is expected to be regulated due to exposure to X. Furthermore, CTD information of chemical-gene (T_{C2G}), gene-pathway (T_{C2P}) and gene-disease associations (T_{C2D}) can be transformed to UMLS-annotated genesets and applied in an overrepresentation analysis with* `webGestaltR`*. Significantly enriched genesets ($FDR \leq 0.05$) may be associations of X to the chemical X itself or pathways or diseases. With help of T_{C2P} and T_{C2D} the overlaps of enrichment results to CTD can be determined and interpreted. In this study, functional enrichment analysis was applied on pathway level only.*

ITED or *UNREGULATED* with the help of f. In consequence, we determined chemical wise lists of regulated (*STIMULATED* or *INHIBITED*) genes.

The list of the ten most frequent CUIs in T_{C2G} comprised three chemical compounds: Bisphenol A (three CUIs), Benzo(a)pyrene (one CUI) and Valproic acid (six CUIs). For all three chemical compounds, lists of regulated genes with model $A^{\star}$ were predicted and filtered with threshold filter $f = 0.9$. Out of the UMLS-annotated background with 6779 genes, the respective genesets comprised 219, 321, and 1063 as *REGULATED* predicted genes to Benzo(a)pyrene, Bisphenol A and Valproic acid, respectively. The functional enrichment across 2283 UMLS-annotated pathways retrieved from T_{G2P} resulted in 4, 7 and 2 overrepresented pathways (see table S3-4).

Table S3-4. ***Chemical wise enrichment results for genes predicted as regulated.*** *For the most frequent represented CUIs in T_{C2G}, regulated genes across all genes available in the model were determined with the chosen threshold of $f = 0.9$ (see table S3-3). A unique gene set was generated for CUIs, which represent the same chemical. An overrepresentation analysis was performed applying* `webGestaltR`*, and significantly enriched terms ($FDR \leq 0.05$) of biological pathways were determined. Most of the chemical associated gene sets were represented in T_{C2P}.*

Chemical	Pathway	FDR	inCTD
Bisphenol A	Nuclear receptor transcription pathway	0.0376	✓
	Intrinsic pathway for apoptosis	0.0376	✓
	Synthesis of PG	0.0376	
	Activation of NOXA and translocation to mitochondria	0.0376	
Benzo(a)pyrene	Synthesis of PG	0.0153	
	Intrinsic pathway for apoptosis	0.0153	✓
	Nuclear receptor transcription pathway	0.0153	✓
	Apoptosis multiple species	0.0153	✓
	Longevity regulation pathway	0.0451	✓
	cAMP signaling pathway	0.0451	✓
	Mitochondrial biogenesis	0.0451	✓
Valproic acid	Biological oxidations	0.0027	✓
	Steroid hormone biosynthesis	0.0027	✓

Although the number of enriched terms was small and marginal to the number of associated pathways in T_{C2P} (Benzo(a)pyrene: 1629, Bisphenol A: 1668, Valproic acid: 1551), 2, 6, and

2 enriched pathways were covered in T_{C2P}. Thus, empirical measurements supported our prediction-based outcomes.

S3.3 Reduction of learning rate in a model with large word embedding vectors.

The training in figure 4.5 had a smaller number of epochs, the larger the length of the word embedding vector (n) was. However, the loss was slightly better for the largest n with *binary crossentropy* = 0.601. In comparison, the loss for the chosen vector size $n = 100$ valued *binary crossentropy* = 0.636. We chose the smallest word embedding size when considering an unseen test set I_E, as the accuracy was the largest.
To proof, whether the rapid drop in binary cross-entropy for word embeddings with large n might be not inflicted by the selected learning rate, the word embedding model with $n = 5000$ was trained with $\bigcup_{i=2}^{5} I_{Ti}$ and validated with I_{T1}. Nothing was changed on the word embedding model architecture but the learning rate (reduced from $1e-5$ to $1e-6$). The results of the model training are shown in Figure S3-5.

As can be seen in the left plot, the binary cross-entropy decreased steadily and rather fast within the first 100 epochs for both the training and validation set. After that the difference in loss per epoch became smaller and the loss curve reached a plateau until stopping after 214 epochs. The minimal loss valued *binary crossentropy* = 0.614. Also, the accuracy increased in the training and validation data. In the first 50 epochs, the training did not result in any accuracy improvement, albeit loss decreased. The validation accuracy diverged to 0.693.
Compared to the word embedding model with $n = 5000$ in section 4.2.2, the binary cross-entropy was not smaller with a lower learning rate. Furthermore, the accuracy for the unseen test data valued 0.685 and was smaller than for the models with higher learning rates ($n =$ 5000:0.696, $n =$ 100:0.697). As expected, the choice of the learning rate influenced the training duration in terms of epochs. However, the change in the learning rate did not improve the model performance in training or evaluation.

S3.4 Horizontal augmentation without tail-padding

To proof, whether the order has an influence, the model performance for training with horizontally augmented input was performed with zero-masked tokens but without tail-padding to conserve the semantically identical order across all samples. The input preparation remained identical as before for horizontally augmented inputs, except that sequences were not

tail-padded where parental or grandparental terms were missing. Consequently, when samples were integer encoded, the zero-values are not at the end but on its semantically specified position (<Subject, Subject Parent, Subject Grandparent, Object, Object Parent, Object Grandparent>).

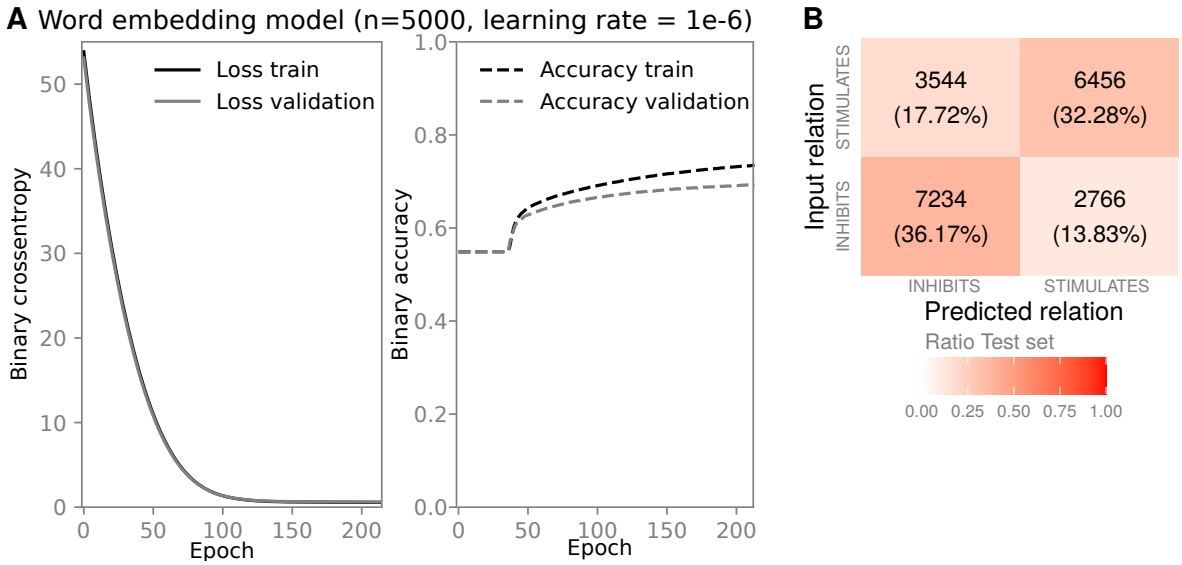

Figure S3-5. ***Training and validation curves of word embedding model (n=5000) and lower learning rate*** *($\alpha = 1e-6$).* *The deep learning model was trained with UMLS-annotated chemical-biomolecule-pairs retrieved from SemMedDB to predict their toxicogenomic relationship. The model has a word embedding layer, an LSTM layer, a time distributed dense layer and five subsequent dense layers decreasing in size.* ***A:*** *The binary accuracy (dashed line) and binary cross-entropy (solid line) for training and validation. We determined the training performance across epochs for a validation set with approximately $71T$ relations (20% of training). The loss decreased over time and reached a saturated state at approximately* 0.6. *The maximal accuracy was valued at* 0.68. ***B:*** *The confusion matrix presents the prediction results for the test set $I\ E$ ($n = 20T$) resulting in a true-positive rate of* 0.684.

The input was considered for a 5-fold cross-validated model training with model architecture A and with B. Supplemental figure S3-6 shows the training curves during training for both model architectures and the confusion matrices considering the unseen test data set.

The validation loss decreased rather quickly within the first five epochs and converged to a value at minimally 0.69. The binary accuracy curve for validation data increased within the first five epochs and converged to 0.65.

The evaluation with unseen test data resulted in true-positive rates of 65% for both model

architectures. The values of relationship type-specific performance measures were similar across models (see table S3-5) and compared to the previous experiments with tail-padded sequences (see table 4.4 A^H and B^H). Consequently, preserving a semantically meaningful order of the horizontally augmented input did not significantly affect the model performance.

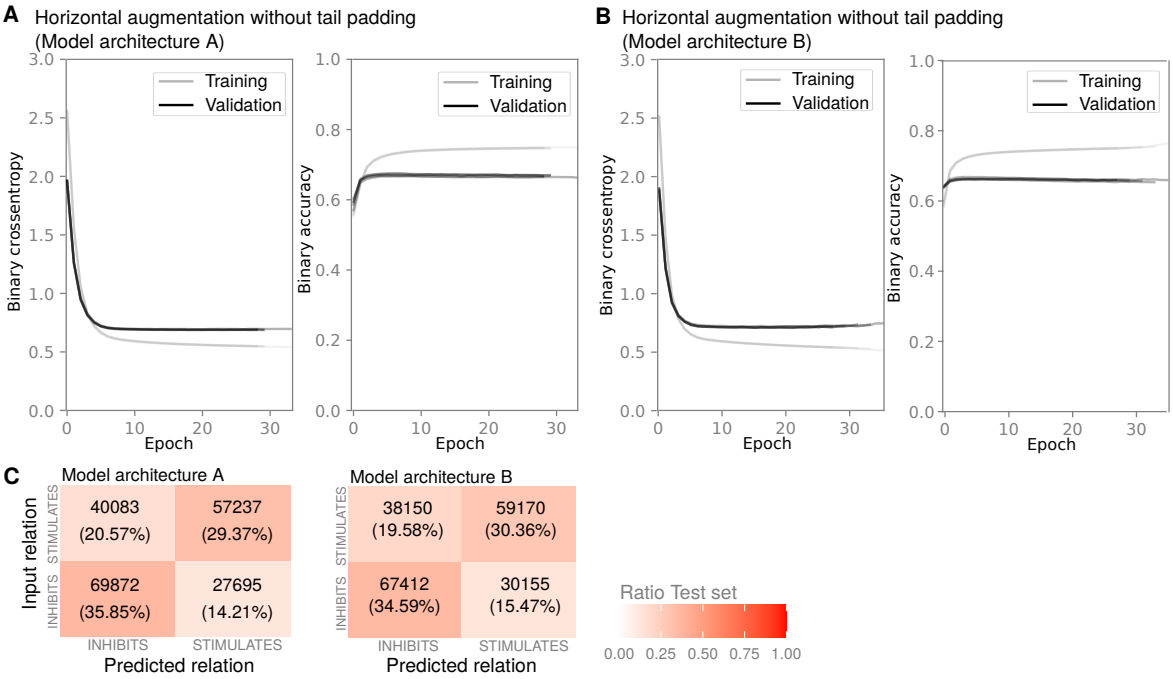

Figure S3-6. ***Performance of model trained with horizontally augmented, but not tail padded input.*** *The SemMedDB input set I was horizontally augmented to I_E^H considering the UMLS semantic network. Some subject and object concepts have no semantic parents or grandparents and were zero-masked. In contrast to the trained models in figure 4.7, the horizontally augmented input was not tail-padded.* ***A:*** *Curves of binary accuracy and loss for 5-fold cross-validated training with model architecture A.* ***B:*** *Curves of binary accuracy and loss for 5-fold cross-validated training with model architecture B.* ***C:*** *Confusion matrices for not tail-padded test data I_E^H for one trainings fold of model architecture A (top) and B (bottom).*

S3.5 Four-relationship classification

We performed training of deep learning models with SemMedDB chemical-gene interactions considering four predicates (see section 4.2.6). The input data I_4 was considered a four-class classification problem and a binarised classification problem. Furthermore, we trained models with model architecture A and B, respectively. The results of the four 5-fold cross-validations

Table S3-5. ***Test performances of models with not tail-padded horizontally augmented input.*** *Performance measures were calculated based on not tail-padded relations in the test sets (I_E^H). Next to binary accuracy (Acc), the other measures (precision (Prec), recall (Rec) and the F1-score (F1)) were considered per relationship type.*

	Relationship	Acc	Prec	Rec	F1
$A^H_{ordered}$	INHIBITS	0.65	0.73	0.64	0.67
	STIMULATES		0.59	0.69	0.63
$B^H_{ordered}$	INHIBITS	0.65	0.69	0.64	0.66
	STIMULATES		0.61	0.66	0.63

Table S3-6. ***Number of chemical-biomolecule-relations per predicate.*** *The not entirely preprocessed input I (with contradictions) consisted of triplets with twenty-five predicates. The numbers varied extremely and made the dat set very unbalanced considering predicates.*

Predicate	Number of Relations	Predicate	Number of Relations
INTERACTS_WITH	320845	CONVERTS_TO	6245
INHIBITS	250468	TREATS	5442
STIMULATES	215934	PREVENTS	4344
AFFECTS	182786	ASSOCIATED_WITH	4171
COEXISTS_WITH	148957	lower_than	2556
PART_OF	135495	PREDISPOSES	2325
DISRUPTS	130933	same_as	2011
AUGMENTS	106087	PRODUCES	1364
compared_with	49511	USES	1249
CAUSES	44716	LOCATION_OF	285
ADMINISTERED_TO	18498	COMPLICATES	257
higher_than	13964	PROCESS_OF	3
ISA	9067		

are shown in supplemental figure S3-7. In summary, neither considering a categorical nor a binary classification of the expanded input I_4 improved overall performance.

Independent of the classification task and the model architecture, the five folds of a cross-validation training behaved similarly and were visually not distinct (see supplemental figure S3-7 A-D). The validation loss decreased relatively fast in the first twenty epochs and converged to approximately 0.67 after 60 to 120 epochs. The validation accuracies reached maximal values of 0.68. Considering the test set $I_{4,E}$ with 63 287 relations, true-positive rates of 0.683, 0.664, 0.683 and 0.680 were measured for the categorical case with model architecture A and B and binary case with model architecture A and B (see supplemental figure S3-7). Consequently, the previous experiments performed marginally better when considering input I and two predicates.

However, the categorical consideration revealed an exciting characteristic in confusion matrices. Although overall classification performed nearly identical, the model distinguished pharmacogenomic relationship types (*AUGMENTS* and *DISRUPTS*) from substance interactions (*STIMULATES* and *INHIBITS*), with a true-positive rate of 0.994 with both model architectures. Thus, the models recognised the general semantic meanings of the relationship types and respective suitable chemical and biomolecular semantic concepts. Furthermore, the applied binary classification task presented that the models could distinguish negative regulations from positive relations with 0.68% accuracy. Albeit not that accurate as for the interaction types, the predictive model was still able to predict positive or negative directions of chemical-biomolecule interactions.
The very accurate distinction between pharmacogenomic and substance interactive relationships probably originated in the input data itself. The subject concepts overlapped 30% across predicate types (see supplemental figure S3-8) *STIMULATES* and *INHIBITS* share less than 2% of object concepts with *DISRUPTS* and *AUGMENTS*. However, predicate groups shared 37.78% and 13.12% of objects. Consequently, the models have learned to differentiate the relationship types most likely regarding the object concept.

S3.6 Interpreting loss observations for SemMedDB trained models.

We observed a meaningful trend of the validation loss curve. In the beginning, the loss decreased relatively fast. In the second phase, the loss decreases still steadily. However the changes get more minor over training duration. Finally, a plateau was reached, where minimal decreases or even smaller fluctuations were observed. The training and validation

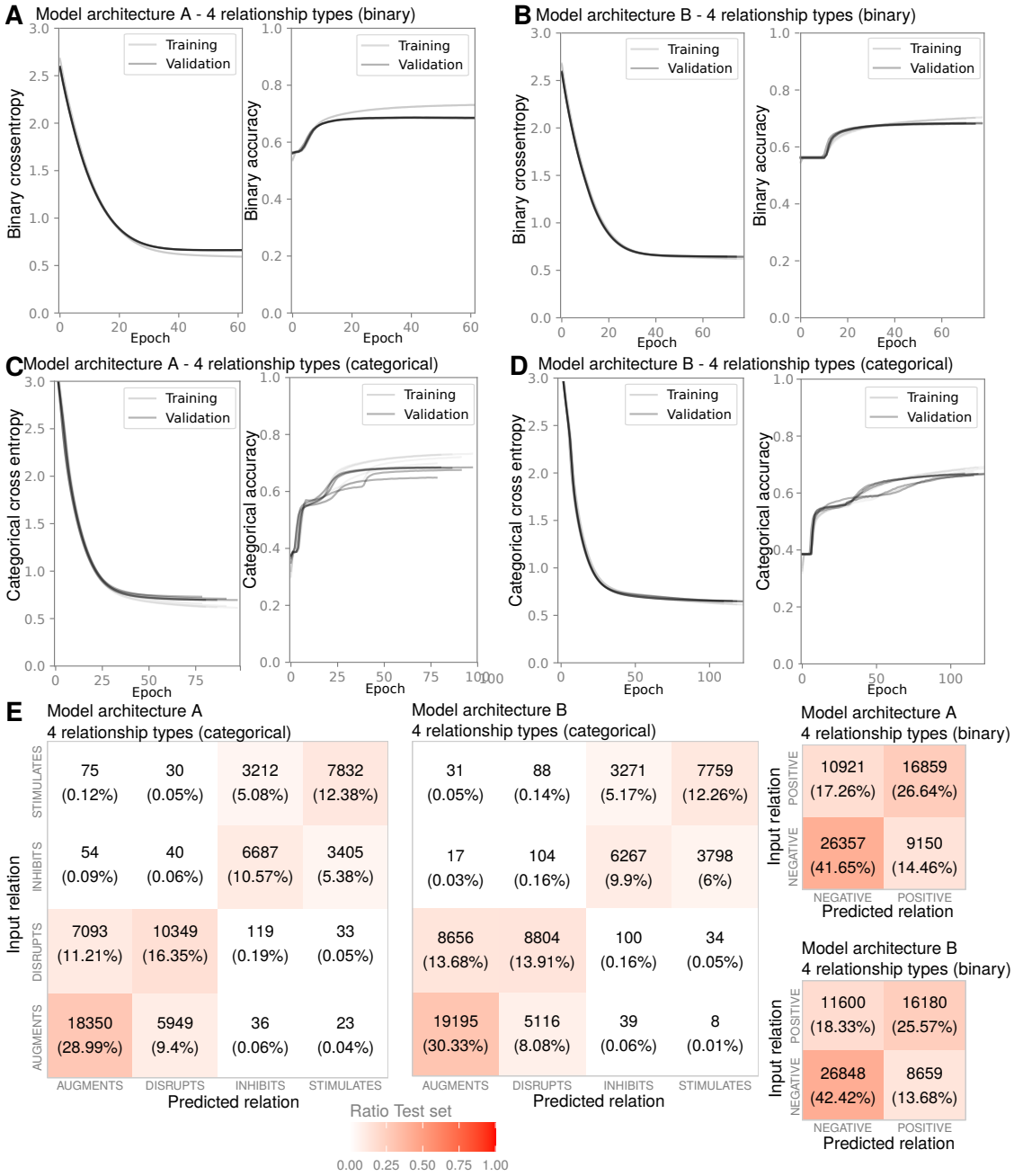

Figure S3-7. ***Model performance considering four relationship types*** *Models were 5-fold cross-validated trained based on architecture A and B with chemical-biomolecule interactions from SemMedDB. A binary (0: NEGATIVE, 1: POSITIVE) or categorical classification task ((INHIBITS, STIMULATES, AUGMENTS and DISRUPTS)) was considered. The accuracy and cross-entropy were tracked for five folds of training and validation (20% of training).* ***A:*** *Binary with architecture A.* ***B:*** *Binary with architecture B.* ***C:*** *Categorical with architecture A.* ***D:*** *Categorical with architecture B.* ***E:*** *Confusion matrices for best fold model from A-D with test set (n=63 287).*

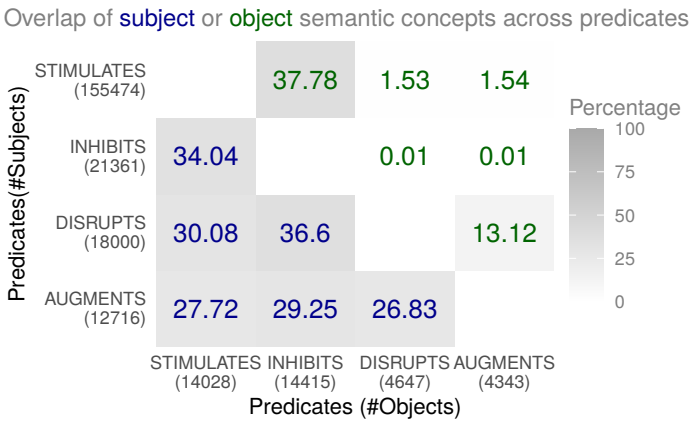

Figure S3-8. ***Concept overlap in input across four predicates.*** *The input data set* I_4 *contains* 632 864 *relations with* 32 527 *chemical and* 24 550 *molecular concepts which are not equally distributed across predicates. The matrix presents the percentage of overlapping subjects (blue) and objects (green). Number of unique subject (y) and object (x) concepts per predicates are given with the axis labels.*

curves performed somewhat similar. In most cases, the validation curves reached a slightly higher loss plateau, which is an expected training behaviour in deep learning. Consequently, the trained models learned with the given input. The as best selected model predicted 70% of relationships correctly.

However, the minimal values of loss for the best model were at values of binary cross-entropy of 0.6. The tracked binary cross-entropy represented the average per-sample loss per batch. Thus, in average the prediction probability for the correct label values $p = 10^{-0.6} = 0.2512$ *. Thus, the probability value calculated from the average loss hints at misleading predictions. According to these estimations, the binary cross-entropy values below 0.3 would indicate a

* The average per-sample-loss per batch can be calculated as following:

$$\overline{binary\ accuracy}_{batch} = -\frac{1}{N}\sum_{i=1}^{N} y_i * log_{10}\hat{y_i} + (1 - y_i) * log_{10}1 - \hat{y_i} \tag{S3.1}$$

In the case of a binary classification the true label has a probability of 1, whereas the predicted value for this label is p, thus the formula can be reduced to:

$$\overline{binary\ accuracy}_{batch} = -\frac{1}{N}\sum_{i=1}^{N} log_{10}p = -log_{10}\bar{p} \tag{S3.2}$$

, where $\bar{p}$ presents the average prediction probability of the true label. In consequence, the average prediction probability can be estimated as:

$$\bar{p} = 10^{-p} \tag{S3.3}$$

trend towards true predictions. As shown in the results, the average loss value underestimates the hard classification of the model, as accuracy values highlighted a true positive rate of 70%. The loss value considered probabilities on a log-scale, and probabilities closer to zero may affect the average more than values closer to one. Thus, few outlying false predictions with probabilities close to zero may lead to higher average loss values. In consequence, albeit the minimally tracked binary accuracy was still rather high with a value of 0.6, it should not be misinterpreted that the model might perform not well. Especially when the validation loss curve showed an expected behaviour in training.

Glossary

AO Adverse outcome. 10, 100

AOP Adverse Outcome Pathway. 9–11, 20, 22, 23, 44, 91, 100, 102, 120, 127, 128, 130, 136, 140, 143, 146–149

AOPwiki An open access database of the recent known and hypothesized adverse outcome pathways. 10, 11, 22, 100, 136, 147

AR Association Rule mining. 16, 32, 36, 57, 58, 69, 72, 74, 75, 81, 92–94, 97, 146, 168, 169

BERT Bidirectional Encoder Representations from Transformers. 26, 101, 136

BioGRID Biological General Repository for Interaction Datasets . 135

BioWordVec Retro-fitted wordembedding with UMLS-vocabularies. 26, 137

CA Concentration addition. 5

cDNA Complementary Deoxyribonucleic Acid. 12, 13, 157, 158

CG Chemical compound groups. 34, 69, 74, 75, 77–84, 86, 87, 89–92, 96

CTD Comparative Toxicogenomic Database. iv, v, vii, viii, III, IV, 20–22, 25, 29, 31, 40, 41, 44, 52–55, 57, 79, 82–86, 89–91, 97–100, 102–104, 107–109, 118–120, 123–127, 133–136, 139, 143–146, 148, 149, 168–172, 177, 179

ctdp1 Carboxy-terminal domain phosphatase subunit1. 88

CUI Concept unique identifier. 42, 44, 53, 54, 105, 107, 180

DE Differentially expressed. 36, 66, 68, 69, 75, 84, 90, 95

DEA Differential gene expression analysis. 18, 32, 35, 55–58, 66, 68, 69, 75–77, 81–84, 86, 87, 90, 92–95, 97, 98, 142, 143, 146, 170, 171

DNA Deoxyribonucleic Acid. 154–157

E2 17-β-estradiol. 7, 152

ECOSAR Ecological Structure-Activity Relationship - predictive class program for aquatic toxicity. 33

ECOTOX Ecotoxicological database curated by the Environmental Policy Agency of the United States of America. 33

EE2 17-α-ethinyl-estradiol. 33, 66, 78, 79, 81–87, 89, 90, 92, 97, 152, 163, 168–172

ES Enrichment score. 40

ET Environmental Toxicology. 2–5, 7–14, 18–22, 25, 27, 29, 136, 140, 143–147, 149

fastText Word embedding approach that interprets word-internal structures (subword embedding). 26

FDR False discovery rate. 36, 67, 70, 89

FHM Fathead minnow - an ecological and aquatic vertebrate model species. 56–58, 61, 65, 79

FI Falsely inhibited predictions. 51

FS Falsely stimulated predictions. 51

GO Gene Ontology. 19

GS Gene significance. 38, 39, 77, 78, 80

GSEA Gene set enrichment analysis. 19, 40, 77, 78, 87, 88, 91, 92, 94

IA Independent action. 5

IQR Interquartile-range. 64

KE Key event of an adverse outcome pathway. 9, 10, 22, 100, 147

KEGG Kyoto Encyclopedia of Genes and Genomes. 19, 40, 41, 57, 68, 147

KER Key event relationship. 10

LBD Literature-based knowledge discovery. 23–25, 28, 127

LC50 Lethal concentration for half of the tested biological systems. 33

LOBO Level of biological organisation. 9, 10, 43, 45, 52, 53, 100, 107, 120, 128, 135–137, 144, 146, 147

log2FC 2-Logarithmic fold change of expression. 40

LSTM Long short term memory. 27, 48, 50, 51, 101, 102, 104, 113–116, 125, 127, 131–134, 137, 174–177

MalaCards Integrated database of human maladies and their annotations, modeled on the architecture and richness of the popular GeneCards database of human genes. 135

MEDLINE Part of PubMed database and contains journal citations and abstracts for biomedical literature.. 25, 101

MeOA Mechanism of (toxic) action. 8, 9, 11

MESH MEdical Subject Headings. 26, 129, 137

MESH-ID Unique term identifier of the MESH ontology. 40, 41

MetaMap UMLS word recognition tool to map biomedical text to semantic concepts. 24, 25

MIE Molecular initiating event. 9, 100, 102

MM Module membership. 38, 39, 77, 78, 80

MOA Mode of (toxic) action. 8–13, 154, 155

MPCA Minnesota Pollution Control Agency. 58

mRNA messenger Ribonucleic Acid. 12, 155–157

MsigDB Molecular Signatures Database (repository of biological signaling pathways). 19

MTC Module-Trait-correlation. 38, 39, 77, 78, 81, 87, 88, 91, 92

ncoa2 Nuclear receptor coactivator 2. 88

ncRNA non-coding Ribonucleic Acid. 156, 157

NLP Natural language processing. 23–27, 127, 139, 147

nr3c1 Nuclear receptor subfamily 3 group C member 1. 88

NSAID Non-steroidal anti-inflammatory agent. 46

OECD Organisation for Economic Cooperation and Development. 10

ORA Overrepresentation analysis. 19, 39, 54, 94, 179

PubMed Free full-text archive of biomedical and life sciences journal literature. 25, 26, 42–44, 91, 101, 102, 105, 129, 137, 139, 147

QSAR Quantitative Structure-Activity Relationship. 33

ReLU Rectified Linear Unifier. 48

RNA Ribonucleic Acid. 13, 154, 156–158

RNA-seq RNA-sequencing. 13, 158

RNN Recurrent neural network. 27, 48, 50

ROS Reactive oxygen species. 154

rRNA Ribosomal Ribonucleic Acid. 156

SemMedDB Semantic MEDLINE Database - repository of semantic predications extracted by SemRep. iv, viii, 24, 25, 29, 31, 42–45, 53, 54, 99, 101–105, 107, 111, 114, 117, 120, 122–137, 139, 143–149, 175, 176, 182, 183, 185

SemRep UMLS-based natural language processing tool that extracts semantic predications. 25, 31, 42, 44, 101, 105, 129, 145, 147

STITCH Search Tool for InTeracting CHemicals. viii, 20–22, 29, 40, 41, 79, 82, 83, 85, 97, 136, 144, 148, 168–172

STRING Search Tool for the Retrieval of Interacting Genes/Proteins. 21

TG-Gates Japanese public toxicogenomics database. 22

TI Truly inhibited predictions. 51

TOM Topological Overlap Measure. 38, 39, 76

ToxCast Toxicity forecaster - data set and predictive models on chemicals of interest to the Environmental Policy Agency of the United States. 20, 22, 100, 136, 148, 163

TPR True positive rate. 177, 178

TransE Translating Embeddings for Modeling Multi-relational Data. 135

tRNA transport Ribonucleic Acid. 156, 157

TS Truly stimulated predictions. 51

TU Toxic unit. 33, 66

UMLS Unified medical language system. 23–27, 31, 42–46, 52–54, 101, 103, 105–107, 109, 111, 114, 118, 126–130, 133–137, 140, 143–148, 175, 176, 179, 180, 182, 183

UMLSBERT A contextualized knowledge representation for the UMLS ontology. 26, 136, 137

WGCNA Weighted Gene Correlation Network Analysis. 17, 37–39, 55, 57, 58, 76, 78, 81, 86–88, 90–94, 98, 142, 143, 146, 172

WikiPathways Repository of biological signaling pathways. 19

List of Figures

1 Transfer of the exposome definition to environmental toxicology. X

1.1 The triangle of environmental toxicology 3

1.2 Classification types for chemical exposures in the environment. 4

1.3 The adverse outcome pathway framework 10

1.4 Overview of transriptomics analyses in environmental toxicology. 15

1.5 Overview network inference analysis 18

1.6 Tasks in natural language processing 25

2.1 UMLS-Annotation of lexical words to levels of biological organisation 45

2.2 Input data augmentation for chemical biomolecule relations 48

2.3 Architectures of neural networks. 51

2.4 Deep learning model architecture with word embedding and long-short-term memory 52

3.1 Project workflow 61

3.2 Overview of detected chemicals in streams in Minnesota. 64

3.3 Exposure patterns and correlation matrices of single compound and compound group exposure patterns 65

3.4 Preprocessing of microarray samples 66

3.5 Differential gene expression models based on four exposure scenarios across stream sites 67

3.6 Results of differential gene expression analysis with geneset enrichment analysis 70

3.7 Comparison of differential gene expression contrasts to general treatment on gene and biological pathway level 73

3.8 Overview of association rule mining approach 74

3.9 Distribution of AR-metrics across pairwise exposure-gene-rules 74

3.10 Number of exposure-related genes grouped by exposure scenarios 75

3.11 Boxplot of support ratio distribution per ANTECEDENT grouped by exposure scenarios 77

3.12 Scale free topology estimation for weighted gene correlation network analysis 78

3.13 The gene dendrogram was clustered in 21 modules 79

3.14 Overview of module-trait-correlation 81

3.15 Number of exposure-associated genes identified in weighted gene correlation network analysis grouped by exposure scenarios and decreasingly ordered . . 82

3.16 Relation of module memebership and gene significance for an example of a significant module-trait-correlation. 83

3.17 Method comparison with external reference bases 86

4.1 Workflow of the deep learning prediction approach 106

4.2 Preprocessing of chemical-biomolecule interactions from SemMedDB 107

4.3 Coverage of SemMedDB input in comparative toxicogenomic database 109

4.4 Coverage of input data in CTD 111

4.5 Comparison of the model performance with different word embedding vector sizes 113

4.6 Comparison of the training performances for different number of neurons in LSTM layer 116

4.7 Model performance and validation 120

4.8 Performance of models trained with CTD input 126

S1-1 Example of a cell-based bioassay 155

S1-2 The central dogma of molecular biology 158

S2-1 Pairwise Jaccard similarity and overlap of DEA results for single compound exposure scenarios. 168

S2-2 Module-trait-correlations across all exposure traits and all modules 169

S3-1 Model performance of selected word embedding model 177

S3-2 Training performance of selected model with LSTM layer 178

S3-3 Relationship grouped violin plots of prediction probability for chemical-gene interactions from T_{C2G} 180

S3-4 Toxicological evaluation applying functional enrichment on CTD reference sets 181

S3-5 Training and validation curves of word embedding model (n=5000) and lower learning rate. 184

S3-6 Performance of model trained with horizontally augmented, but not tail padded input 185

S3-7 Model performance considering four relationship types 188

S3-8 Concept overlap in input across four predicates 189

List of Tables

2.1 List of toxicological reference data sets (including URL) 43

2.2 List of downloaded data from CTD (including URL) 46

3.1 Selected single detected compounds from chemical analytical data 63

3.2 Significantly functionally enriched terms for the differential gene expression analysis I 69

3.3 Significantly functionally enriched terms for differential gene expression analysis II 72

3.4 Significantly enriched terms for purple module with EE2 or compound group4 80

3.5 Significantly functionally enriched terms for the differential gene expression analysis IV 87

3.6 Significantly functionally enriched terms for the differential gene expression analysis III 88

3.7 Significantly functionally enriched terms to Iopamidol 89

3.8 Significantly enriched terms for method integration of WGCNA and DEA. . . 90

4.1 Evaluation of deep learning models with different sizes of word embedding vectors. 114

4.2 Evaluation of deep learning models with varying numbers of LSTM neurons. 117

4.3 Used symbols for model comparison 119

4.4 Test performances of best folds of 5-fold cross-validated models. 121

4.5 Comparison of SemMedDB trained models with chemical-gene interactions from CTD. 122

4.6 Comparison of selected model architectures with different training inputs from CTD. 127

4.7 Application case of SemMedDB data to CTD trained model with horizontal augmentation 128

S2-1 Compound list for chemical analysis in stream water samples. 162

S2-2 Site information of 10 selected streams in Minnesota. 164

S2-3 Data flags in chemical analytics. 165

S2-4 List of detected chemical compounds with additional information. 166

S2-5 Estrogen and androgen receptor-related activities in water samples of ten selected streams. 167

S2-6 Overlap of single compound-gene association rule mining results to CTD and STITCH 170

S2-7 Overlap of association rule mined compound group-gene interactions to CTD and STITCH 171

S2-8 Overlap of chemical-gene interactions to CTD and STITCH applying differential gene expression analysis 172

S2-9 Overlap of compound group-gene interactions to CTD and STITCH applying differntial gene expression analysis 173

S2-10 Overlap of chemical-gene interactions to CTD and STITCH applying network inference 174

S3-1 Hyperparameter tuning for word embedding model 176

S3-2 Hyperparameter tuning for model with LSTM 177

S3-3 Influence of probability threshold on true-positive rate. 181

S3-4 Chemical wise enrichment results for genes predicted as regulated. 182

S3-5 Test performances of models with not tail-padded horizontally augmented input. 186

S3-6 Number of chemical-biomolecule-relations per predicate. 186

Bibliography

M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: A system for Large-Scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283, Savannah, GA, Nov. 2016. USENIX Association. ISBN 978-1-931971-33-1. URL `https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi`.

M. D. M. AbdulHameed, G. J. Tawa, K. Kumar, D. L. Ippolito, J. A. Lewis, J. D. Stallings, and A. Wallqvist. Systems level analysis and identification of pathways and networks associated with liver fibrosis. *PLoS ONE*, 9(11), 2014. ISSN 19326203. DOI: 10.1371/JOURNAL.PONE.0112193.

M. D. M. AbdulHameed, D. L. Ippolito, J. D. Stallings, and A. Wallqvist. Mining kidney toxicogenomic data by using gene co-expression modules. *BMC Genomics*, 17(1):1–17, 2016. ISSN 14712164. DOI: 10.1186/S12864-016-3143-Y.

S. G. Abernethy, D. MacKay, and L. S. McCarty. Volume fraction correlation for narcosis in aquatic organisms: The key role of partitioning. *Environmental Toxicology and Chemistry*, 7(6):469–481, 1988. ISSN 15528618. DOI: 10.1002/ETC.5620070607.

A. Acharjee, Z. Ament, J. A. West, E. Stanley, and J. L. Griffin. Integration of metabolomics, lipidomics and clinical data using a machine learning method. *BMC Bioinformatics*, 17, 2016. ISSN 14712105. DOI: 10.1186/s12859-016-1292-2.

A. F. Agarap. Deep learning using rectified linear units (relu), 2018. URL `https://arxiv.org/abs/1803.08375`.

A. Aguayo-Orozco, K. Audouze, T. Siggaard, R. Barouki, S. Brunak, and O. Taboureau. sAOP: linking chemical stressors to adverse outcomes pathway networks. *Bioinformatics*, 35(24):1–2, Jul 2019. ISSN 1367-4803. DOI: 10.1093/BIOINFORMATICS/BTZ570.

C. B. Ahlers, M. Fiszman, D. Demner-Fushman, F.-M. Lang, and T. C. Rindflesch. Extracting semantic predications from medline citations for pharmacogenomics. In *Biocomputing 2007*, pages 209–220. World Scientific, 2007. URL `http://psb.stanford.edu/psb-online/proceedings/psb07/abstracts/2007_p209.html`.

T. Akutsu, S. Miyano, and S. Kuhara. Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics*, 16(8):727–734, 2000. ISSN 13674803. DOI: 10.1093/BIOINFORMATICS/16.8.727.

M. Alawad, S. M. Hasan, J. Blair Christian, and G. Tourassi. Retrofitting Word Embeddings with the UMLS Metathesaurus for Clinical Information Extraction. *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, pages 2838–2846, 2019. DOI: 10.1109/BIGDATA.2018.8621999.

B. Alexander-Dann, L. L. Pruteanu, E. Oerton, N. Sharma, I. Berindan-Neagoe, D. Módos, and A. Bender. Developments in toxicogenomics: Understanding and predicting compound-induced toxicity from gene expression data. *Molecular Omics*, 14(4):218–236, 2018. ISSN 25154184. DOI: 10.1039/C8MO00042E.

E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. B. A. McDermott. Publicly Available Clinical BERT Embeddings, 2019. URL `http://arxiv.org/abs/1904.03323`.

G. T. Ankley and D. L. Villeneuve. The fathead minnow in aquatic toxicology: Past, present and future, 2006. ISSN 0166445X. URL `https://www.sciencedirect.com/science/article/pii/S0166445X06000488`.

G. T. Ankley, R. S. Bennett, R. J. Erickson, D. J. Hoff, M. W. Hornung, R. D. Johnson, D. R. Mount, J. W. Nichols, C. L. Russom, P. K. Schmieder, J. A. Serrrano, J. E. Tietge, and D. L. Villeneuve. Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment. *Environmental Toxicology and Chemistry*, 29(3):730–741, 2010. ISSN 15528618. DOI: 10.1002/ETC.34.

P. Antczak, H. J. Jo, S. Woo, L. Scanlan, H. Poynton, A. Loguinov, S. Chan, F. Falciani, and C. Vulpe. Molecular toxicity identification evaluation (mTIE) approach predicts chemical exposure in Daphnia magna. *Environmental Science and Technology*, 47(20):11747–11756, 2013. ISSN 0013936X. DOI: 10.1021/ES402819C.

A. R. Aronson. Metamap: Mapping text to the umls metathesaurus. *Bethesda MD NLM NIH DHHS*, pages 1–26, 2006. URL `http://0-skr.nlm.nih.gov.library.law.suffolk.edu/papers/references/metamap06.pdf`.

J. Arstikaitis, F. Gagné, and D. G. Cyr. Exposure of fathead minnows to municipal wastewater effluent affects intracellular signaling pathways in the liver. *Comparative Biochemistry and Physiology Part - C: Toxicology and Pharmacology*, 164:1–10, 2014. ISSN 18781659. DOI: 10.1016/J.CBPC.2014.04.002.

J. Asselman, M. E. Pfrender, J. A. Lopez, J. R. Shaw, and K. A. De Schamphelaere. Gene Coexpression Networks Drive and Predict Reproductive Effects in Daphnia in Response to Environmental Disturbances. *Environmental Science and Technology*, 52(1):317–326, 2018. ISSN 15205851. DOI: 10.1021/ACS.EST.7B05256.

G. Bakal, P. Talari, E. V. Kakani, and R. Kavuluru. Exploiting semantic patterns over biomedical knowledge graphs for predicting treatment and causative relations. *Journal of Biomedical Informatics*, 82:189–199, Jan 2018. ISSN 15320464. DOI: 10.1016/J.JBI.2018.05.003.

G. Barel and R. Herwig. Network and pathway analysis of toxicogenomics data. *Frontiers in Genetics*, 9, Oct 2018. ISSN 16648021. DOI: 10.3389/FGENE.2018.00484.

J. Barrera-Gómez, L. Agier, L. Portengen, M. Chadeau-Hyam, L. Giorgis-Allemand, V. Siroux, O. Robinson, J. Vlaanderen, J. R. González, M. Nieuwenhuijsen, P. Vineis, M. Vrijheid, R. Vermeulen, R. Slama, and X. Basagaña. A systematic comparison of statistical methods to detect interactions in exposome-health associations. *Environmental Health: A Global Access Science Source*, 16(1):1–13, 2017. ISSN 1476069X. DOI: 10.1186/s12940-017-0277-6.

M. G. Barron, C. R. Lilavois, and T. M. Martin. MOAtox: A comprehensive mode of action and acute aquatic toxicity database for predictive model development. *Aquatic Toxicology*, 161:102–107, 2015. ISSN 18791514. DOI: 10.1016/j.aquatox.2015.02.001.

C. A. Bejan and J. C. Denny. Learning to identify treatment relations in clinical text. In *AMIA Annual Symposium Proceedings*, volume 2014, page 282. American Medical Informatics Association, 2014. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4419965/`.

S. M. Bell, M. M. Angrish, C. E. Wood, and S. W. Edwards. Integrating publicly available data to generate computationally predicted adverse outcome pathways for fatty liver. *Toxicological Sciences*, 150(2):510–520, 02 2016. ISSN 1096-6080. DOI: 10.1093/toxsci/kfw017.

J. P. Berninger, D. Martinović-Weigelt, N. Garcia-Reyero, L. Escalon, E. J. Perkins, G. T. Ankley, and D. L. Villeneuve. Using transcriptomic tools to evaluate biological effects across effluent gradients at a diverse set of study sites in Minnesota, USA. *Environmental Science and Technology*, 48(4):2404–2412, 2014. ISSN 0013936X. DOI: 10.1021/es4040254.

P. Bjerregaard, K. L. Kinnberg, M. P. Mose, and H. Holbech. Investigation of the potential endocrine effect of nitrate in zebrafish Danio rerio and brown trout Salmo trutta. *Comparative Biochemistry and Physiology Part - C: Toxicology and Pharmacology*, 211:32–40, 2018. ISSN 18781659. DOI: 10.1016/j.cbpc.2018.05.006.

B. R. Blackwell, G. T. Ankley, P. M. Bradley, K. A. Houck, S. S. Makarov, A. V. Medvedev, J. Swintek, and D. L. Villeneuve. Potential Toxicity of Complex Mixtures in Surface Waters from a Nationwide Survey of United States Streams: Identifying in Vitro Bioactivities and Causative Chemicals. *Environmental Science and Technology*, 53(2):973–983, 2019. ISSN 15205851. DOI: 10.1021/acs.est.8b05304.

O. Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004. DOI: 10.1093/nar/gkh061.

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. ISSN 2307-387X. DOI: 10.1162/tacl_a_00051.

P. M. Bradley, C. A. Journey, K. M. Romanok, L. B. Barber, H. T. Buxton, W. T. Foreman, E. T. Furlong, S. T. Glassmeyer, M. L. Hladik, L. R. Iwanowicz, D. K. Jones, D. W. Kolpin, K. M. Kuivila, K. A. Loftin, M. A. Mills, M. T. Meyer, J. L. Orlando, T. J. Reilly, K. L. Smalling, and D. L. Villeneuve. Expanded Target-Chemical Analysis Reveals Extensive Mixed-Organic-Contaminant Exposure in U.S. Streams. *Environmental Science and Technology*, 51(9):4792–4802, 2017. ISSN 15205851. DOI: 10.1021/acs.est.7b00012.

P. M. Bradley, C. A. Journey, J. P. Berninger, D. T. Button, J. M. Clark, S. R. Corsi, L. A. DeCicco, K. G. Hopkins, B. J. Huffman, N. Nakagaki, J. E. Norman, L. H. Nowell, S. L. Qi, P. C. VanMetre, and I. R. Waite. Mixed-chemical exposure and predicted effects potential in wadeable southeastern USA streams. *Science of the Total Environment*, 655:70–83, Mar 2019. ISSN 18791026. DOI: 10.1016/j.scitotenv.2018.11.186.

W. Busch, S. Schmidt, R. Kühne, T. Schulze, M. Krauss, and R. Altenburger. Micropollutants in European rivers: A mode of action survey to support the development of effect-based tools for water monitoring. *Environmental Toxicology and Chemistry*, 35(8):1887–1899, 2016. ISSN 15528618. DOI: 10.1002/etc.3460.

I. C. Calderón-Delgado, D. A. Mora-Solarte, and Y. M. Velasco-Santamaría. Physiological and enzymatic responses of Chlorella vulgaris exposed to produced water and its potential for bioremediation. *Environmental Monitoring and Assessment*, 191(6), 2019. ISSN 15732959. DOI: 10.1007/s10661-019-7519-8.

T. J. Callahan, I. J. Tripodi, H. Pielke-Lombardo, and L. E. Hunter. Knowledge-based biomedical Data Science. *Annual Review of Biomedical Data Science*, 1(1-2):23–41, 2017. ISSN 24518484. DOI: 10.3233/ds-170001.

D. Cameron, O. Bodenreider, H. Yalamanchili, T. Danh, S. Vallabhaneni, K. Thirunarayan, A. P. Sheth, and T. C. Rindflesch. A graph-based recovery and decomposition of Swanson's hypothesis using semantic predications. *Journal of Biomedical Informatics*, 46(2):238–251, 2013. ISSN 15320464. DOI: 10.1016/j.jbi.2012.09.004.

S. Carbon, E. Douglass, N. Dunn, B. Good, N. L. Harris, S. E. Lewis, C. J. Mungall, S. Basu, R. L. Chisholm, R. J. Dodson, E. Hartline, P. Fey, P. D. Thomas, L. P. Albou, D. Ebert, M. J. Kesling, H. Mi, A. Muruganujan, X. Huang, S. Poudel, T. Mushayahama, J. C. Hu, S. A. LaBonte, D. A. Siegele, G. Antonazzo, H. Attrill, N. H. Brown, S. Fexova, P. Garapati, T. E. Jones, S. J. Marygold, G. H. Millburn, A. J. Rey, V. Trovisco, G. Dos Santos, D. B. Emmert, K. Falls, P. Zhou, J. L. Goodman, V. B. Strelets, J. Thurmond, M. Courtot, D. S. Osumi, H. Parkinson, P. Roncaglia, M. L. Acencio, M. Kuiper, A. Lreid, C. Logie, R. C. Lovering, R. P. Huntley, P. Denny, N. H. Campbell, B. Kramarz, V. Acquaah, S. H. Ahmad, H. Chen, J. H. Rawson, M. C. Chibucos, M. Giglio, S. Nadendla, R. Tauber, M. J. Duesbury, N. T. Del, B. H. Meldal, L. Perfetto, P. Porras, S. Orchard, A. Shrivastava, Z. Xie, H. Y. Chang, R. D. Finn, A. L. Mitchell, N. D. Rawlings, L. Richardson, A. Sangrador-Vegas, J. A. Blake, K. R. Christie, M. E. Dolan, H. J. Drabkin, D. P. Hill, L. Ni, D. Sitnikov, M. A. Harris, S. G. Oliver, K. Rutherford, V. Wood, J. Hayles, J. Bahler, A. Lock, E. R. Bolton, J. De Pons, M. Dwinell, G. T. Hayman, S. J. Laulederkind, M. Shimoyama, M. Tutaj, S. J. Wang, P. D'Eustachio, L. Matthews, J. P. Balhoff, S. A. Aleksander, G. Binkley, B. L. Dunn, J. M. Cherry, S. R. Engel, F. Gondwe, K. Karra, K. A. MacPherson, S. R. Miyasato, R. S. Nash, P. C. Ng, T. K. Sheppard, A. Shrivatsav Vp, M. Simison, M. S. Skrzypek, S. Weng, E. D. Wong, M. Feuermann, P. Gaudet, E. Bakker, T. Z. Berardini, L. Reiser, S. Subramaniam, E. Huala, C. Arighi, A. Auchincloss, K. Axelsen, G. P. Argoud, A. Bateman, B. Bely, M. C. Blatter, E. Boutet, L. Breuza, A. Bridge, R. Britto, H. Bye-A-Jee, C. Casals-Casas, E. Coudert, A. Estreicher, L. Famiglietti, P. Garmiri, G. Georghiou, A. Gos, N. Gruaz-Gumowski, E. Hatton-Ellis, U. Hinz, C. Hulo, A. Ignatchenko, F. Jungo, G. Keller, K. Laiho, P. Lemercier, D. Lieberherr, Y. Lussi, A. Mac-Dougall, M. Magrane, M. J. Martin, P. Masson, D. A. Natale, N. N. Hyka, I. Pedruzzi, K. Pichler, S. Poux, C. Rivoire, M. Rodriguez-Lopez, T. Sawford, E. Speretta, A. Shypitsyna, A. Stutz, S. Sundaram, M. Tognolli, N. Tyagi, K. Warner, R. Zaru, C. Wu, J. Chan, J. Cho, S. Gao, C. Grove, M. C. Harrison, K. Howe, R. Lee, J. Mendel, H. M. Muller, D. Raciti, K. Van Auken, M. Berriman, L. Stein, P. W. Sternberg, D. Howe, S. Toro, and M. Westerfield. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, 47(D1):D330–D338, 2019. ISSN 13624962. DOI: 10.1093/nar/gky1055.

M. Carlson. *org.Dr.eg.db: Genome wide annotation for Zebrafish*, 2019. URL `http://bioconductor.org/packages/release/data/annotation/html/org.Dr.eg.db.html`. R package version 3.10.0.

R. Carson. *Silent spring*. Houghton Mifflin Harcourt, 2002.

J. C. Carvaillo, R. Barouki, X. Coumoul, and K. Audouze. Linking Bisphenol S to Adverse Outcome Pathways Using a Combined Text Mining and Systems Biology Approach. *Environmental health perspectives*, 127(4): 47005, 2019. ISSN 15529924. DOI: 10.1289/EHP4200.

N. L. Catlett, A. J. Bargnesi, S. Ungerer, T. Seagaran, W. Ladd, K. O. Elliston, and D. Pratt. Reverse causal reasoning: applying qualitative causal knowledge to the interpretation of high-throughput data. *BMC Bioinformatics*, 14(1):340, 2013. DOI: 10.1186/1471-2105-14-340.

N. Cedergreen. Quantifying synergy: a systematic review of mixture toxicity studies within environmental toxicology. *PloS one*, 9(5):e96580, 2014. DOI: 10.1371/JOURNAL.PONE.0096580.

D. Chen, F. Zhang, Q. Zhao, and J. Xu. OmicsARules: a R package for integration of multi-omics datasets via association rules mining. *BMC Bioinformatics*, 20(1), 2019. ISSN 14712105. DOI: 10.1186/S12859-019-3171-0.

L. Chen, B. Chen, Y. Ren, and D. Ji. Long short-term memory RNN for biomedical named entity recognition. *BMC Bioinformatics*, 18(1):1–11, 2017. ISSN 14712105. DOI: 10.1186/s12859-017-1868-5.

T. Chen. *BERT Argues : How Attention Informs Argument Mining*. University of Richmond, 2021. URL `https://scholarship.richmond.edu/honors-theses/1589/`.

W. Choi and H. Lee. Inference of Biomedical Relations among Chemicals, Genes, Diseases, and Symptoms Using Knowledge Representation Learning. *IEEE Access*, 7:179373–179384, 2019. ISSN 21693536. DOI: 10.1109/ACCESS.2019.2957812.

F. Chollet et al. Keras. `https://keras.io`, 2015.

Q. Cong, Z. Feng, F. Li, L. Zhang, G. Rao, and C. Tao. Constructing Biomedical Knowledge Graph Based on SemMedDB and Linked Open Data. *Proceedings - 2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018*, pages 1628–1631, 2019. DOI: 10.1109/BIBM.2018.8621568.

I. Cordero-Herrera, D. D. Guimarães, C. Moretti, Z. Zhuge, H. Han, S. McCann Haworth, A. E. Uribe Gonzalez, D. C. Andersson, E. Weitzberg, J. O. Lundberg, and M. Carlström. Head-to-head comparison of inorganic nitrate and metformin in a mouse model of cardiometabolic disease. *Nitric Oxide - Biology and Chemistry*, 97:48–56, Feb 2020. ISSN 10898611. DOI: 10.1016/J.NIOX.2020.01.013.

S. L. Costigan, J. Werner, J. D. Ouellet, L. G. Hill, and R. D. Law. Expression profiling and gene ontology analysis in fathead minnow (Pimephales promelas) liver following exposure to pulp and paper mill effluents. *Aquatic Toxicology*, 122-123:44–55, 2012. ISSN 0166445X. DOI: 10.1016/J.AQUATOX.2012.05.011.

C. Creighton and S. Hanash. Mining gene expression databases for association rules. *Bioinformatics*, 19(1): 79–86, 2003. ISSN 13674803. DOI: 10.1093/BIOINFORMATICS/19.1.79.

F. Crick. Central Dogma of Molecular Biology. *Nature*, 227, 1970. DOI: 10.1038/227561A0.

K. Dale, M. B. Müller, Z. Tairova, E. A. Khan, K. Hatlen, M. Grung, F. Yadetie, R. Lille-Langøy, N. Blaser, H. J. Skaug, J. L. Lyche, A. Arukwe, K. Hylland, O. A. Karlsen, and A. Goksøyr. Contaminant accumulation and biological responses in Atlantic cod (Gadus morhua) caged at a capped waste disposal site in Kollevåg, Western Norway. *Marine Environmental Research*, 145:39–51, 2019. ISSN 18790291. DOI: 10.1016/J.MARENVRES.2019.02.003.

A. P. Davis, C. J. Grondin, R. J. Johnson, D. Sciaky, B. L. King, R. McMorran, J. Wiegers, T. C. Wiegers, and C. J. Mattingly. The Comparative Toxicogenomics Database: Update 2017. *Nucleic Acids Research*, 45(D1):D972–D978, 2017. ISSN 13624962. DOI: 10.1093/NAR/GKW838.

A. P. Davis, T. C. Wiegers, J. Wiegers, R. J. Johnson, D. Sciaky, C. J. Grondin, and C. J. Mattingly. Chemical-Induced phenotypes at CTD help inform the predisease state and construct adverse outcome pathways. *Toxicological Sciences*, 165(1):145–156, 2018. ISSN 10960929. DOI: 10.1093/TOXSCI/KFY131.

A. P. Davis, C. J. Grondin, R. J. Johnson, D. Sciaky, R. McMorran, J. Wiegers, T. C. Wiegers, and C. J. Mattingly. The Comparative Toxicogenomics Database: Update 2019. *Nucleic Acids Research*, 47(D1): D948–D954, 2019. ISSN 13624962. DOI: 10.1093/NAR/GKY868.

A. P. Davis, C. J. Grondin, R. J. Johnson, D. Sciaky, J. Wiegers, T. C. Wiegers, and C. J. Mattingly. Comparative Toxicogenomics Database (CTD): Update 2021. *Nucleic Acids Research*, 49(D1):D1138–D1143, 2021. ISSN 13624962. DOI: 10.1093/NAR/GKAA891.

D. Degli Esposti, C. Almunia, M. A. Guery, N. Koenig, J. Armengaud, A. Chaumot, and O. Geffard. Co-expression network analysis identifies gonad- and embryo-associated protein modules in the sentinel species Gammarus fossarum. *Scientific Reports*, 9(1), 2019. ISSN 20452322. DOI: 10.1038/s41598-019-44203-5.

J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding, 2019. URL `https://arxiv.org/abs/1810.04805`.

D. J. Dix, K. A. Houck, M. T. Martin, A. M. Richard, R. W. Setzer, and R. J. Kavlock. The toxcast program for prioritizing toxicity testing of environmental chemicals. *Toxicological Sciences*, 95(1):5–12, 2007. ISSN 10966080. DOI: 10.1093/TOXSCI/KFL103.

T. Y. Doktorova, N. O. Oki, T. Mohorič, T. E. Exner, and B. Hardy. A semi-automated workflow for adverse outcome pathway hypothesis generation: The use case of non-genotoxic induced hepatocellular carcinoma. *Regulatory Toxicology and Pharmacology*, 114, Jul 2020. ISSN 10960295. DOI: 10.1016/J.YRTPH.2020.104652.

R. Dollah and M. Aono. Ontology based Approach for Classifying Biomedical Text Abstracts. *International Journal of Data Engineering (IJDE)*, 2(1):1–15, 2011. URL `https://www.semanticscholar.org/paper/Ontology-based-Approach-for-Classifying-Biomedical-Dollah-Aono/a543162b14ef236c41426b16dc8715054a281238`.

W. Dong, P. P. Simeonova, R. Gallucci, J. Matheson, R. Fannin, P. Montuschi, L. Flood, and M. I. Luster. Cytokine expression in hepatocytes: Role of oxidant stress. *Journal of Interferon and Cytokine Research*, 18(8):629–638, 1998. ISSN 10799907. DOI: 10.1089/JIR.1998.18.629.

S. Durinck, Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma, and W. Huber. BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16):3439–3440, 2005. ISSN 13674803. DOI: 10.1093/BIOINFORMATICS/BTI525.

J. L. Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990. DOI: 10.1207/S15516709COG1402_1.

B. I. Escher, R. Ashauer, S. Dyer, J. L. Hermens, J. H. Lee, H. A. Leslie, P. Mayer, J. P. Meador, and M. S. Warnekk. Crucial role of mechanisms and modes of toxic action for understanding tissue residue toxicity and internal effect concentrations of organic chemicals. *Integrated Environmental Assessment and Management*, 7(1):28–49, 2011. ISSN 15513793. DOI: 10.1002/IEAM.100.

B. I. Escher, C. Van Daele, M. Dutt, J. Y. Tang, and R. Altenburger. Most oxidative stress response in water samples comes from unknown chemicals: The need for effect-based water quality trigger values. *Environmental Science and Technology*, 47(13):7002–7011, 2013. ISSN 0013936X. DOI: 10.1021/ES304793H.

B. I. Escher, J. Hackermüller, T. Polte, S. Scholz, A. Aigner, R. Altenburger, A. Böhme, S. K. Bopp, W. Brack, W. Busch, M. Chadeau-Hyam, A. Covaci, A. Eisenträger, J. J. Galligan, N. Garcia-Reyero, T. Hartung, M. Hein, G. Herberth, A. Jahnke, J. Kleinjans, N. Klüver, M. Krauss, M. Lamoree, I. Lehmann, T. Luckenbach, G. W. Miller, A. Müller, D. H. Phillips, T. Reemtsma, U. Rolle-Kampczyk, G. Schüürmann, B. Schwikowski, Y.-M. Tan, S. Trump, S. Walter-Rohde, and J. F. Wambaugh. From the exposome to mechanistic understanding of chemical-induced adverse effects. *Environment International*, 99:97–106, 2017. ISSN 0160-4120. DOI: HTTPS://DOI.ORG/10.1016/J.ENVINT.2016.11.029.

J. D. Ewald, O. Soufan, D. Crump, M. Hecker, J. Xia, and N. Basu. EcoToxModules: Custom Gene Sets to Organize and Analyze Toxicogenomics Data from Ecological Species. *Environmental Science and Technology*, 54(7):4376–4387, 2020. ISSN 15205851. DOI: 10.1021/ACS.EST.9B06607.

A. Fabregat, S. Jupe, L. Matthews, K. Sidiropoulos, M. Gillespie, P. Garapati, R. Haw, B. Jassal, F. Korninger, B. May, M. Milacic, C. D. Roca, K. Rothfels, C. Sevilla, V. Shamovsky, S. Shorser, T. Varusai, G. Viteri, J. Weiser, G. Wu, L. Stein, H. Hermjakob, and P. D'Eustachio. The Reactome Pathway Knowledgebase. *Nucleic Acids Research*, 46(D1):D649–D655, 2018. ISSN 13624962. DOI: 10.1093/NAR/GKX1132.

H. Feng, H. Cao, J. Li, H. Zhang, Q. Xue, X. Liu, A. Zhang, and J. Fu. Estrogenic activity of benzotriazole UV stabilizers evaluated through in vitro assays and computational studies. *Science of the Total Environment*, 727:138549, 2020. ISSN 18791026. DOI: 10.1016/J.SCITOTENV.2020.138549.

M. Ferrey, D. Martinovic, W. Backe, and A. Andrews. Pharmaceuticals and chemicals of concern in rivers: Occurrence and biological effects. *Minnesota Pollution Control Agency*, pages 1–70, 2017. URL `https://www.pca.state.mn.us/sites/default/files/tdr-g1-20.pdf`.

A. Feswick, J. R. Loughery, M. A. Isaacs, K. R. Munkittrick, and C. J. Martyniuk. Molecular initiating events of the intersex phenotype: Low-dose exposure to 17α-ethinylestradiol rapidly regulates molecular networks associated with gonad differentiation in the adult fathead minnow testis. *Aquatic Toxicology*, 181:46–56, 2016. ISSN 18791514. DOI: 10.1016/J.AQUATOX.2016.10.021.

A. Feswick, M. Isaacs, A. Biales, R. W. Flick, D. C. Bencic, R.-L. Wang, C. Vulpe, M. Brown-Augustine, A. Loguinov, F. Falciani, P. Antczak, J. Herbert, L. Brown, N. D. Denslow, K. J. Kroll, C. Lavelle, V. Dang, L. Escalon, N. Garcia-Reyero, C. J. Martyniuk, and K. R. Munkittrick. How consistent are we? interlaboratory comparison study in fathead minnows using the model estrogen 17-ethinylestradiol to develop recommendations for environmental transcriptomics. *Environmental Toxicology and Chemistry*, 36 (10):2614–2623, 2017. DOI: HTTPS://DOI.ORG/10.1002/ETC.3799.

M. W. E. J. Fiers, L. Minnoye, S. Aibar, C. Bravo González-Blas, Z. Kalender Atak, and S. Aerts. Mapping gene regulatory networks from single-cell omics data. *Briefings in Functional Genomics*, 17:246–254, Jan 2018. ISSN 2041-2649. DOI: 10.1093/BFGP/ELX046.

Y. Fu, D. Dominissini, G. Rechavi, and C. He. Gene expression regulation mediated through reversible m 6 A RNA methylation. *Nature Reviews Genetics*, 15(5):293–306, 2014. ISSN 14710064. DOI: 10.1038/NRG3724.

A. Gandar, P. Laffaille, C. Canlet, M. Tremblay-Franco, R. Gautier, A. Perrault, L. Gress, P. Mormède, N. Tapie, H. Budzinski, and S. Jean. Adaptive response under multiple stress exposure in fish: From the molecular to individual level. *Chemosphere*, 188:60–72, 2017. ISSN 18791298. DOI: 10.1016/j.chemosphere.2017.08.089.

S. Gao, O. Kotevska, A. Sorokine, and J. B. Christian. A pre-training and self-training approach for biomedical named entity recognition. *PLoS ONE*, 16:1–23, Feb 2021. ISSN 19326203. DOI: 10.1371/journal.pone.0246310.

N. Garcia-Reyero and E. J. Perkins. Systems biology: Leading the revolution in ecotoxicology. *Environmental Toxicology and Chemistry*, 30(2):265–273, 2011. ISSN 07307268. DOI: 10.1002/etc.401.

N. Garcia-Reyero, I. R. Adelman, D. Martinović, L. Liu, and N. D. Denslow. Site-specific impacts on gene expression and behavior in fathead minnows (Pimephales promelas) exposed in situ to streams adjacent to sewage treatment plants. *BMC Bioinformatics*, 10(SUPPL. 11), 2009. ISSN 14712105. DOI: 10.1186/1471-2105-10-S11-S11.

N. Garcia-Reyero, C. M. Lavelle, B. L. Escalon, D. Martinović, K. J. Kroll, P. W. Sorensen, and N. D. Denslow. Behavioral and genomic impacts of a wastewater effluent on the fathead minnow. *Aquatic Toxicology*, 101 (1):38–48, 2011. ISSN 0166445X. DOI: 10.1016/j.aquatox.2010.08.014.

Gene Ontology Consortium. The Gene Oncology (GO) database and informatics resource. *Nucleic Acids Research*, 32:D258–D261, 2004. ISSN 03051048. DOI: 10.1093/nar/gkh036.

J. Gu, F. Sun, L. Qian, and G. Zhou. Chemical-induced disease relation extraction via attention-based distant supervision. *BMC Bioinformatics*, 20(1):1–14, 2019. ISSN 14712105. DOI: 10.1186/s12859-019-2884-4.

M. Hahsler. arulesViz: Interactive visualization of association rules with R. *R Journal*, 9(2):163–175, 2017. ISSN 20734859. DOI: 10.32614/rj-2017-047.

M. Hahsler, B. Grün, and K. Hornik. Arules - A computational environment for mining association rules and frequent item sets. *Journal of Statistical Software*, 14, 2005. ISSN 15487660. DOI: 10.18637/jss.v014.i15.

T. Hartung, R. E. FitzGerald, P. Jennings, G. R. Mirams, M. C. Peitsch, A. Rostami-Hodjegan, I. Shah, M. F. Wilks, and S. J. Sturla. Systems Toxicology: Real World Applications and Opportunities. *Chemical Research in Toxicology*, 30(4):870–882, 2017. ISSN 15205010. DOI: 10.1021/acs.chemrestox.7b00003.

S. A. Hermsen, T. E. Pronk, E. J. van den Brandhof, L. T. van der Ven, and A. H. Piersma. Concentration-response analysis of differential gene expression in the zebrafish embryotoxicity test following flusilazole exposure. *Toxicological Sciences*, 127(1):303–312, 2012. ISSN 10960929. DOI: 10.1093/toxsci/kfs092.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. DOI: 10.1162/neco.1997.9.8.1735.

T. F. Holth, R. Nourizadeh-Lillabadi, M. Blaesbjerg, M. Grung, H. Holbech, G. I. Petersen, P. Aleström, and K. Hylland. Differential gene expression and biomarkers in zebrafish (Danio rerio) following exposure to produced water components. *Aquatic Toxicology*, 90(4):277–291, 2008. ISSN 0166445X. DOI: 10.1016/j.aquatox.2008.08.020.

P. Hou, O. Jolliet, J. Zhu, and M. Xu. Estimate ecotoxicity characterization factors for chemicals in life cycle assessment using machine learning models. *Environment International*, 135:105393, Feb 2020. ISSN 18736750. DOI: 10.1016/j.envint.2019.105393.

D. Hristovski, A. Kastrin, D. Dinevski, and T. C. Rindflesch. Towards implementing semantic literature-based discovery with a graph database. In *The Seventh International Conference on Advances in Databases, Knowledge, andData Applications*, pages 180–184, 2015. ISBN 9781612084084. URL `https://www.semanticscholar.org/paper/Towards-Implementing-Semantic-Literature-Based-with-Hristovski-Kastrin/90284ada9c2bf21721a1329053482a707f584fdf`.

D. W. Huang, B. T. Sherman, X. Zheng, J. Yang, T. Imamichi, R. Stephens, and R. A. Lempicki. Extracting biological meaning from large gene lists with david. *Current protocols in bioinformatics*, 27(1):13–11, 2009. DOI: 10.1002/0471250953.bi1311s27.

B. L. Humphreys, D. A. B. Lindberg, H. M. Schoolman, and G. O. Barnett. The Unified Medical Language System: An Informatics Research Collaboration. *Journal of the American Medical Informatics Association*, 5(1):1–11, 01 1998. ISSN 1067-5027. DOI: 10.1136/jamia.1998.0050001.

Y. Igarashi, N. Nakatsu, T. Yamashita, A. Ono, Y. Ohno, T. Urushidani, and H. Yamada. Open TG-GATEs: a large-scale toxicogenomics database. *Nucleic acids research*, 43(D1):D921–D927, 2015. DOI: 10.1093/nar/gku955.

P. Jain, P. Vineis, B. Liquet, J. Vlaanderen, B. Bodinier, K. Van Veldhoven, M. Kogevinas, T. J. Athersuch, L. Font-Ribera, C. M. Villanueva, R. Vermeulen, and M. Chadeau-Hyam. A multivariate approach to investigate the combined biological effects of multiple exposures. *Journal of Epidemiology and Community Health*, 72(7):564–571, 2018. ISSN 14702738. DOI: 10.1136/jech-2017-210061.

B. Jassal, L. Matthews, G. Viteri, C. Gong, P. Lorente, A. Fabregat, K. Sidiropoulos, J. Cook, M. Gillespie, R. Haw, F. Loney, B. May, M. Milacic, K. Rothfels, C. Sevilla, V. Shamovsky, S. Shorser, T. Varusai, J. Weiser, G. Wu, L. Stein, H. Hermjakob, and P. D'Eustachio. The reactome pathway knowledgebase. *Nucleic Acids Research*, 48(D1):D498–D503, jan 2020. ISSN 13624962. DOI: 10.1093/nar/gkz1031.

J. Jeong and J. Choi. Development of AOP relevant to microplastics based on toxicity mechanisms of chemical additives using ToxCast and deep learning models combined approach. *Environment International*, 137: 105557, Apr 2020. ISSN 18736750. DOI: 10.1016/j.envint.2020.105557.

R. D. Jimenez, D. C. Martins-Jr, and C. S. Santos. One genetic algorithm per gene to infer gene networks from expression data. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 4(1):1–22, 2015. ISSN 21926670. DOI: 10.1007/s13721-015-0092-3.

A. Jimeno Yepes. Word embeddings and recurrent neural networks based on Long-Short Term Memory nodes in supervised biomedical word sense disambiguation. *Journal of Biomedical Informatics*, 73:137–147, 2017. ISSN 15320464. DOI: 10.1016/j.jbi.2017.08.001.

F. Jornod, M. Rugard, L. Tamisier, X. Coumoul, H. R. Andersen, R. Barouki, and K. Audouze. AOP4EUpest: Mapping of pesticides in adverse outcome pathways using a text mining tool. *Bioinformatics*, 36(15): 4379–4381, 2020. ISSN 14602059. DOI: 10.1093/bioinformatics/btaa545.

F. Jornod, T. Jaylet, L. Blaha, D. Sarigiannis, L. Tamisier, and K. Audouze. AOP-helpFinder webserver: a tool for comprehensive analysis of the literature to support adverse outcome pathways development . *Bioinformatics*, 2021. DOI: 10.1093/BIOINFORMATICS/BTAB750.

M. Kanehisa and S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28 (1):27–30, 2000. ISSN 03051048. DOI: 10.1093/NAR/28.1.27.

M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. The KEGG resource for deciphering the genome. *Nucleic Acids Research*, 32(suppl_1), 2004. ISSN 03051048. DOI: 10.1093/NAR/GKH063.

D. F. Kapraun, J. F. Wambaugh, C. L. Ring, R. Tornero-Velez, and R. Woodrow Setzer. A method for identifying prevalent chemical combinations in the U.S. population. *Environmental Health Perspectives*, 125 (8), 2017. ISSN 15529924. DOI: 10.1289/EHP1265.

F. Karel and J. Klema. Quantitative association rule mining in genomics using apriori knowledge. In *Workshop on Prior Conceptual Knowledge in Machine Learning and Knowledge Discovery (PriCKL'07)*, pages 53–64, 2007. URL `https://www.researchgate.net/profile/Agnieszka-Lawrynowicz/publication/227118943_A_study_of_the_SEMINTEC_approach_to_frequent_pattern_mining/links/0046353af43881e1aa000000/A-study-of-the-SEMINTEC-approach-to-frequent-pattern-mining.pdf#page=63`.

A. Kastrin, P. Ferk, and B. Leskošek. Predicting potential drug-drug interactions on topological and semantic similarity features using statistical learning. *PLoS ONE*, 13(5):1–23, 2018. ISSN 19326203. DOI: 10.1371/JOURNAL.PONE.0196865.

S. Kauffman, C. Peterson, B. Samuelsson, and C. Troein. Genetic networks with canalyzing Boolean rules are always stable. In *Proceedings of the National Academy of Sciences*, volume 101, pages 17102–17107, 2004. DOI: 10.1073/PNAS.0407783101.

K. A. Kellock, A. P. Moore, and R. B. Bringolf. Chronic nitrate exposure alters reproductive physiology in fathead minnows. *Environmental Pollution*, 232:322–328, 2018. ISSN 18736424. DOI: 10.1016/J.ENVPOL.2017.08.004.

K. A. Kidd, P. J. Blanchfield, K. H. Mills, V. P. Palace, R. E. Evans, J. M. Lazorchak, and R. W. Flick. Collapse of a fish population after exposure to a synthetic estrogen. *Proceedings of the National Academy of Sciences of the United States of America*, 104(21):8897–8901, 2007. ISSN 00278424. DOI: 10.1073/PNAS.0609568104.

H. Kilicoglu, G. Rosemblat, M. Fiszman, and T. C. Rindflesch. Constructing a semantic predication gold standard from the biomedical literature. *BMC Bioinformatics*, 12(1), 2011. ISSN 14712105. DOI: 10.1186/1471-2105-12-486.

H. Kilicoglu, D. Shin, M. Fiszman, G. Rosemblat, and T. C. Rindflesch. SemMedDB: A PubMed-scale repository of biomedical semantic predications. *Bioinformatics*, 28(23):3158–3160, 2012. ISSN 13674803. DOI: 10.1093/BIOINFORMATICS/BTS591.

H. Kilicoglu, G. Rosemblat, M. Fiszman, and D. Shin. Broad-coverage biomedical relation extraction with SemRep. *BMC Bioinformatics*, 21(1):1–28, 2020. ISSN 14712105. DOI: 10.1186/s12859-020-3517-7.

D. Knapen, M. M. Angrish, M. C. Fortin, I. Katsiadaki, M. Leonard, L. Margiotta-Casaluci, S. Munn, J. M. O'Brien, N. Pollesch, L. C. Smith, X. Zhang, and D. L. Villeneuve. Adverse outcome pathway networks

I: Development and applications. *Environmental Toxicology and Chemistry*, 37(6):1723–1733, 2018. ISSN 15528618. DOI: 10.1002/ETC.4125.

E. J. Koh and S. Y. Hwang. Multi-omics approaches for understanding environmental exposure and human health. *Molecular and Cellular Toxicology*, 15(1):1–7, 2019. ISSN 20928467. DOI: 10.1007/S13273-019-0001-4.

R. Kolde. *pheatmap: Pretty Heatmaps*, 2019. URL `https://CRAN.R-project.org/package=pheatmap`. R package version 1.0.12.

D. W. Kolpin, E. T. Furlong, M. T. Meyer, E. M. Thurman, S. D. Zaugg, L. B. Barber, and H. T. Buxton. Pharmaceuticals, hormones, and other organic wastewater contaminants in U.S. streams, 1999-2000: A national reconnaissance. *Environmental Science and Technology*, 36(6):1202–1211, 2002. ISSN 0013936X. DOI: 10.1021/ES011055J.

K. Kongsbak, N. Hadrup, K. Audouze, and A. M. Vinggaard. Applicability of computational systems biology in toxicology. *Basic and Clinical Pharmacology and Toxicology*, 115(1):45–49, 2014. ISSN 17427843. DOI: 10.1111/BCPT.12216.

A. Kortenkamp and M. Faust. Regulate to reduce chemical mixture risk. *Science*, 361(6399):224–226, 2018. ISSN 10959203. DOI: 10.1126/SCIENCE.AAT9219.

Kraak, Michiel. Chapter 6.4: Diagnostic risk assessment approaches and tools. In A. van Gestel Cornelis, F. G. Van Belleghem, N. W. van den Brink, S. T. J. Droge, T. Hamers, J. L. Hermens, M. H. Kraak, A. J. Löhr, J. R. Parsons, A. M. Ragas, N. M. van Straalen, and M. G. Vijver, editors, *Environmental Toxicology, an open online textbook*, pages 481–486. Elsevier, 2021. URL `https://research.ou.nl/en/publications/environmental-toxicology-an-open-online-textbook`.

A. Krämer, J. Green, J. Pollard, and S. Tugendreich. Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics*, 30(4):523–530, 2014. ISSN 13674803. DOI: 10.1093/BIOINFORMATICS/BTT703.

S. Krämer, W. Busch, and A. Schüttler. A Self-Organizing Map of the Fathead Minnow Liver Transcriptome to Identify Consistent Toxicogenomic Patterns across Chemical Fingerprints. *Environmental Toxicology and Chemistry*, 39(3):526–537, 2020. ISSN 15528618. DOI: 10.1002/ETC.4646.

M. Kuhn, C. von Mering, M. Campillos, L. J. Jensen, and P. Bork. STITCH: Interaction networks of chemicals and proteins. *Nucleic Acids Research*, 36(SUPPL. 1), 2008. ISSN 03051048. DOI: 10.1093/NAR/GKM795.

E. Kumar. *Natural language processing*. IK International Pvt Ltd, 2011.

K. Kwak, K. Ji, Y. Kho, P. Kim, J. Lee, J. Ryu, and K. Choi. Chronic toxicity and endocrine disruption of naproxen in freshwater waterfleas and fish, and steroidogenic alteration using H295R cell assay. *Chemosphere*, 204:156–162, 2018. ISSN 18791298. DOI: 10.1016/J.CHEMOSPHERE.2018.04.035.

C. Lai, D. Martinović-Weigelt, A. Serrao De Filippo, S. Krämer, and C. Poschen. Extracting Semantics of Predicates From Millions of Bio-Medical Abstracts for Inferencing New Biological Key Events and Relationships. In *The 5th International Workshop on Deep Learning in Bioinformatics, Biomedicine, and Healthcare Informatics*, 2021. (Accepted on 21.10.2021).

K. S. Lakshmi and G. Vadivu. A novel approach for disease comorbidity prediction using weighted association rule mining. *Journal of Ambient Intelligence and Humanized Computing*, 0(0):0, 2019. ISSN 18685145. DOI: 10.1007/s12652-019-01217-1.

C. A. Lalone, D. L. Villeneuve, J. A. Doering, B. R. Blackwell, T. R. Transue, C. W. Simmons, J. Swintek, S. J. Degitz, A. J. Williams, and G. T. Ankley. Evidence for Cross Species Extrapolation of Mammalian-Based High-Throughput Screening Assay Results. *Environmental Science and Technology*, 52(23):13960–13971, 2018. ISSN 15205851. DOI: 10.1021/acs.est.8b04587.

P. Langfelder and S. Horvath. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, 9, 2008. ISSN 14712105. DOI: 10.1186/1471-2105-9-559.

P. Langfelder and S. Horvath. Fast R Functions for Robust Correlations and Hierarchical Clustering. *Journal of Statistical Software*, 46(11), 2012. DOI: 10.18637/jss.v046.i11.

P. Langfelder and S. Horvath. WGCNA package FAQ, 2017. URL `https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA`.

P. Langfelder, B. Zhang, and S. Horvath. Dynamic Tree Cut : in-depth description , tests and applications. *Bioinformatics*, 24(5):1–12, 2007. URL `http://bioinformatics.oxfordjournals.org/cgi/content/abstract/24/5/719`.

J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020. DOI: 10.1093/bioinformatics/btz682.

K. Lee, H. Schoenfuss, L. Barber, V. Writer, J.H.and Blazer, R. Kiesling, and M. Ferrey. Endocrine active chemicals and endocrine disruption in minnesota streams and lakes – implications for aquatic resources, 19942008, 2010. URL `https://pubs.usgs.gov/sir/2010/5107/`.

L.-H. Lee, Y. Lu, P.-H. Chen, P.-L. Lee, and K.-K. Shyu. NCUEE at MEDIQA 2019: Medical Text Inference Using Ensemble BERT-BiLSTM-Attention Model. In *Proceedings of the BioNLP 2019 workshop*, pages 528–532, 2019. DOI: 10.18653/v1/w19-4426.

J. T. Leek, E. Monsen, A. R. Dabney, and J. D. Storey. EDGE: Extraction and analysis of differential gene expression. *Bioinformatics*, 22(4):507–508, 2006. ISSN 13674803. DOI: 10.1093/bioinformatics/btk005.

H. Li. Microbiome, Metagenomics, and High-Dimensional Compositional Data Analysis. *Annual Review of Statistics and Its Application*, 2:73–94, 2015. ISSN 2326831X. DOI: 10.1146/annurev-statistics-010814-020351.

R. Li, A. Zupanic, M. Talikka, V. Belcastro, S. Madan, J. Dörpinghaus, C. vom Berg, J. Szostak, F. Martin, M. C. Peitsch, and J. Hoeng. Systems Toxicology Approach for Testing Chemical Cardiotoxicity in Larval Zebrafish. *Chemical Research in Toxicology*, 33(10):2550–2564, 2020. ISSN 15205010. DOI: 10.1021/acs.chemrestox.0c00095.

X. Liang, M. Wang, X. Chen, J. Zha, H. Chen, L. Zhu, and Z. Wang. Endocrine disrupting effects of benzotriazole in rare minnow (Gobiocypris rarus) in a sex-dependent manner. *Chemosphere*, 112:154–162, 2014. ISSN 18791298. DOI: 10.1016/j.chemosphere.2014.03.106.

Y. Liao, J. Wang, E. J. Jaehnig, Z. Shi, and B. Zhang. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Research*, 47(W1):W199–W205, 05 2019. ISSN 0305-1048. DOI: 10.1093/NAR/GKZ401.

W. Lichtensteiger, C. Bassetti-Gaille, O. Faass, M. Axelstad, J. Boberg, S. Christiansen, H. Rehrauer, J. K. Georgijevic, U. Hass, A. Kortenkamp, and M. Schlumpf. Differential gene expression patterns in developing sexually dimorphic rat brain regions exposed to antiandrogenic, estrogenic, or complex endocrine disruptor mixtures: Glutamatergic synapses as target. *Endocrinology*, 156(4):1477–1493, 2015. ISSN 19457170. DOI: 10.1210/EN.2014-1504.

G. Limonta, A. Mancia, A. Benkhalqui, C. Bertolucci, L. Abelli, M. C. Fossi, and C. Panti. Microplastics induce transcriptional changes, immune response and behavioral alterations in adult zebrafish. *Scientific Reports*, 9(1):1–11, 2019. ISSN 20452322. DOI: 10.1038/S41598-019-52292-5.

W. Lin, Y. Yan, S. Ping, P. Li, D. Li, J. Hu, W. Liu, X. Wen, and Y. Ren. Metformin-Induced Epigenetic Toxicity in Zebrafish: Experimental and Molecular Dynamics Simulation Studies. *Environmental Science and Technology*, 2020. ISSN 0013-936X. DOI: 10.1021/ACS.EST.0C06052.

P. J. Lioy and K. R. Smith. A discussion of exposure science in the 21st century: A vision and a strategy. *Environmental Health Perspectives*, 121(4):405–409, 2013. ISSN 00916765. DOI: 10.1289/EHP.1206170.

I. Liška, F. Wagner, M. Sengl, K. Deutsch, and J. Slobodník. Joint Danube Survey 3: A Comprehensive Analysis of Danube Water Quality, 2015. URL `http://www.danubesurvey.org/results`.

Y. C. Liu, C. P. Cheng, and V. S. Tseng. Mining differential top-k co-expression patterns from time course comparative gene expression datasets. *BMC Bioinformatics*, 14(1):1–13, 2013. ISSN 14712105. DOI: 10.1186/1471-2105-14-230.

J. R. Loughery, J. R. Marentette, R. A. Frank, L. M. Hewitt, J. L. Parrott, and C. J. Martyniuk. Transcriptome Profiling in Larval Fathead Minnow Exposed to Commercial Naphthenic Acids and Extracts from Fresh and Aged Oil Sands Process-Affected Water. *Environmental Science and Technology*, 53(17):10435–10444, 2019. ISSN 15205851. DOI: 10.1021/ACS.EST.9B01493.

M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014. ISSN 1465-6906. DOI: 10.1186/s13059-014-0550-8.

T. Luechtefeld, D. Marsh, C. Rowlands, and T. Hartung. Machine learning of toxicological big data enables read-across structure activity relationships (RASAR) outperforming animal test reproducibility. *Toxicological Sciences*, 165(1):198–212, 2018. ISSN 10960929. DOI: 10.1093/TOXSCI/KFY152.

C. Lyu, B. Chen, Y. Ren, and D. Ji. Long short-term memory RNN for biomedical named entity recognition. *BMC Bioinformatics*, 18(1):1–11, 2017. ISSN 14712105. DOI: 10.1186/s12859-017-1868-5.

A. Maertens, V. Tran, A. Kleensang, and T. Hartung. Weighted Gene Correlation Network Analysis (WGCNA) Reveals Novel Transcription Factors Associated With Bisphenol A Dose-Response. *Frontiers in Genetics*, 9, 2018. ISSN 16648021. DOI: 10.3389/FGENE.2018.00508.

S. Mallik and Z. Zhao. TrapRM: Transcriptomic and proteomic rule mining using weighted shortest distance based multiple minimum supports for multi-omics dataset. In *Proceedings - 2017 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2017*, pages 2187–2194, Jan 2017. ISBN 9781509030491. DOI: 10.1109/BIBM.2017.8217997.

Y. Mao and K. W. Fung. Use of word and graph embedding to measure semantic relatedness between unified medical language system concepts. *Journal of the American Medical Informatics Association*, 27(10):1538–1546, 2020. ISSN 1527974X. DOI: 10.1093/JAMIA/OCAA136.

M. Martens, C. T. Evelo, and E. L. Willighagen. Providing adverse outcome pathways from the AOP-Wiki in semantic web format to increase usability and accessibility of the content . *ChemRxiv*, page 16, 2021. DOI: 10.26434/CHEMRXIV.13524191.V1.

D. Martinović-Weigelt, A. C. Mehinto, G. T. Ankley, N. D. Denslow, L. B. Barber, K. E. Lee, R. J. King, H. L. Schoenfuss, A. L. Schroeder, and D. L. Villeneuve. Transcriptomic effects-based monitoring for endocrine active chemicals: Assessing relative contribution of treated wastewater to downstream pollution. *Environmental Science and Technology*, 48(4):2385–2394, 2014. ISSN 0013936X. DOI: 10.1021/ES404027N.

C. Martins, K. Dreij, and P. M. Costa. The state-of-the art of environmental toxicogenomics: Challenges and perspectives of omics approaches directed to toxicant mixtures. *International Journal of Environmental Research and Public Health*, 16(23):1–16, 2019. ISSN 16604601. DOI: 10.3390/IJERPH16234718.

C. J. Mattingly, G. T. Colby, J. N. Forrest, and J. L. Boyer. The Comparative Toxicogenomics Database (CTD). *Environmental Health Perspectives*, 111(6):793, 2003. ISSN 00916765. DOI: 10.1289/EHP.6028.

S. McGovarin, T. Sultana, and C. Metcalfe. Biological Responses in Brook Trout (Salvelinus fontinalis) Caged Downstream from Municipal Wastewater Treatment Plants in the Credit River, ON, Canada. *Bulletin of Environmental Contamination and Toxicology*, 100(1):106–111, 2018. ISSN 14320800. DOI: 10.1007/s00128-017-2242-z.

G. Michalopoulos, Y. Wang, H. Kaka, H. Chen, and A. Wong. UmlsBERT: Clinical Domain Knowledge Augmentation of Contextual Embeddings Using the Unified Medical Language System Metathesaurus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1744–1753. Association for Computational Linguistics, Jun 2021. DOI: 10.18653/V1/2021.NAACL-MAIN.139.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space, 2013. URL `https://arxiv.org/abs/1301.3781`.

G. W. Miller and D. P. Jones. The nature of nurture: Refining the definition of the exposome. *Toxicological Sciences*, 137(1):1–2, 2014. ISSN 10966080. DOI: 10.1093/TOXSCI/KFT251.

A. L. Miracle, G. P. Toth, and D. L. Lattier. The Path from Molecular Indicators of Exposure to Describing Dynamic Biological Systems in an Aquatic Organism: Microarrays and the Fathead Minnow. *Ecotoxicology*, 12(6):457–462, 2003. ISSN 09639292. DOI: 10.1023/B:ECTX.0000003030.67752.04.

S. J. Moe, R. Wolf, L. Xie, W. G. Landis, N. Kotamäki, and K. E. Tollefsen. Quantification of an Adverse Outcome Pathway Network by Bayesian Regression and Bayesian Network Modeling. *Integrated Environmental Assessment and Management*, 17(1):147–164, 2021. ISSN 15513793. DOI: 10.1002/IEAM.4348.

V. K. Mootha, C. M. Lindgren, E. Karl-Frederik, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstrale, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, M. J. P., T. R. Golub, P. Tamajo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler, and L. C. Groop. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34:267–273, 2003. DOI: 10.1038/NG1180.

H. M. Mortensen, J. Senn, T. Levey, P. Langley, and A. J. Williams. The 2021 update of the EPA's adverse outcome pathway database. *Scientific Data*, 8(1):1–9, 2021. ISSN 20524463. DOI: 10.1038/S41597-021-00962-3.

J. Mower, D. Subramanian, N. Shang, and T. Cohen. Classification-by-Analogy: Using Vector Representations of Implicit Relationships to Identify Plausibly Causal Drug/Side-effect Relationships. *AMIA Annual Symposium proceedings. AMIA Symposium*, pages 1940–1949, Feb 2017. ISSN 1942597X. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5333205/`.

J. Mower, D. Subramanian, and T. Cohen. Learning predictive models of drug side-effect relationships from distributed representations of literature-derived semantic predications. *Journal of the American Medical Informatics Association*, 25(10):1339–1350, 2018. ISSN 1527974X. DOI: 10.1093/JAMIA/OCY077.

K. Nagata, T. Washio, Y. Kawahara, and A. Unami. Toxicity prediction from toxicogenomic data based on class association rule mining. *Toxicology Reports*, 1:1133–1142, 2014. ISSN 22147500. DOI: 10.1016/J.TOXREP.2014.10.014.

S. K. Nair, C. Eeles, C. Ho, G. Beri, E. Yoo, D. Tkachuk, A. Tang, P. Nijrabi, P. Smirnov, H. Seo, D. Jennen, and B. Haibe-Kains. ToxicoDB: an integrated database to mine and visualize large-scale toxicogenomic datasets. *Nucleic acids research*, 48(W1):W455–W462, 2020. ISSN 13624962. DOI: 10.1093/NAR/GKAA390.

National Research Council. *Exposure science in the 21st century: a vision and a strategy*. National Academies Press, 2012. URL `https://www.ncbi.nlm.nih.gov/books/NBK206806/`.

M. Niemira, F. Collin, A. Szalkowska, A. Bielska, K. Chwialkowska, J. Reszec, J. Niklinski, M. Kwasniewski, and A. Kretowski. Molecular signature of subtypes of non-small-cell lung cancer by large-scale transcriptional profiling: Identification of key modules and genes by weighted gene co-expression network analysis (WGCNA). *Cancers*, 12(1), 2020. ISSN 20726694. DOI: 10.3390/CANCERS12010037.

P. Nymark, L. Rieswijk, F. Ehrhart, N. Jeliazkova, G. Tsiliki, H. Sarimveis, C. T. Evelo, V. Hongisto, P. Kohonen, E. Willighagen, and R. C. Grafström. A Data Fusion Pipeline for Generating and Enriching Adverse Outcome Pathway Descriptions. *Toxicological sciences*, 162(1):264–275, 2018. ISSN 10960929. DOI: 10.1093/TOXSCI/KFX252.

N. O. Oki and S. W. Edwards. An integrative data mining approach to identifying adverse outcome pathway signatures. *Toxicology*, 350-352:49–61, 2016. ISSN 18793185. DOI: 10.1016/J.TOX.2016.04.004.

T. O'Malley, E. Bursztein, J. Long, F. Chollet, H. Jin, L. Invernizzi, et al. Keras Tuner. `https://github.com/keras-team/keras-tuner`, 2019.

M. Orešič, A. McGlinchey, C. E. Wheelock, and T. Hyötyläinen. Metabolic signatures of the exposomequantifying the impact of exposure to environmental chemicals on human health. *Metabolites*, 10(11):1–31, 2020. ISSN 22181989. DOI: 10.3390/METABO10110454.

A. Ornostay, A. M. Cowie, M. Hindle, C. J. Baker, and C. J. Martyniuk. Classifying chemical mode of action using gene networks and machine learning: A case study with the herbicide linuron. *Comparative Biochemistry and Physiology - Part D: Genomics and Proteomics*, 8(4):263–274, 2013. ISSN 18780407. DOI: 10.1016/j.cbd.2013.08.001.

L. Orsini, J. B. Brown, O. Shams Solari, D. Li, S. He, R. Podicheti, M. H. Stoiber, K. I. Spanier, D. Gilbert, M. Jansen, D. B. Rusch, M. E. Pfrender, J. K. Colbourne, M. J. Frilander, J. Kvist, E. Decaestecker, K. A. De Schamphelaere, and L. De Meester. Early transcriptional response pathways in Daphnia magna are coordinated in networks of crustacean-specific genes. *Molecular Ecology*, 27(4):886–897, 2018. ISSN 1365294X. DOI: 10.1111/mec.14261.

R. Oughtred, C. Stark, B.-J. Breitkreutz, J. Rust, L. Boucher, C. Chang, N. Kolas, L. ODonnell, G. Leung, R. McAdam, et al. The BioGRID interaction database: 2019 update. *Nucleic acids research*, 47(D1): D529–D541, 2019. DOI: 10.1093/nar/gky1079.

G. Pallocca and M. Leist. EU TOXRISK. *ALTEX - ALTERNATIVES TO ANIMAL EXPERIMENTATION*, 36(1):152–153, 2021.

T. Paracelsus. Die dritte Defension wegen des Schreibens der neuen Rezepte, Septem Defensiones 1538. *Werke Bd*, 2:510, 1965. URL `http://www.zeno.org/Philosophie/M/Paracelsus/Septem+Defensiones/Die+dritte+Defension+wegen+des+Schreibens+der+neuen+Rezepte`.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. URL `https://www.jmlr.org/papers/v12/pedregosa11a.html`.

S. Peng, R. You, H. Wang, C. Zhai, H. Mamitsuka, and S. Zhu. DeepMeSH: Deep semantic representation for improving large-scale MeSH indexing. *Bioinformatics*, 32(12):i70–i79, 2016. ISSN 14602059. DOI: 10.1093/bioinformatics/btw294.

Y. Peng, A. Rios, R. Kavuluru, and Z. Lu. Chemical-protein relation extraction with ensembles of SVM, CNN, and RNN models, 2018. ISSN 2331-8422. URL `http://arxiv.org/abs/1802.01255`.

Y. Peng, S. Yan, and Z. Lu. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy, 2019. Association for Computational Linguistics. DOI: 10.18653/v1/W19-5006.

J. Pennington, R. Socher, and M. C. D. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014. DOI: 10.3115/v1/D14-1162.

M. R. Pérez, A. S. Rossi, C. Bacchetta, Y. Elorriaga, P. Carriquiriborde, and J. Cazenave. In situ evaluation of the toxicological impact of a wastewater effluent on the fish Prochilodus lineatus: biochemical and histological assessment. *Ecological Indicators*, 84:345–353, 2018. ISSN 1470160X. DOI: 10.1016/j.ecolind.2017.09.004.

E. J. Perkins, J. K. Chipman, S. Edwards, T. Habib, F. Falciani, R. Taylor, G. Van Aggelen, C. Vulpe, P. Antczak, and A. Loguinov. Reverse engineering adverse outcome pathways. *Environmental Toxicology and Chemistry*, 30(1):22–38, 2011. ISSN 07307268. DOI: 10.1002/ETC.374.

E. J. Perkins, T. Habib, B. L. Escalon, J. E. Cavallin, L. Thomas, M. Weberg, M. N. Hughes, K. M. Jensen, M. D. Kahl, D. L. Villeneuve, G. T. Ankley, and N. Garcia-Reyero. Prioritization of Contaminants of Emerging Concern in Wastewater Treatment Plant Discharges Using Chemical:Gene Interactions in Caged Fish. *Environmental Science and Technology*, 51(15):8701–8712, 2017. ISSN 15205851. DOI: 10.1021/ACS.EST.7B01567.

M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, 2018. Association for Computational Linguistics. DOI: 10.18653/V1/N18-1202.

G. Piatetsky-Shapiro. Discovery, analysis and presentation of strong rules. In G. Piatetsky-Shapiro and W. J. Frawley, editors, *Knowledge Discovery in Databases*, pages 229–248. AAAI Press, 1991.

A. R. Pico, T. Kelder, M. P. Van Iersel, K. Hanspers, B. R. Conklin, and C. Evelo. WikiPathways: Pathway Editing for the People. *PLoS Biology*, 6(7):1403–1407, 2008. ISSN 15449173. DOI: 10.1371/JOURNAL.PBIO.0060184.

Y. V. Pinyaga, T. M. Prokopiv, A. V. Petrishin, O. V. Khalimonchuk, O. V. Protchenko, D. V. Fedorovich, and Y. R. Boretsky. Restoration of the wild-type phenotype in Pichia guilliermondii transformants. *Microbiology*, 71(3):314–318, 2002. ISSN 00262617. DOI: 10.1023/A:1015858712461.

M. E. Pittman, S. W. Edwards, C. Ives, and H. M. Mortensen. AOP-DB: A database resource for the exploration of Adverse Outcome Pathways through integrated association networks. *Toxicology and Applied Pharmacology*, 343:71–83, Feb 2018. ISSN 10960333. DOI: 10.1016/J.TAAP.2018.02.006.

N. Polavarapu, S. B. Navathe, R. Ramnarayanan, A. Ul Haque, S. Sahay, and Y. Liu. Investigation into biomedical literature classification using support vector machines. In *Proceedings - 2005 IEEE Computational Systems Bioinformatics Conference, CSB 2005*, pages 366–374, 2005. ISBN 0769523447. DOI: 10.1109/CSB.2005.36.

N. L. Pollesch, D. L. Villeneuve, and J. M. O'Brien. Extracting and Benchmarking Emerging Adverse Outcome Pathway Knowledge. *Toxicological Sciences*, 168(2):349–364, 2019. ISSN 10960929. DOI: 10.1093/TOXSCI/KFZ006.

T. G. Pottinger. Modulation of the stress response in wild fish is associated with variation in dissolved nitrate and nitrite. *Environmental Pollution*, 225:550–558, 2017. ISSN 18736424. DOI: 10.1016/J.ENVPOL.2017.03.021.

X. Qian, Y. Ba, Q. Zhuang, and G. Zhong. RNA-seq technology and its application in fish transcriptomics. *OMICS A Journal of Integrative Biology*, 18(2):98–110, 2014. ISSN 15362310. DOI: 10.1089/OMI.2013.0110.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org/.

Ragas, Ad. Chapter 1: Environmental Toxicology. In A. van Gestel Cornelis, F. G. Van Belleghem, N. W. van den Brink, S. T. J. Droge, T. Hamers, J. L. Hermens, M. H. Kraak, A. J. Löhr, J. R. Parsons, A. M. Ragas, N. M. van Straalen, and M. G. Vijver, editors, *Environmental Toxicology, an open online textbook*, page 834. Elsevier, 2021. URL https://research.ou.nl/en/publications/environmental-toxicology-an-open-online-textbook.

F. Rapaport, R. Khanin, Y. Liang, M. Pirun, A. Krek, P. Zumbo, C. E. Mason, N. D. Socci, and D. Betel. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biology*, 14(9), 2013. ISSN 1474760X. DOI: 10.1186/GB-2013-14-9-R95.

M. Rastegar-Mojarad, R. K. Elayavilli, L. Wang, R. Prasad, and H. Liu. Prioritizing adverse drug reaction and drug repositioning candidates generated by literature-based discovery. In *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 289–296, 2016. DOI: 10.1145/2975167.2975197.

T. C. Rindflesch and M. Fiszman. The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics*, 36(6):462–477, 2003. ISSN 15320464. DOI: 10.1016/J.JBI.2003.11.003.

T. C. Rindflesch, M. Fiszman, and L. Bisharah. Semantic interpretation for the biomedical research literature. In *Medical informatics: Knowledge Management and Data Mining in Biomedicine*, pages 399–422. Springer: New York, 2005. URL https://lhncbc.nlm.nih.gov/LHC-publications/pubs/SemanticInterpretationfortheBiomedicalResearchLiterature.html.

M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 2015. ISSN 13624962. DOI: 10.1093/NAR/GKV007.

I. A. Rodriguez-Jorquera, K. J. Kroll, G. S. Toor, and N. D. Denslow. Transcriptional and physiological response of fathead minnows (Pimephales promelas) exposed to urban waters entering into wildlife protected areas. *Environmental Pollution*, 199:155–165, 2015. ISSN 18736424. DOI: 10.1016/J.ENVPOL.2015.01.021.

I. A. Rodríguez-Jorquera, R. C. Colli-Dula, K. Kroll, B. S. Jayasinghe, M. V. Parachu Marco, C. Silva-Sanchez, G. S. Toor, and N. D. Denslow. Blood Transcriptomics Analysis of Fish Exposed to Perfluoro Alkyls Substances: Assessment of a Non-Lethal Sampling Technique for Advancing Aquatic Toxicology Research. *Environmental Science and Technology*, 53(3):1441–1452, 2019. ISSN 15205851. DOI: 10.1021/ACS.EST.8B03603.

Roelofs, Dick. Chapter 4.2.13: Adverse outcome pathways. In A. van Gestel Cornelis, F. G. Van Belleghem, N. W. van den Brink, S. T. J. Droge, T. Hamers, J. L. Hermens, M. H. Kraak, A. J. Löhr, J. R. Parsons, A. M. Ragas, N. M. van Straalen, and M. G. Vijver, editors, *Environmental Toxicology, an open online textbook*, pages 481–486. Elsevier, 2021. URL https://research.ou.nl/en/publications/environmental-toxicology-an-open-online-textbook.

G. Rosemblat, M. P. Resnick, I. Auston, D. Shin, C. Sneiderman, M. Fizsman, and T. C. Rindflesch. Extending SemRep to the public health domain. *Journal of the American Society for Information Science and Technology*, 64(10):1963–1974, 2013a. DOI: 10.1002/ASI.22899.

G. Rosemblat, D. Shin, H. Kilicoglu, C. Sneiderman, and T. C. Rindflesch. A methodology for extending domain coverage in SemRep. *Journal of Biomedical Informatics*, 46(6):1099–1107, 2013b. ISSN 15320464. DOI: 10.1016/j.jbi.2013.08.005.

M. Rotmensch, Y. Halpern, A. Tlimat, S. Horng, and D. Sontag. Learning a health knowledge graph from electronic medical records. *Scientific reports*, 7(1):1–11, 2017. DOI: 10.1038/s41598-017-05778-z.

M. Rugard, X. Coumoul, J. C. Carvaillo, R. Barouki, and K. Audouze. Deciphering Adverse Outcome Pathway Network Linked to Bisphenol F Using Text Mining and Systems Toxicology Approaches. *Toxicological Sciences*, 173(1):32–40, 2020. ISSN 10960929. DOI: 10.1093/toxsci/kfz214.

SAAOP. Aop-wiki, 2021. URL `http://aopwiki.org`. last accessed on 01.12.2021.

S. Santos, L. Maitre, C. Warembourg, L. Agier, L. Richiardi, X. Basagaña, and M. Vrijheid. Applying the exposome concept in birth cohort research: a review of statistical approaches. *European Journal of Epidemiology*, 35(3):193–204, 2020. ISSN 15737284. DOI: 10.1007/s10654-020-00625-4.

S. Scholz, J. W. Nichols, B. I. Escher, G. T. Ankley, R. Altenburger, B. Blackwell, W. Brack, L. Burkhard, T. W. Collette, J. A. Doering, D. Ekman, K. Fay, F. Fischer, J. Hackermüller, J. C. Hoffman, C. Lai, D. Leuthold, D. MartinovicWeigelt, T. Reemtsma, N. Pollesch, A. Schroeder, G. Schüürmann, and M. Bergen. The EcoExposome concept: Supporting an Integrated Assessment of Mixtures of Environmental Chemicals. *Environmental Toxicology and Chemistry*, 2021. ISSN 0730-7268. DOI: 10.1002/etc.5242.

A. L. Schroeder, G. T. Ankley, K. A. Houck, and D. L. Villeneuve. Environmental surveillance and monitoring-The next frontiers for high-throughput toxicology. *Environmental Toxicology and Chemistry*, 35(3):513–525, 2016. ISSN 15528618. DOI: 10.1002/etc.3309.

A. L. Schroeder, D. Martinović-Weigelt, G. T. Ankley, K. E. Lee, N. Garcia-Reyero, E. J. Perkins, H. L. Schoenfuss, and D. L. Villeneuve. Prior knowledge-based approach for associating contaminants with biological effects: A case study in the St. Croix River basin, MN, WI, USA. *Environmental Pollution*, 221: 427–436, 2017. ISSN 18736424. DOI: 10.1016/j.envpol.2016.12.005.

A. Schüttler, K. Reiche, R. Altenburger, and W. Busch. The transcriptome of the zebrafish embryo after chemical exposure: A meta-analysis. *Toxicological Sciences*, 157(2):291–304, 2017. ISSN 10960929. DOI: 10.1093/toxsci/kfx045.

A. Schüttler, R. Altenburger, M. Ammar, M. Bader-Blukott, G. Jakobs, J. Knapp, J. Krüger, K. Reiche, G.-M. Wu, and W. Busch. Map and model - moving from observation to prediction in toxicogenomics. *GigaScience*, 8(6), 05 2019. ISSN 2047-217X. DOI: 10.1093/gigascience/giz057.

M. K. Sellin Jeffries, A. C. Mehinto, B. J. Carter, N. D. Denslow, and A. S. Kolok. Taking microarrays to the field: Differential hepatic gene expression of caged fathead minnows from Nebraska watersheds. *Environmental Science and Technology*, 46(3):1877–1885, 2012. ISSN 0013936X. DOI: 10.1021/es2039097.

A. Sergushichev. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *bioRxiv*, 2016. DOI: 10.1101/060012. URL `http://biorxiv.org/content/early/2016/06/20/060012`.

J. Shendure. The beginning of the end for microarrays? *Nature Methods*, 5(7):585–587, 2008. ISSN 15487091. DOI: 10.1038/nmeth0708-585.

J. Shi and M. Walker. Gene Set Enrichment Analysis (GSEA) for Interpreting Gene Expression Profiles. *Current Bioinformatics*, 2(2):133–137, 2008. ISSN 15748936. DOI: 10.2174/157489307780618231.

Z. Q. Shi, Y. S. Liu, Q. Xiong, W. W. Cai, and G. G. Ying. Occurrence, toxicity and transformation of six typical benzotriazoles in the environment: A review. *Science of the Total Environment*, 661:407–421, 2019. ISSN 18791026. DOI: 10.1016/j.scitotenv.2019.01.138.

T. Simões, S. C. Novais, T. Natal-da Luz, B. Devreese, T. de Boer, D. Roelofs, J. P. Sousa, N. M. van Straalen, and M. F. Lemos. An integrative omics approach to unravel toxicity mechanisms of environmental chemicals: effects of a formulated herbicide. *Scientific Reports*, 8(1):1–12, 2018. ISSN 20452322. DOI: 10.1038/s41598-018-29662-6.

D. M. Skelton, D. R. Ekman, D. Martinović-Weigelt, G. T. Ankley, D. L. Villeneuve, Q. Teng, and T. W. Collette. Metabolomics for in situ environmental monitoring of surface waters impacted by contaminants from both point and nonpoint sources. *Environmental Science and Technology*, 48(4):2395–2403, 2014. ISSN 0013936X. DOI: 10.1021/es404021f.

N. R. Smalheiser and G. Bonifield. Unsupervised low-dimensional vector representations for words, phrases and text that are transparent, scalable, and produce similarity metrics that are complementary to neural embeddings, 2018. ISSN 23318422. URL `https://arxiv.org/abs/1801.01884`.

T. W. Snell, S. E. Brogdon, and M. B. Morgan. Gene expression profiling in ecotoxicology. *Ecotoxicology*, 12: 529–475–483, 2003. DOI: 10.1007/978-1-4614-7357-2_36.

J. O. Straub, D. J. Caldwell, T. Davidson, V. D'Aco, K. Kappler, P. F. Robinson, B. Simon-Hettich, and J. Tell. Environmental risk assessment of metformin and its transformation product guanylurea. I. Environmental fate. *Chemosphere*, 216:844–854, 2019. ISSN 18791298. DOI: 10.1016/j.chemosphere.2018.10.036.

R. Su, H. Wu, B. Xu, X. Liu, and L. Wei. Developing a multi-dose computational model for drug-induced hepatotoxicity prediction based on toxicogenomics data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(4):1231–1239, 2019. ISSN 15579964. DOI: 10.1109/TCBB.2018.2858756.

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 102, pages 15545–15550, 2005. DOI: 10.1073/pnas.0506580102.

A. Subramanian, R. Narayan, S. M. Corsello, D. D. Peck, T. E. Natoli, X. Lu, J. Gould, J. F. Davis, A. A. Tubelli, J. K. Asiedu, D. L. Lahr, J. E. Hirschman, Z. Liu, M. Donahue, B. Julian, M. Khan, D. Wadden, I. C. Smith, D. Lam, A. Liberzon, C. Toder, M. Bagul, M. Orzechowski, O. M. Enache, F. Piccioni, S. A. Johnson, N. J. Lyons, A. H. Berger, A. F. Shamji, A. N. Brooks, A. Vrcic, C. Flynn, J. Rosains, D. Y. Takeda, R. Hu, D. Davison, J. Lamb, K. Ardlie, L. Hogstrom, P. Greenside, N. S. Gray, P. A. Clemons, S. Silver, X. Wu, W. N. Zhao, W. Read-Button, X. Wu, S. J. Haggarty, L. V. Ronco, J. S. Boehm, S. L. Schreiber, J. G. Doench, J. A. Bittker, D. E. Root, B. Wong, and T. R. Golub. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell*, 171(6):1437–1452.e17, 2017. ISSN 10974172. DOI: 10.1016/j.cell.2017.10.049.

J. J. Sutherland, Y. W. Webster, J. A. Willy, G. H. Searfoss, K. M. Goldstein, A. R. Irizarry, D. G. Hall, and J. L. Stevens. Toxicogenomic module associations with pathogenesis: A network-based approach to understanding drug toxicity. *Pharmacogenomics Journal*, 18(3):377–390, 2018. ISSN 14731150. DOI: 10.1038/TPJ.2017.17.

D. R. Swanson. Fish oil, raynaud's syndrome, and undiscovered public knowledge. *Perspectives in biology and medicine*, 30(1):7–18, 1986. DOI: 10.1353/PBM.1986.0087.

D. Szklarczyk, A. Santos, C. Von Mering, L. J. Jensen, P. Bork, and M. Kuhn. STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data. *Nucleic acids research*, 44(D1):D380–D384, 2016. DOI: 10.1093/NAR/GKV1277.

D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork, L. J. Jensen, and C. Von Mering. STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1):D607–D613, 2019. ISSN 13624962. DOI: 10.1093/NAR/GKY1131.

D. Szklarczyk, A. L. Gable, K. C. Nastou, D. Lyon, R. Kirsch, S. Pyysalo, N. T. Doncheva, M. Legeay, T. Fang, P. Bork, L. J. Jensen, and C. von Mering. The STRING database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research*, 49(D1):D605–D612, 2021. ISSN 13624962. DOI: 10.1093/NAR/GKAA1074.

O. Taboureau, U. P. Jacobsen, C. Kalhauge, D. Edsgärd, O. Rigina, R. Gupta, and K. Audouze. HExpoChem: a systems biology resource to explore human exposure to chemicals. *Bioinformatics (Oxford, England)*, 29 (9):1231–1232, 2013. ISSN 13674811. DOI: 10.1093/BIOINFORMATICS/BTT112.

O. Taboureau, W. El M'Selmi, and K. Audouze. Integrative systems toxicology to predict human biological systems affected by exposure to environmental chemicals. *Toxicology and Applied Pharmacology*, 405:115210, Aug 2020. ISSN 10960333. DOI: 10.1016/J.TAAP.2020.115210.

G. J. Tawa, M. D. M. AbdulHameed, X. Yu, K. Kumar, D. L. Ippolito, J. A. Lewis, J. D. Stallings, and A. Wallqvist. Characterization of chemically induced liver injuries using gene co-expression modules. *PLoS ONE*, 9(9), 2014. ISSN 19326203. DOI: 10.1371/JOURNAL.PONE.0107230.

M. A. Thomas, L. Yang, B. J. Carter, and R. D. Klaper. Gene set enrichment analysis of microarray data from Pimephales promelas (Rafinesque), a non-mammalian model organism. *BMC Genomics*, 12, 2011. ISSN 14712164. DOI: 10.1186/1471-2164-12-66.

P. J. Thul, L. Akesson, M. Wiking, D. Mahdessian, A. Geladaki, H. Ait Blal, T. Alm, A. Asplund, L. Björk, L. M. Breckels, A. Bäckström, F. Danielsson, L. Fagerberg, J. Fall, L. Gatto, C. Gnann, S. Hober, M. Hjelmare, F. Johansson, S. Lee, C. Lindskog, J. Mulder, C. M. Mulvey, P. Nilsson, P. Oksvold, J. Rockberg, R. Schutten, J. M. Schwenk, A. Sivertsson, E. Sjöstedt, M. Skogs, C. Stadler, D. P. Sullivan, H. Tegel, C. Winsnes, C. Zhang, M. Zwahlen, A. Mardinoglu, F. Pontén, K. Von Feilitzen, K. S. Lilley, M. Uhlén, and E. Lundberg. A subcellular map of the human proteome. *Science*, 356(6340), 2017. ISSN 10959203. DOI: 10.1126/SCIENCE.AAL3321.

Z. Tian, W. He, J. Tang, X. Liao, Q. Yang, Y. Wu, and G. Wu. Identification of important modules and biomarkers in breast cancer based on WGCNA. *OncoTargets and Therapy*, 13:6805–6817, 2020. ISSN 11786930. DOI: 10.2147/OTT.S258439.

G. Toti, R. Vilalta, P. Lindner, B. Lefer, C. Macias, and D. Price. Analysis of correlation between pediatric asthma exacerbation and exposure to pollutant mixtures with association rule mining. *Artificial Intelligence in Medicine*, 74:44–52, 2016. ISSN 18732860. DOI: 10.1016/J.ARTMED.2016.11.003.

U.S. Environmental Protection Agency. ECOTOX User Guide: ECOTOXicology Knowledgebase System, 2021. URL `https://cfpub.epa.gov/ecotox/`. Version 5.3.

S. van Dam, U. Võsa, A. van der Graaf, L. Franke, and J. P. de Magalhães. Gene co-expression analysis for functional classification and genedisease predictions. *Briefings in Bioinformatics*, 19, Jul 2018. ISSN 1467-5463. DOI: 10.1093/BIB/BBW139.

van Duursen, Majorie. Chapter 4.2.8: Endocrine Disruption. In A. van Gestel Cornelis, F. G. Van Belleghem, N. W. van den Brink, S. T. J. Droge, T. Hamers, J. L. Hermens, M. H. Kraak, A. J. Löhr, J. R. Parsons, A. M. Ragas, N. M. van Straalen, and M. G. Vijver, editors, *Environmental Toxicology, an open online textbook*, pages 481–486. Elsevier, 2021. URL `https://research.ou.nl/en/publications/environmental-toxicology-an-open-online-textbook`.

G. Van Rossum and F. L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697.

van Straalen, Nico M. Chapter 4.1.4: Xenobiotic defence and metabolism. In A. van Gestel Cornelis, F. G. Van Belleghem, N. W. van den Brink, S. T. J. Droge, T. Hamers, J. L. Hermens, M. H. Kraak, A. J. Löhr, J. R. Parsons, A. M. Ragas, N. M. van Straalen, and M. G. Vijver, editors, *Environmental Toxicology, an open online textbook*, pages 351–359. Elsevier, 2021a. URL `https://research.ou.nl/en/publications/environmental-toxicology-an-open-online-textbook`.

van Straalen, Nico M. Chapter 4.3.12: Gene Expression. In A. van Gestel Cornelis, F. G. Van Belleghem, N. W. van den Brink, S. T. J. Droge, T. Hamers, J. L. Hermens, M. H. Kraak, A. J. Löhr, J. R. Parsons, A. M. Ragas, N. M. van Straalen, and M. G. Vijver, editors, *Environmental Toxicology, an open online textbook*, pages 481–486. Elsevier, 2021b. URL `https://research.ou.nl/en/publications/environmental-toxicology-an-open-online-textbook`.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, G. A. N., K. Lukasz, and I. Polosukhin. Attention is all you need. In *31st Conference on Neural Information Processing Systems*, pages 1–11, 2017. DOI: 10.1109/2943.974352.

D. E. Vidal-Dorsch, R. C. Colli-Dula, S. M. Bay, D. J. Greenstein, L. Wiborg, D. Petschauer, and N. D. Denslow. Gene expression of fathead minnows (Pimephales promelas) exposed to two types of treated municipal wastewater effluents. *Environmental Science and Technology*, 47(19):11268–11277, 2013. ISSN 0013936X. DOI: 10.1021/ES401942N.

D. L. Villeneuve, D. Crump, N. Garcia-Reyero, M. Hecker, T. H. Hutchinson, C. A. LaLone, B. Landesmann, T. Lettieri, S. Munn, M. Nepelska, M. A. Ottinger, L. Vergauwen, and M. Whelan. Adverse outcome pathway (AOP) development I: Strategies and principles. *Toxicological Sciences*, 142(2):312–320, 2014. ISSN 10960929. DOI: 10.1093/TOXSCI/KFU199.

D. L. Villeneuve, M. M. Angrish, M. C. Fortin, I. Katsiadaki, M. Leonard, L. Margiotta-Casaluci, S. Munn, J. M. O'Brien, N. L. Pollesch, L. C. Smith, X. Zhang, and D. Knapen. Adverse outcome pathway networks

II: Network analytics. *Environmental Toxicology and Chemistry*, 37(6):1734–1748, 2018. ISSN 15528618. DOI: 10.1002/ETC.4124.

H. Wang, R. Liu, P. Schyman, and A. Wallqvist. Deep neural network models for predicting chemically induced liver toxicity endpoints from transcriptomic responses. *Frontiers in Pharmacology*, 10:42, 2019. ISSN 1663-9812. DOI: 10.3389/FPHAR.2019.00042.

J. Wang and Y. Liao. *WebGestaltR: Gene Set Analysis Toolkit WebGestaltR*, 2020. URL `https://CRAN.R-project.org/package=WebGestaltR`. R package version 0.4.3.

P. Wang, P. Xia, Z. Wang, and X. Zhang. Evidence-based assessment on environmental mixture using a concentration-dependent transcriptomics approach. *Environmental Pollution*, 265, 2020. ISSN 18736424. DOI: 10.1016/J.ENVPOL.2020.114839.

R. L. Wang, A. D. Biales, N. Garcia-Reyero, E. J. Perkins, D. L. Villeneuve, G. T. Ankley, and D. C. Bencic. Fish connectivity mapping: Linking chemical stressors by their mechanisms of action-driven transcriptomic profiles. *BMC Genomics*, 17(1), 2016. ISSN 14712164. DOI: 10.1186/S12864-016-2406-Y.

Y. Wang, S. Liu, N. Afzal, M. Rastegar-Mojarad, L. Wang, F. Shen, P. Kingsbury, and H. Liu. A comparison of word embeddings for the biomedical natural language processing. *Journal of Biomedical Informatics*, 87: 12–20, Jul 2018. ISSN 15320464. DOI: 10.1016/J.JBI.2018.09.008.

S. M. Watford, R. G. Grashow, V. Y. D. L. Rosa, R. A. Rudel, K. P. Friedman, and M. T. Martin. Novel application of normalized pointwise mutual information (NPMI) to mine biomedical literature for gene sets associated with disease: use case in breast carcinogenesis. *Computational Toxicology*, 7:46–57, 2018. DOI: 10.1016/J.COMTOX.2018.06.003.

J. D. Watson and F. H. C. Crick. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature 1953 171:4356*, 171(4356):737–738, apr 1953. ISSN 1476-4687. DOI: 10.1038/171737A0.

W.-Q. Wei, R. M. Cronin, H. Xu, T. A. Lasko, L. Bastarache, and J. C. Denny. Development and evaluation of an ensemble resource linking medications to their indications. *Journal of the American Medical Informatics Association*, 20(5):954–961, 2013. DOI: 10.1136/AMIAJNL-2012-001431.

R. R. Wilcox. *Introduction to Robust Estimation and Hypothesis Testing: 4th Edition*. Elsevier Academic Press, 2016. ISBN 9780128047330. DOI: 10.1016/C2010-0-67044-1.

C. P. Wild. Complementing the genome with an "exposome": The outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiology Biomarkers and Prevention*, 14(8): 1847–1850, 2005. ISSN 10559965. DOI: 10.1158/1055-9965.EPI-05-0456.

C. P. Wild. The exposome: From concept to utility. *International Journal of Epidemiology*, 41(1):24–32, 2012. ISSN 03005771. DOI: 10.1093/IJE/DYR236.

T. D. Williams, N. Turan, A. M. Diab, H. Wu, C. Mackenzie, K. L. Bartie, O. Hrydziuszko, B. P. Lyons, G. D. Stentiford, J. M. Herbert, J. K. Abraham, I. Katsiadaki, M. J. Leaver, J. B. Taggart, S. G. George, M. R. Viant, K. J. Chipman, and F. Falciani. Towards a system level understanding of non-model organisms sampled from the environment: A network biology approach. *PLoS Computational Biology*, 7(8):e1002126, 2011. ISSN 1553734X. DOI: 10.1371/JOURNAL.PCBI.1002126.

H. Wirth, M. Löffler, M. von Bergen, and H. Binder. Expression cartography of human tissues using self organizing maps. *BMC Bioinformatics*, 12, 2011. ISSN 14712105. DOI: 10.1186/1471-2105-12-306.

S. B. Wiseman, Y. He, M. Gamal-El Din, J. W. Martin, P. D. Jones, M. Hecker, and J. P. Giesy. Transcriptional responses of male fathead minnows exposed to oil sands process-affected water. *Comparative Biochemistry and Physiology - C Toxicology and Pharmacology*, 157(2):227–235, 2013. ISSN 18781659. DOI: 10.1016/J.CBPC.2012.12.002.

C. Wu, R. Carta, and L. Zhang. Sequence dependence of cross-hybridization on short oligo microarrays. *Nucleic Acids Research*, 33(9):e84–e84, 01 2005. ISSN 0305-1048. DOI: 10.1093/NAR/GNI082.

A. M. Yip and S. Horvath. Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics*, 8:1–14, 2007. ISSN 14712105. DOI: 10.1186/1471-2105-8-22.

A. E. Yosim and R. Fry. *Systems biology in toxicology and environmental health*. Haley Mica, Academic Press, 2015. URL `https://www.elsevier.com/books/systems-biology-in-toxicology-and-environmental-health/fry/978-0-12-801564-3`.

Z. Yuxuan, G. Peiyong, W. Yanmei, Z. Xiaoyan, W. Meixian, Y. Simin, S. Yinshi, D. Jun, and S. Haitao. Evaluation of the subtle effects and oxidative stress response of chloramphenicol, thiamphenicol, and florfenicol in Daphnia magna. *Environmental Toxicology and Chemistry*, 38(3):575–584, 2019. ISSN 15528618. DOI: 10.1002/ETC.4344.

A. Zare, D. Henry, G. Chua, P. Gordon, and H. R. Habibi. Differential Hepatic Gene Expression Profile of Male Fathead Minnows Exposed to Daily Varying Dose of Environmental Contaminants Individually and in Mixture. *Frontiers in Endocrinology*, 9, Dec 2018. ISSN 16642392. DOI: 10.3389/FENDO.2018.00749.

E. Zgheib, M. J. Kim, F. Jornod, K. Bernal, C. Tomkiewicz, S. Bortoli, X. Coumoul, R. Barouki, K. De Jesus, E. Grignard, P. Hubert, E. S. Katsanou, F. Busquet, and K. Audouze. Identification of non-validated endocrine disrupting chemical characterization methods by screening of the literature using artificial intelligence and by database exploration. *Environment International*, 154, 2021. ISSN 18736750. DOI: 10.1016/J.ENVINT.2021.106574.

B. Zhang and S. Horvath. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1), 2005. URL `10.2202/1544-6115.1128`.

D. Zhang, D. He, N. Zou, X. Zhou, and F. Pei. Automatic Relationship Verification in Online Medical Knowledge Base: A Large Scale Study in SemMedDB. In *Proceedings - 2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018*, pages 1673–1680, 2019a. ISBN 9781538654880. DOI: 10.1109/BIBM.2018.8621316.

Y. Zhang, Q. Chen, Z. Yang, H. Lin, and Z. Lu. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific Data*, 6(1):1–10, 2019b. ISSN 20524463. DOI: 10.1038/s41597-019-0055-0.

D. Zhao, J. Wang, H. Lin, Z. Yang, and Y. Zhang. Extracting drugdrug interactions with hybrid bidirectional gated recurrent unit and graph convolutional network. *Journal of Biomedical Informatics*, 99:103295, Sep 2019. ISSN 15320464. DOI: 10.1016/J.JBI.2019.103295.

S. Zhao, C. Su, Z. Lu, and F. Wang. Recent advances in biomedical literature mining. *Briefings in bioinformatics*, 22(3):1–19, 2021. ISSN 14774054. DOI: 10.1093/BIB/BBAA057.

W. Zhao, P. Langfelder, T. Fuller, J. Dong, A. Li, and S. Hovarth. Weighted gene coexpression network analysis: State of the art. *Journal of Biopharmaceutical Statistics*, 20(2):281–300, 2010. ISSN 10543406. DOI: 10.1080/10543400903572753.

Curriculum Scientiae

Education:

since 03/2022	Scientific employee at the Federal Environmental Agency Germany
10/2018 – 10/2021	PhD student at the University of Leipzig and Helmholtz-Centre for Environmental Research UFZ, Germany
	• Supervised by Prof. Dr. Jörg Hackermüller and Prof. Peter F. Stadler, Chair of Bioinformatics
	• Thesis: *Computationally Linking Chemical Exposure to Molecular Effects with Complex Data: Comparing Methods to Disentangle Chemical Drivers in Environmental Mixtures and Knowledge-based Deep Learning for Predictions in Environmental Toxicology*
10/2015 – 05/2018	Master student at the University of Leipzig, Germany
	• Master of Science Bioinformatics
	• Thesis: *The ecotoxicogenomic (self-organising) Map of Fathead Minnow*
10/2013 – 10/2015	Master student at the University of Leipzig, Germany
	• Master of Education in Physics and Biology
	• Thesis: *Biogeography of Impatiens L. in South-East-Asia*
09/2013 – 09/2016	Bachelor student at the University of Leipzig, Germany
	• Bachelor of Science in Physics and Biology (teaching qualification)
	• Thesis: *Phylogeny of Gentiana L. in New Guinea*

Practical Training:

08/2017 – 07/2018	Research assistant at the Helmholtz-Centre for Environmental Research UFZ, Germany • Department Bioanalytical Ecotoxicology - working group Dr. Wibke Busch • Topic: *Toxicogenomic profiling with the self-organising map approach for meta-analyses in aquatic environmental model organisms*

IT-Knowledge:

Operating systems:	UNIX, Mac, Linux, Windows
Programming:	Python, R
Markup languages:	Latex
Database languages:	SQL

Language Skills:

German:	native speaker
English:	fluent
French:	basic knowledge

Publications:

Krämer, S., Busch, W., & Schüttler, A. (2020). A Self-Organizing Map of the Fathead Minnow Liver Transcriptome to Identify Consistent Toxicogenomic Patterns across Chemical Fingerprints. Environmental toxicology and chemistry, 39(3), 526-537.

Lai, C., Martinović-Weigelt, D., Serrao De Filippo, A., Krämer, Stefan and Poschen, C. (2021). **Extracting Semantics of Predicates From Millions of Bio-Medical Abstracts for Inferencing New Biological Key Events and Relationships.** The 5th International Workshop on Deep Learning in Bioinformatics, Biomedicine, and Healthcare Informatics (Accepted on 21.10.2021).

Krämer, S., Lai, C., Martinović, Busch, W., Schor, J. & Hackermüller, J. **A method comparison to disentangle chemical specific effects for fish exposed to environmental mixtures.** In preparation.

Krämer, S., Lai, C., Schor, J. & Hackermüller, J. **Evaluation of predicted chemical-biomolecule interactions as Molecular Initiating-like Events with recent toxicological knowledge in literature and databases.** In preparation.

Conferences / Seminars:

Seminar PhD-Colleg 'Proxies of the Eco-Exposome' (Presenter)
Krämer, S.: *Update report project T4: Linking chemical exposure to biological effects*
04/2021; Leipzig, Germany

Department Seminar Molecular Systems Biology (Presenter)
Krämer, S.: *Discover transcriptional effects of single chemicals in complex mixtures*
05/2019; Leipzig, Germany

SETAC Annual Meeting Europe 2020 (Presenter)
Krämer, S.: *Predicting key events and key event relationships - a knowledge based deep learning approach*
05/2020; Online

IP Exposome Plenary Meeting (Presenter)
Scholz S., Krämer, S., and Dann J.: *PhD-Colleg 'Proxies of the Eco-Exposome'*
10/2019; Leipzig, Germany

17th Bioinformatik Herbstseminar (Presenter)
Krämer, S.: *Link Biological Effects to Chemical Exposure*
10/2019; Doubice, Czech Republic

Mittelerde Meeting 2019 (Poster)
Krämer, S., Schor J., and Hackermüller J.: *Linking Exposure Data to transcriptomic Effects for the aquatic environment*
06/2019; Dresden, Germany

Department Seminar Molecular Systems Biology (Presenter)
Krämer, S.: *Linking Chemical Exposure to Biological Effects*
05/2019; Leipzig, Germany

SETAC Annual Meeting Europe 2019 (Presenter, Poster)
Krämer, S.: *A machine learning approach for transcriptomic data in ecotoxicology to improve consistency of microarray sample sets of chemically treated fathead minnows*
05/2019; Helsinki, Finnland

Department Seminar Molecular Systems Biology (Presenter)
Krämer, S.: *Proxies of the Eco-Exposome: Link omics to exposure data*
10/2018; Leipzig, Germany

IP Exposome Plenary Meeting (Presenter)
Scholz S., Jakobs, G., Wernicke, T., Krämer, S., and Dann J.: *Introducing the PhD-Colleg 'Proxies of the Eco-Exposome'*
10/2018; Leipzig, Germany

Seminar PhD-Colleg 'Proxies of the Eco-Exposome' (Presenter)
Krämer, S.: *WGCNA for exposure-related microarray samples of 10 streams in Minnesota*
11/2018; Leipzig, Germany

Selbständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialen oder erbrachten Dienstleistungen als solche gekennzeichnet.

Stefan Krämer

Leipzig, May 27, 2022