

Adverse outcome pathways coming to life

Citation for published version (APA):

Martens, M. T. L. J. (2024). *Adverse outcome pathways coming to life: exploring new ways to support risk assessments*. [Doctoral Thesis, Maastricht University]. Maastricht University.
<https://doi.org/10.26481/dis.20240129mm>

Document status and date:

Published: 01/01/2024

DOI:

[10.26481/dis.20240129mm](https://doi.org/10.26481/dis.20240129mm)

Document Version:

Publisher's PDF, also known as Version of record

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.umlib.nl/taverne-license

Take down policy

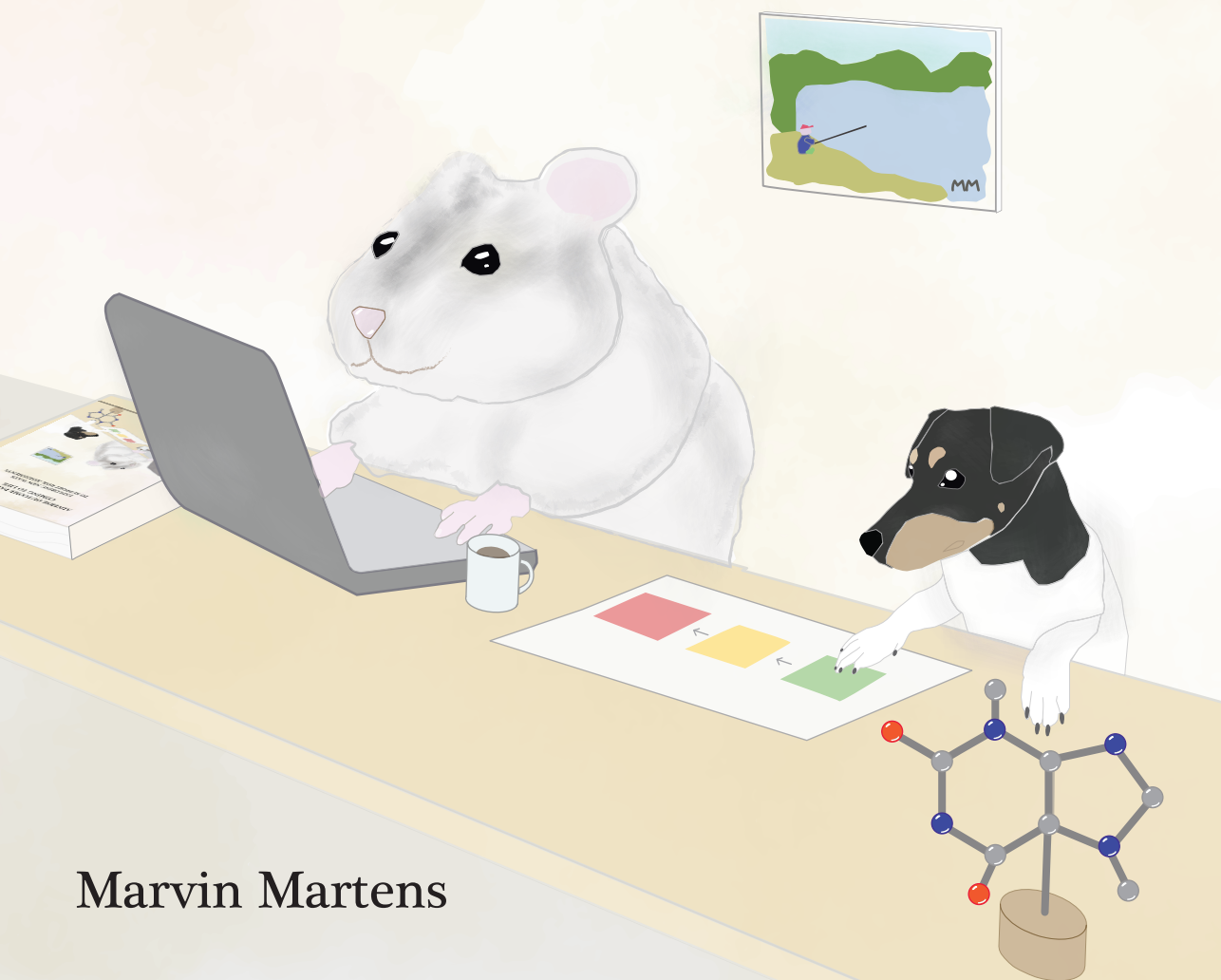
If you believe that this document breaches copyright please contact us at:

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.

ADVERSE OUTCOME PATHWAYS COMING TO LIFE

EXPLORING NEW WAYS TO SUPPORT RISK ASSESSMENTS



Marvin Martens

**ADVERSE OUTCOME PATHWAYS
COMING TO LIFE**

**EXPLORING NEW WAYS TO SUPPORT
RISK ASSESSMENTS**

© Marvin Martens, Maastricht, 2024

Cover design: Estefania Carolina Prado Clavijo

Layout by: Marvin Martens

Printed by: Total Copy Service

ISBN/EAN: 978-94-6469-725-4

ADVERSE OUTCOME PATHWAYS COMING TO LIFE

EXPLORING NEW WAYS TO SUPPORT RISK ASSESSMENTS

DISSERTATION

To obtain the degree of Doctor at Maastricht University,
on the authority of the Rector Magnificus, Prof. dr. P. Habibović,
in accordance with the decision of the Board of Deans,
to be defended in public on
Monday, January 29th, 2024, at 16:00hours

by

Marvin Theodoor Leonardus Johannes Martens

Supervisor

Prof. dr. Chris T. Evelo

Co-supervisor

Dr. Egon L. Willighagen

Assessment Committee

Prof. dr. Theo de Kok (Chair)

Dr. Remzi Çelebi

Dr. Roger Godschalk

Prof. dr. Paul Groth (University of Amsterdam)

Prof. dr. Juliette Legler (Utrecht University)

To my parents

Contents

1	Introduction	1
1.1	Novel technologies	3
1.2	Adverse Outcome Pathways	4
1.3	Thesis outline	9
2	WikiPathways: connecting communities	19
2.1	Introduction	22
2.2	Content and general updates	23
2.3	Pathway curation communities	26
2.4	Connections to other initiatives	30
2.5	Future work	36
2.6	Conclusion	37
2.7	Data availability	37
3	Introducing WikiPathways as a Data-Source to Support Adverse Outcome Pathways for Regulatory Risk Assessment of Chemicals and Nanomaterials	43
3.1	Introduction	46
3.2	Materials and Methods	50
3.3	Results	53
3.4	Discussion	56
4	Providing adverse outcome pathways from the AOP-Wiki in a semantic web format to increase usability and accessibility of the content.	65
4.1	Introduction	67
4.2	Methods	69
4.3	Results	75
4.4	Discussion	83
4.5	Data links	88

5	The AOP-DB RDF: Applying FAIR Principles to the Semantic Integration of AOP Data Using the Research Description Framework	95
5.1	Introduction	97
5.2	Materials and Methods	99
5.3	Results	100
5.4	Conclusion	103
6	AOPLink Jupyter Notebook: Extracting and analysing data related to an AOP of interest	107
6.1	Introduction	108
6.2	Jupyter notebook	110
7	Molecular Adverse Outcome Pathways: towards the implementation of transcriptomics data in risk assessments	143
7.1	Introduction	145
7.2	Methods	148
7.3	Results	152
7.4	Discussion	161
8	Discussion	173
8.1	WikiPathways	175
8.2	Toxicological data	183
8.3	Adverse Outcome Pathways as a tool in risk assessments	189
8.4	Conclusion	191
	Impact	201
	Summary	209
	Samenvatting	213
	List of Abbreviations	219
	Acknowledgments	221

About the author	227
Published work	229

List of Figures

2.1	WikiPathways Graphical Abstract.	21
2.2	Recent growth of WikiPathways.	24
2.3	The number of revisions and contributors for all pathways in the human pathway analysis collection.	25
2.4	The COVID-19 Portal on WikiPathways (covid.wikipathways.org).	28
2.5	The WikiPathways SNORQL user interface (sparql.wikipathways.org).	32
3.1	Illustrative description of the linkage of KEs of an AOP with molecular pathways described in WikiPathways and the practical application of transcriptomics.	50
3.2	Ontology usage of AOP-Wiki for KEs on the molecular, cellular, tissue, and organ level of biological organization.	54
3.3	WikiPathways statistics.	55
3.4	AOP-Wiki statistics on KEs and KERs, identifier mapping with HGNC identifiers and links to molecular pathways in WikiPathways.	56
4.1	General overview of the AOP-Wiki RDF scheme.	70
4.2	The AOP-Wiki SNORQL User Interface.	74
4.3	Adverse Outcome Pathways and their properties in RDF.	78
4.4	Key Events and their properties in RDF.	79
4.5	Stressors and chemicals and their properties in RDF.	81
4.6	Ontology annotations and molecular identifiers.	82
5.1	The OECD funded AOP-KB currently support the AOP-Wiki.	98
5.2	AOP-DB Semantic Mapping.	102
7.1	The AOP Portal on WikiPathways.org.	149

7.2	Conceptual illustration of a molecular AOP in WikiPathways.	149
7.3	Gene expression data after Rifampicin exposure visualised on PXR AOP.	154
7.4	Gene expression data after SR12813 exposure visualised on PXR AOP.	155
7.5	Gene expression data after T0901317 exposure visualised on PXR AOP.	156
7.6	Gene expression data visualised on wikipathways:WP2876 (Pregnane X receptor pathway) for the three PXR agonists.	157
7.7	Gene expression data after 4 hour exposure to GW3965 (LXR agonist) visualised on LXR AOP.	158
7.8	Gene expression data after 24 hour exposure to GW3965 (LXR agonist) visualised on LXR AOP.	158
7.9	Gene expression data after exposure to GW3965 (LXR agonist) visualised on the SREBP signaling pathway (wikipathways:WP1982).	159
7.10	Rotenone (50nM) data visualised on mitochondrial complex I inhibition AOP.	159
7.11	Rotenone (100nM) data visualised on mitochondrial complex I inhibition AOP.	160

List of Tables

4.1	Ontologies and vocabularies used in the RDF.	72
4.2	Prefixes in the RDF for the Key Event Component annotations.	94
5.1	Overview of ontologies, consistent vocabularies and databases included in the AOP-DB RDF.	101
7.1	Overview of datasets used and detailed information on relevant samples.	150
7.2	Enrichment Scores for KEs by PXR agonists.	153
7.3	Enrichment Scores of KEs after exposure to GW3965 (LXR agonist).	155
7.4	Enrichment Scores for KEs by exposure to Rotenone at 50nM and 100nM.	161
8.1	Recent projects in which Adverse Outcome Pathways (AOPs) are a central theme.	191

1

Introduction

The field of toxicology originates from the science of poisons, with the very first observed toxic effects likely observed by accident in ancient times. With humanity learning about harmful substances came practical applications, such as poisoning enemies through exposure to toxicants [1]. Later, around the sixteenth century, it was determined that any substance that can harm or kill at a high enough dose is a poison, with a relationship between the dose and the effect: the dose-response relationship [2]. With the publication of “*Traite des Poisons*” in 1814 by Orfila [3], toxicology as a field of science was born, defined by scientific investigation and evaluation of toxic exposures. With the improved scientific understanding and development of analytical methodologies, the field of toxicology has progressed to an era of understanding toxicants and their effects on the molecular level [4].

Nowadays, the toxicology domain can be subdivided into clinical, forensic, and regulatory toxicology, the latter of which includes occupational and environmental toxicology, and workspace drug testing [4]. Regulatory toxicology involves the assessment of potential hazards and risks of exposure to toxicants and the regulation

thereof, based on exposure and potency. By systematic evaluation and integration of qualitative and quantitative information, risk assessment aims to identify potential adverse health effects resulting from exposure to hazardous stressors [5].

The origin of toxicological data for regulatory purposes can originate from various sources, including human *in vivo* databases, animal experimentation, *in vitro* cell cultures, and *in silico* methods. Traditionally, regulatory hazard and safety assessments on chemicals heavily rely on animal experimentation [6]. However, these come with a high cost in resources and time, and therefore the throughput of testing was limited, which cannot keep up with the increasing number of new chemicals and nanomaterials that require assessment. Furthermore, animal models introduce uncertainty in assessing hazards for humans [7], and the use of animal models encounter ethical and societal concerns [8, 9].

In 1959, there was the first description of the widely implemented 3Rs: Reduction, Refinement and Replacement [10, 11]. These are aimed to improve experimental design to minimize the number of animals used, minimize animal suffering and improve welfare. Ultimately, a transition towards non-animal approaches would be made, replacing animal experimentation with *in vitro* and *in silico* methods [8, 12].

With the development of so-called New Approach Methodologies (NAMs), this issue would be addressed. NAMs can drive the paradigm shift from animal experiments to robust, targeted, mechanism-based *in vitro* and *in silico* methods. These, particularly as a complementary testing battery, can be used for hazard assessment, prioritization, read-across, and screening of chemicals [13–15]. However, while many efforts focus on the development of NAMs, procedures for the validation and approval for risk assessment purposes should still be defined for optimal uptake by regulators [16–18]. For example, a major challenge lies in the design of a set of *in vitro* assays to produce results that can be used as good predictors of *in vivo* toxicities [19, 20].

1.1 Novel technologies

In recent decades, various new technologies have emerged to allow large-scale examination of biological responses upon exposure to a stressor and elucidate the underlying mechanisms of action. These so-called “omics-technologies” include the characterization of a wide range of biomarkers on the level of DNA (genomics), mRNA (transcriptomics), protein (proteomics) and small molecules (metabolomics), among others. Toxicogenomics, a recent branch in the domain of toxicology, focuses on the use of these omics technologies to study the molecular and cellular processes caused by toxicants [21–25].

Although toxicogenomic, specifically transcriptomic approaches such as microarrays are powerful tools for risk and hazard identification [26–29], they are not yet widely applied in risk assessments. Various issues have been expressed such as the lower specificity and sensitivity on individual genes when compared to RT-PCR, the statistics involving the false discovery rate, and the large scale of single datasets [30]. Furthermore, there is no consensus on the storage, curation, processing, and normalisation of the data, and there is uncertainty introduced in the interpretation of the data. Another concern is the reproducibility of the experiments and processing. Therefore, omics approaches are often regarded as hypothesis generation, and there is a lack of confidence in their implementation in risk assessments [31, 32]. There is also a validation barrier that needs to be addressed, where omics approaches should be validated on their consistency and reliability, the software used to collect and analyse the data, the application and relevance for the biological endpoint to assess, and the ability to generalize or specify for smaller target populations [32].

To improve the standardization, reliability, and transparency of transcriptomics, recent efforts have focused on providing guidelines and best practices for omics data generation, handling and the conversion of raw datasets into biological observations [33–36]. This has also led

to the recent development of the Transcriptomics Reporting Framework by the Organisation for Economic Co-operation and Development (OECD) Extended Advisory Group on Molecular Screening and Toxicogenomics (EAGMST), focused on the reporting of omics studies in toxicology to increase transparency and standardization in reporting of data and associated metadata [37]. Furthermore, case-study-driven studies focus on the potential applications of transcriptomics data and address aspects that currently limit their uptake in regulatory risk assessments. For instance, the generation of a predictive toxicogenomics space can explain dose-dependent cytotoxicity and provide a probability score for the induction of drug-induced liver injury [38].

After processing transcriptomics data, paired samples with different gene expression patterns are compared to generate lists of differentially expressed genes (DEGs). These can provide insights into the potency of stressors on the test system under various treatment conditions involving concentration, time, and exposure patterns [33]. By grouping genes based on their functional annotations, pathway-specific perturbation estimates can be generated. This can be done based on gene sets annotated in Gene Ontology [39, 40] for the function of genes, or molecular pathway databases that describe the involvement of genes in the complexity of biological pathways in detail [35]. One of the molecular pathway databases with such functional annotations of genes and pathways is the community-driven database WikiPathways [41]. Together with its accompanying pathway drawing and analysis tool PathVisio [42], WikiPathways allows for the tailored development and curation of molecular pathways and the analysis of a variety of omics datasets.

1.2 Adverse Outcome Pathways

Since their first description in 2010 [43], AOPs have become a central concept in the field of risk assessments, particularly to drive the paradigm from using animal models for safety assessments towards *in vitro* systems [44–46]. This is in line with the previously described

3Rs, which have been a goal in toxicological risk assessments for a long time. The purpose of AOPs is to capture and reuse existing mechanistic information on toxicological processes and identify gaps of knowledge to be investigated.

The concept of AOPs describes a simplified biological description of toxicological processes in response to stressors such as chemicals, nanomaterials, or types of radiation. The complete toxicological pathway is broken down into smaller processes called Key Events (KEs), which are defined as measurable endpoints and essential for progression towards the Adverse Outcome (AO), the apical endpoint that is relevant for risk assessment [47]. The most upstream, i.e. the first KE in the AOP, is the Molecular Initiating Event (MIE), which contains evidence for the activation by their specific (group of) stressors. All KEs, including MIE and AO, are connected by Key Event Relationships (KERs), which describe the biological basis of causation based on biological understanding and empirical support. These connect the molecular processes in response to stressor exposure, through increasing levels of biology (cell, tissue, organ), with the adverse effects on the level of the individual or population. By design, AOPs are stressor agnostic, meaning that any trigger of a particular MIE can potentially activate a cascade of downstream KEs [48], and KEs can be shared among AOPs, creating a larger AOP Networks [49, 50]. Qualitative descriptions of AOPs are generally captured and stored within the AOP-Wiki, a core part of the AOP Knowledge Base (AOP-KB) and an initiative by the OECD to serve as a platform for collaborative development and communication with regulators [48]. Within this thesis, the AOP-Wiki is a central theme of integration, improving its interoperability.

To be implemented in risk assessments, AOPs can be used to inform hypothesis-driven Integrated Approaches to Testing and Assessment (IATA), pragmatic, science-based solutions to efficient and cost-effective chemicals hazard characterization [51, 52]. For example, skin sensitization was the first apical endpoint for which IATA, consisting of *in silico*, *in chemico* and *in vitro* tests, were

generated on the basis of an AOP [52–57]. However, the acceptance of AOPs as an information resource faces challenges in the overall confidence, completeness, and usefulness [58]. To increase confidence, the weight of evidence considerations have been described based on the Bradford-Hill criteria [59, 60].

1.2.1 Quantification of Adverse Outcome Pathways

While AOPs are generally based on scientific literature, many efforts have focused on the quantification of AOPs, single KEs, and KERs with experimental data or computational predictions. These are generally referred to as various types of quantitative AOPs (qAOPs). Besides the quantitative weighing of evidence in weight-of-evidence qAOPs, for example, using the extended Bradford-Hil criteria [60], the more data-driven qAOPs are empirical dose-response based qAOPs, probabilistic qAOPs, and mechanistic qAOPs [61, 62]. In dose-response-based qAOPs, equations are fitted to measures of each KE at exposures of increasing dose, which are mathematically adjusted to obtain chemical-independent KER quantification [62]. Probabilistic qAOPs provide predictive relationships between KEs and can cover complete AOPs. For example, AOP Bayesian networks have been implemented to predict the probability of chemicals to cause liver steatosis [63], renal toxicity [62], or ATP production associated growth inhibition [64], among others. Finally, mechanistic (or systems toxicology) qAOPs provide the most insights into biological processes. Unlike the other qAOP types, mechanistic qAOPs describe the biological complexity with feedback and feed-forward loops, regulatory processes and modulating factors [61, 62]. Generally, the conversion of a complete AOP into a qAOP is challenging, especially when working with public data, with limiting factors such as data availability and usefulness, the separation of data across different studies, and the accessibility and transferability of established quantitative models [65]. AOPs are also regarded as a pushing force in the design and development of computational models that can predict the activation of particular measurable endpoints related to KEs [66]. For example, the AOP of aromatase in-

hibition leading to reproductive impairment is supported by a set of predictive computational models on molecular, cellular, and population level KEs [67].

1.2.2 Adverse Outcome Pathways and transcriptomics

Another type of quantification in AOP involves large-scale omics approaches, with various applications in the development and support of AOPs. For example, it is described that omics approaches can aid AOP development by defining MIEs and KEs, supporting analysis or validation of KEs by providing supportive evidence and providing biomarkers for hazard identification [68, 69]. Omics can serve in the understanding of biological networks and assess or verify the mode of action of chemicals [70, 71], which can be useful for grouping chemicals and performing read-across [69]. To structure the process of AOP integration with transcriptomics data, a data fusion pipeline has been developed to generate AOP-based molecular pathway networks based on a variety of toxicology databases [72]. Its purpose is to be used for the analysis of transcriptomics data in the regulatory context. Using WikiPathways as a tool to describe the biological complexity of the AOP, it was possible to identify the dysregulation of genes and processes associated with the KEs of the AOP [72]. Besides the support of existing AOPs, transcriptomics data can also drive the generation of computationally predicted AOP (cpAOP) scaffolds to support the development of AOPs [73].

1.2.3 Integration of data and resources

Since AOPs describe an aggregation of toxicological information on all levels of biological organization which is stored in public resources [74], their development can be supported by data-driven approaches. For example, data from the Comparative Toxicogenomics Database (CTD) [75] and ToxCast can be utilized to generate cpAOPs based on the chemicals of interest [76]. It is expected that the integration of public resources through computational approaches

can speed up the mostly expert-reliable development of AOPs. By combining automated tools for AOP development with the AOP-KB as a knowledge management tool, the continuously expanding source of information can increase the quantity and quality of AOPs [74]. For example, the integration of various techniques for MIE prediction and KER support was tested in thyroid hormone-related toxicity, providing an efficient way to extend AOP knowledge and assess health hazards [77].

With the transition toward more data-driven approaches and the goal to make data more useful by integration and reuse, Linked Open Data solutions have been defined for handling data. For example, the Resource Description Framework (RDF) model describes data in semantic triples, defined by combinations of subjects, predicates and objects which jointly make up statements in a machine-readable way [78]. RDF is supported by annotations with ontologies and persistent identifiers. To standardize ontologies, the Web Ontology Language (OWL) was defined for representing knowledge, grouping, and relations between concepts, adding more extensive semantics to RDF triple statements [79]. Furthermore, identifiers can be standardized by the implementation of Internationalized Resource Identifiers (IRIs), providing unique identification of resources and their contents [80]. The query language to explore the vastness of RDF data and its extensive interoperability capabilities is SPARQL Protocol and RDF Query Language (SPARQL), allowing flexible, reusable queries across resources.

Such technologies to annotate data and improve the semantic meaning of data and knowledge are in line with the so-called Findable, Accessible, Interoperable, and Reusable (FAIR) principles, standing for Findable, Accessible, Interoperable and Reusable [81, 82]. These principles were designed to increase the overall usability of data by applying standards and extensive metadata to describe each dataset. By design, RDF and Ontologies are such standards to create linked data, aiming for consistent reporting of data and metadata across resources and disciplines, focusing on interoperability.

As such, semantic web applications are also applied in toxicology research, to annotate experimental data and metadata, and employ linked data practises in databases [83]. Large initiatives such as OpenTox [84], eTOX [85], eNanoMapper [86] have focused on optimizing data interoperability, both by the creation of centralized data repositories supported by linked data, but also providing resources for data annotation and integration of toxicological data, such as the nanotoxicology-focused eNanoMapper Ontology [87] and FAIRification workflow [88], or the OpenTox Application Programming Interface (API) [89, 90].

The efforts to improve data interoperability also focus on AOPs and related resources, including the AOP-Wiki. With the aim to capture as much as possible of toxicological space as possible, the Key Event Components (KECs) were introduced to the AOP-Wiki [91]. These are focused on the implementation of ontologies to annotate KEs, specifically on biological processes and biological objects, covering all biological levels. The aim of introducing ontologies to the AOP-Wiki was to make linked data, to explore connections with other resources based on standardized annotations of concepts. This aspect will be investigated in more detail in Chapter 3. Another effort to centralize AOP concept annotations is the AOP Ontology (AOPO) [92], which has been implemented in the AOPXplorer, a Cytoscape plugin, for the annotation of AOP-related concepts. This can be used for creating AOP networks, as has been investigated for hepatotoxicity [93] and neurotoxicity [94].

1.3 Thesis outline

With the advancements in scientific methodologies over the past decades, the risk assessments of chemicals and other types of stressors are moving toward more extensive, high-throughput screenings. Not only do we learn about the hazards and risks of these, but the focus is on understanding the mechanism of the interactions of stressors with the biological system, the resulting cellular responses,

and the consequential adversities with sufficient exposure. With this thesis, we expect that the application of high-content data, such as transcriptomics, together with AOPs, may be an effective way to both assess adverse effects caused by stressors, and inform us about the mechanistic effects as well in detail. By addressing the integration of AOPs with experimental data and molecular pathways, we aimed to facilitate data analysis and the eventual transition toward high-throughput, high-content data applications in risk assessments.

Within Chapter 2, we show the community-driven developments within the molecular pathway database called WikiPathways, a resource that allows for interoperability with other resources and can serve as a platform for omics data analysis with PathVisio. Because our aim is to integrate resources for omics analyses with AOPs, we explored the AOP-Wiki and its current coverage of interoperability aspects in Chapter 3. By looking into molecular descriptors and ontologies, we explored ways of linking the AOP-Wiki with WikiPathways.

Following that, we implemented semantic web technologies in the AOP-Wiki datasets and developed an RDF schema for it in Chapter 4. We enriched the data with ontologies, standard vocabularies, and persistent identifiers, and made the data explorable with SPARQL queries. As an extension, the AOP-DB was also converted to an RDF format in Chapter 5, providing additional resources to connect to the KEs of the AOP-Wiki. As an illustration of the usefulness of the conversion to RDF, we developed an AOPLink computational workflow, that utilizes, among other services, the AOP-Wiki RDF and AOP-DB RDF to automatically explore experimental data to support any AOPs that exist in the AOP-Wiki, presented as a Jupyter notebook in Chapter 6.

Following the semantification of the AOP-related resources, we introduced a transcriptomics data analysis framework that combines AOPs and molecular pathways in WikiPathways in Chapter 7, which we called molecular AOPs. We explored and discussed their usefulness

and provided use cases as examples, focusing on various AOPs and transcriptomics data sets. Finally, this work is summarized and discussed, looking at its implications and overall impact.

References

- [1] Loralie J. Langman and Bhushan M. Kapur. "Toxicology: Then and now". *Clinical Biochemistry* 39.5 (May 2006), pp. 498–510. DOI: 10.1016/j.clinbiochem.2006.03.004.
- [2] A.M. Tsatsakis et al. "The dose response principle from philosophy to modern toxicology: The impact of ancient philosophy and medicine in modern toxicology science". *Toxicology Reports* 5 (Jan. 2018), pp. 1107–1113. DOI: 10.1016/j.toxrep.2018.10.001.
- [3] Mathieu Orfila. *Traité des poisons tirés de règnes minéral, végétal et animal, ou toxicologie générale, considérée sous les rapports de la physiologie, de la pathologie et de la médecine légale*, 2 Bände. Vol. I. Paris : Crochard et Gabon, 1818.
- [4] Alex A Pappas et al. "Toxicology: Past, Present, and Future". 29.4 (1999).
- [5] J. M. Barnes and F. A. Denz. "Experimental methods used in determining chronic toxicity; a critical review." *Pharmacological reviews* 6.2 (June 1954), pp. 191–242.
- [6] WS Stokes. "Animals and the 3Rs in toxicology research and testing". *Human & Experimental Toxicology* 34.12 (Dec. 2015), pp. 1297–1303. DOI: 10.1177/0960327115598410.
- [7] Harry Olson et al. "Concordance of the Toxicity of Pharmaceuticals in Humans and in Animals". *Regulatory Toxicology and Pharmacology* 32.1 (Aug. 2000), pp. 56–67. DOI: 10.1006/rtph.2000.1399.
- [8] Natalie Burden et al. "Aligning the 3Rs with new paradigms in the safety assessment of chemicals". *Toxicology* 330 (Apr. 2015), pp. 62–66. DOI: 10.1016/j.tox.2015.01.014.
- [9] Joanne Zurlo. "No Animals Harmed: Toward a Paradigm Shift in Toxicity Testing". *Hastings Center Report* 42.s1 (Nov. 2012), S23–S26. DOI: 10.1002/HAST.104.
- [10] Robert G. W. Kirk. "Recovering The Principles of Humane Experimental Technique". *Science, Technology, & Human Values* 43.4 (July 2018), pp. 622–648. DOI: 10.1177/0162243917726579.
- [11] William Moy Stratton Russell and Rex Leonard Burch. *The principles of humane experimental technique*. Methuen, 1959.
- [12] Paul Flecknell. "Replacement, reduction and refinement." *ALTEX* 19.2 (2002), pp. 73–8.
- [13] Sylvia E. Escher et al. "Towards grouping concepts based on new approach methodologies in chemical hazard assessment: the read-across approach of the EU-ToxRisk project". *Archives of Toxicology* 93.12 (Dec. 2019), pp. 3643–3667. DOI: 10.1007/s00204-019-02591-7.
- [14] R. Graepel et al. "Paradigm shift in safety assessment using new approach methods: The EU-ToxRisk strategy". *Current Opinion in Toxicology* 15 (June 2019), pp. 33–39. DOI: 10.1016/j.cotox.2019.03.005.

- [15] Ans Punt et al. "New approach methodologies (NAMs) for human-relevant biokinetics predictions. Meeting the paradigm shift in toxicology towards an animal-free chemical risk assessment". *ALTEX* 37.4 (Oct. 2020), pp. 607–622. DOI: 10.14573/altex.2003242.
- [16] Stanley T. Parish et al. "An evaluation framework for new approach methodologies (NAMs) for human health safety assessment". *Regulatory Toxicology and Pharmacology* 112 (Apr. 2020), p. 104592. DOI: 10.1016/j.yrtph.2020.104592.
- [17] Francesca Pistollato et al. "Current EU regulatory requirements for the assessment of chemicals and cosmetic products: challenges and opportunities for introducing new approach methodologies". *Archives of Toxicology* 95.6 (June 2021), pp. 1867–1897. DOI: 10.1007/s00204-021-03034-y.
- [18] Matthieu Mondou et al. "Envisioning an international validation process for New Approach Methodologies in chemical hazard and risk assessment". *Environmental Advances* 4 (July 2021), p. 100061. DOI: 10.1016/j.envadv.2021.100061.
- [19] Paul Jennings. "The future of in vitro toxicology". *Toxicology in Vitro* 29.6 (Sept. 2015), pp. 1217–1221. DOI: 10.1016/j.tiv.2014.08.011.
- [20] M.J. Garle, J.H. Fentem, and J.R. Fry. "In vitro cytotoxicity tests for the prediction of acute toxicity in vivo". *Toxicology in Vitro* 8.6 (Dec. 1994), pp. 1303–1312. DOI: 10.1016/0887-2333(94)90123-6.
- [21] William D. Pennie et al. "The principles and practice of toxigenomics: applications and opportunities." *Toxicological Sciences* 54.2 (Apr. 2000), pp. 277–83. DOI: 10.1093/toxsci/54.2.277.
- [22] Sara Shostak. "The Emergence of Toxicogenomics". *Social Studies of Science* 35.3 (June 2005), pp. 367–403. DOI: 10.1177/0306312705049882.
- [23] Marilyn J. Aardema and James T. MacGregor. "Toxicology and Genetic Toxicology in the New Era of "Toxicogenomics": Impact of "-omics" Technologies". *Toxicogenomics*. Tokyo: Springer Japan, 2003, pp. 171–193. DOI: 10.1007/978-4-431-66999-9_22.
- [24] Hisham K. Hamadeh et al. "Discovery in toxicology: Mediation by gene expression array technology". *Journal of Biochemical and Molecular Toxicology* 15.5 (2001), pp. 231–242. DOI: 10.1002/jbt.10006.
- [25] Benjamin Alexander-Dann et al. "Developments in toxicogenomics: understanding and predicting compound-induced toxicity from gene expression data". *Molecular Omics* 14.4 (Aug. 2018), pp. 218–236. DOI: 10.1039/C8MO00042E.
- [26] Benjamin Piña et al. "Functional Data Analysis: Omics for Environmental Risk Assessment". *Comprehensive Analytical Chemistry*. Vol. 82. Elsevier, Jan. 2018, pp. 583–611. DOI: 10.1016/bs.coac.2018.07.007.
- [27] Nabila Haddad et al. "Next generation microbiological risk assessment—Potential of omics data for hazard characterisation". *International Journal of Food Microbiology* 287 (Dec. 2018), pp. 28–39. DOI: 10.1016/j.ijfoodmicro.2018.04.015.
- [28] Kenneth M.Y. Leung. "Joining the dots between omics and environmental management". *Integrated Environmental Assessment and Management* 14.2 (Mar. 2018), pp. 169–173. DOI: 10.1002/ieam.2007.

-
- [29] Carole L. Yauk et al. "Toxicogenomic applications in risk assessment at Health Canada". *Current Opinion in Toxicology* 18 (Dec. 2019), pp. 34–45. DOI: 10.1016/j.cotox.2019.02.005.
- [30] Saura C. Sahu. *Toxicogenomics: A Powerful Tool for Toxicity Assessment*. Ed. by Saura C. Sahu. Chichester, UK: John Wiley & Sons, Ltd, Oct. 2008, pp. 1–409. DOI: 10.1002/9780470699638.
- [31] Roland Buesen et al. "Applying 'omics technologies in chemicals risk assessment: Report of an ECETOC workshop". *Regulatory Toxicology and Pharmacology*. Vol. 91. Academic Press, Dec. 2017, S3–S13. DOI: 10.1016/j.yrtph.2017.09.002.
- [32] Ursula G. Sauer et al. "The challenge of the application of 'omics technologies in chemicals risk assessment: Background and outlook". *Regulatory Toxicology and Pharmacology* 91 (Dec. 2017), S14–S26. DOI: 10.1016/j.yrtph.2017.09.020.
- [33] Joshua Harrill et al. "Considerations for strategic use of high-throughput transcriptomics chemical screening data in regulatory decisions". *Current Opinion in Toxicology* 15 (June 2019), pp. 64–75. DOI: 10.1016/j.cotox.2019.05.004.
- [34] Pia Anneli Sofia Kinaret et al. "Transcriptomics in Toxicogenomics, Part I: Experimental Design, Technologies, Publicly Available Data, and Regulatory Aspects". *Nanomaterials* 10.4 (Apr. 2020), p. 750. DOI: 10.3390/nano10040750.
- [35] Antonio Federico et al. "Transcriptomics in Toxicogenomics, Part II: Preprocessing and Differential Expression Analysis for High Quality Data". *Nanomaterials* 10.5 (May 2020), p. 903. DOI: 10.3390/nano10050903.
- [36] Angela Serra et al. "Transcriptomics in Toxicogenomics, Part III: Data Modelling for Risk Assessment". *Nanomaterials* 10.4 (Apr. 2020), p. 708. DOI: 10.3390/nano10040708.
- [37] Joshua A. Harrill et al. "Progress towards an OECD reporting framework for transcriptomics and metabolomics in regulatory toxicology". *Regulatory Toxicology and Pharmacology* 125 (Oct. 2021), p. 105020. DOI: 10.1016/j.yrtph.2021.105020.
- [38] Pekka Kohonen et al. "A transcriptomics data-driven gene space accurately predicts liver cytopathology and drug-induced liver injury". *Nature Communications* 8.1 (July 2017), pp. 1–15. DOI: 10.1038/ncomms15932.
- [39] Michael Ashburner et al. "Gene Ontology: tool for the unification of biology". *Nature Genetics* 25.1 (May 2000), pp. 25–29. DOI: 10.1038/75556.
- [40] Seth Carbon et al. "The Gene Ontology resource: enriching a GOld mine". *Nucleic acids research* 49.D1 (Jan. 2021), pp. D325–D334.
- [41] Marvin Martens et al. "WikiPathways: connecting communities". *Nucleic Acids Research* 49.D1 (Jan. 2021), pp. D613–D621. DOI: 10.1093/nar/gkaa1024.
- [42] Martina Kutmon et al. "PathVisio 3: An Extendable Pathway Analysis Toolbox". *PLOS Computational Biology* 11.2 (Feb. 2015). Ed. by Robert F. Murphy, e1004085. DOI: 10.1371/journal.pcbi.1004085.
- [43] Gerald T. Ankley et al. "Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment". *Environmental Toxicology and Chemistry* 29.3 (Mar. 2010), pp. 730–741. DOI: 10.1002/etc.34.
- [44] Mathieu Vinken. "The adverse outcome pathway concept: A pragmatic tool in toxicology". *Toxicology* 312.1 (Oct. 2013), pp. 158–165. DOI: 10.1016/j.tox.2013.08.011.

- [45] Marcel Leist et al. "Adverse outcome pathways: opportunities, limitations and open questions". *Archives of Toxicology* 91.11 (Nov. 2017), pp. 3477–3505. DOI: 10.1007/s00204-017-2045-3.
- [46] S. W. Edwards et al. "Adverse Outcome Pathways—Organizing Toxicological Information to Improve Decision Making". *Journal of Pharmacology and Experimental Therapeutics* 356.1 (Dec. 2016), pp. 170–181. DOI: 10.1124/jpet.115.228239.
- [47] Gerald T. Ankley and Stephen W. Edwards. "The adverse outcome pathway: A multifaceted framework supporting 21st century toxicology". *Current Opinion in Toxicology* 9 (June 2018), pp. 1–7. DOI: 10.1016/j.cotox.2018.03.004.
- [48] Daniel L. Villeneuve et al. "Adverse outcome pathway (AOP) development I: strategies and principles." *Toxicological Sciences* 142.2 (Dec. 2014), pp. 312–320. DOI: 10.1093/toxsci/kfu199.
- [49] Dries Knapen et al. "Adverse outcome pathway networks I: Development and applications". *Environmental Toxicology and Chemistry* 37.6 (June 2018), pp. 1723–1733. DOI: 10.1002/etc.4125.
- [50] Thomas Ball et al. "Beyond adverse outcome pathways: Making toxicity predictions from event networks, SAR models, data and knowledge". *Toxicology Research* 10.1 (Feb. 2021), pp. 102–122. DOI: 10.1093/toxres/tfaa099.
- [51] Knut Erik Tollefsen et al. "Applying Adverse Outcome Pathways (AOPs) to support Integrated Approaches to Testing and Assessment (IATA)". *Regulatory Toxicology and Pharmacology* 70.3 (Dec. 2014), pp. 629–640. DOI: 10.1016/j.yrtph.2014.09.009.
- [52] Yuki Sakuratani, Masashi Horie, and Eeva Leinala. "Integrated Approaches to Testing and Assessment: OECD Activities on the Development and Use of Adverse Outcome Pathways and Case Studies". *Basic & Clinical Pharmacology & Toxicology* 123 (Sept. 2018), pp. 20–28. DOI: 10.1111/BCPT.12955.
- [53] Susanne N. Kolle, Robert Landsiedel, and Andreas Natsch. "Replacing the refinement for skin sensitization testing: Considerations to the implementation of adverse outcome pathway (AOP)-based defined approaches (DA) in OECD guidelines". *Regulatory Toxicology and Pharmacology* 115 (Aug. 2020), p. 104713. DOI: 10.1016/j.yrtph.2020.104713.
- [54] Cameron MacKay et al. "From pathways to people: applying the adverse outcome pathway (AOP) for skin sensitization to risk assessment". *ALTEX* 30.4 (Nov. 2013), pp. 473–486. DOI: 10.14573/altex.2013.4.473.
- [55] Gavin Maxwell et al. "Applying the skin sensitisation adverse outcome pathway (AOP) to quantitative risk assessment". *Toxicology in Vitro* 28.1 (Feb. 2014), pp. 8–12. DOI: 10.1016/j.tiv.2013.10.013.
- [56] Jochem W. van der Veen et al. "Anchoring molecular mechanisms to the adverse outcome pathway for skin sensitization: Analysis of existing data". *Critical Reviews in Toxicology* 44.7 (Aug. 2014), pp. 590–599. DOI: 10.3109/10408444.2014.925425.
- [57] Grace Patlewicz et al. "Towards AOP application – Implementation of an integrated approach to testing and assessment (IATA) into a pipeline tool for skin sensitization". *Regulatory Toxicology and Pharmacology* 69.3 (Aug. 2014), pp. 529–545. DOI: 10.1016/j.yrtph.2014.06.001.
- [58] Edward J. Perkins et al. "Adverse Outcome Pathways for Regulatory Applications: Examination of Four Case Studies With Different Degrees of Complete-

-
- ness and Scientific Confidence". *Toxicological Sciences* 148.1 (Nov. 2015), pp. 14–25. DOI: 10.1093/toxsci/kfv181.
- [59] Richard A. Becker et al. "Increasing Scientific Confidence in Adverse Outcome Pathways: Application of Tailored Bradford-Hill Considerations for Evaluating Weight of Evidence". *Regulatory Toxicology and Pharmacology* 72.3 (Aug. 2015), pp. 514–537. DOI: 10.1016/j.yrtph.2015.04.004.
- [60] Zachary A. Collier et al. "A weight of evidence assessment approach for adverse outcome pathways". *Regulatory Toxicology and Pharmacology* 75 (Mar. 2016), pp. 46–57. DOI: 10.1016/j.yrtph.2015.12.014.
- [61] Nicoleta Spinu et al. "Quantitative adverse outcome pathway (qAOP) models for toxicity prediction". *Archives of Toxicology* 94.5 (May 2020), pp. 1497–1510. DOI: 10.1007/s00204-020-02774-7.
- [62] Elias Zgheib et al. "Application of three approaches for quantitative AOP development to renal toxicity". *Computational Toxicology* 11 (Aug. 2019), pp. 1–13. DOI: 10.1016/j.comtox.2019.02.001.
- [63] Lyle D. Burgoon et al. "Predicting the Probability that a Chemical Causes Steatosis Using Adverse Outcome Pathway Bayesian Networks (AOPBNs)". *Risk Analysis* 40.3 (Mar. 2020), pp. 512–523. DOI: 10.1111/RISA.13423.
- [64] S. Jannicke Moe et al. "Quantification of an Adverse Outcome Pathway Network by Bayesian Regression and Bayesian Network Modeling". *Integrated Environmental Assessment and Management* 17.1 (Jan. 2021), pp. 147–164. DOI: 10.1002/ieam.4348.
- [65] Dennis Sinitzyn, Natàlia Garcia-Reyero, and Karen H. Watanabe. "From Qualitative to Quantitative AOP: A Case Study of Neurodegeneration". *Frontiers in Toxicology* 4 (Mar. 2022), p. 30. DOI: 10.3389/ftox.2022.838729.
- [66] Clemens Wittwehr et al. "How adverse outcome pathways can aid the development and use of computational prediction models for regulatory toxicology". *Toxicological Sciences* 155.2 (Feb. 2017), pp. 326–336. DOI: 10.1093/toxsci/kfw207.
- [67] Rory B. Conolly et al. "Quantitative Adverse Outcome Pathways and Their Application to Predictive Toxicology". *Environmental Science and Technology* 51.8 (Apr. 2017), pp. 4661–4672.
- [68] Mathieu Vinken. "Omics-based input and output in the development and use of adverse outcome pathways". *Current Opinion in Toxicology* 18 (Dec. 2019), pp. 8–12. DOI: 10.1016/j.cotox.2019.02.006.
- [69] Erica K. Brockmeier et al. "The Role of Omics in the Application of Adverse Outcome Pathways for Chemical Risk Assessment". *Toxicological Sciences* 158.2 (Aug. 2017), pp. 252–262. DOI: 10.1093/toxsci/kfx097.
- [70] Vinita Chauhan et al. "Bringing together scientific disciplines for collaborative undertakings: a vision for advancing the adverse outcome pathway framework". *International Journal of Radiation Biology* 97.4 (2021), pp. 431–441. DOI: 10.1080/09553002.2021.1884314.
- [71] Noffisat O. Oki et al. "Integrated analysis of in vitro data and the adverse outcome pathway framework for prioritization and regulatory applications: An exploratory case study using publicly available data on piperonyl butoxide and liver models". *Toxicology in Vitro* 54 (Feb. 2019), pp. 23–32. DOI: 10.1016/j.tiv.2018.09.002.

- [72] Penny Nymark et al. "A Data Fusion Pipeline for Generating and Enriching Adverse Outcome Pathway Descriptions". *Toxicological Sciences* 162.1 (Mar. 2018), pp. 264–275. DOI: 10.1093/toxsci/kfx252.
- [73] Shannon M. Bell et al. "Integrating publicly available data to generate computationally predicted adverse outcome pathways for fatty liver". *Toxicological Sciences* 150.2 (Apr. 2016), pp. 510–520. DOI: 10.1093/toxsci/kfw017.
- [74] Noffisat O. Oki et al. "Accelerating Adverse Outcome Pathway Development Using Publicly Available Data Sources". *Current Environmental Health Reports* 3.1 (Mar. 2016), pp. 53–63. DOI: 10.1007/s40572-016-0079-y.
- [75] Allan Peter Davis et al. "Comparative Toxicogenomics Database (CTD): Update 2021". *Nucleic Acids Research* 49.D1 (Jan. 2021), pp. D1138–D1143. DOI: 10.1093/nar/gkaa891.
- [76] Noffisat O. Oki and Stephen W. Edwards. "An integrative data mining approach to identifying adverse outcome pathway signatures". *Toxicology* 350-352 (Mar. 2016), pp. 49–61. DOI: 10.1016/j.tox.2016.04.004.
- [77] Xiaoqing Wang et al. "Integration of Computational Toxicology, Toxicogenomics Data Mining, and Omics Techniques to Unveil Toxicity Pathways". *ACS Sustainable Chemistry & Engineering* 9.11 (Mar. 2021), pp. 4130–4138. DOI: 10.1021/acssuschemeng.0c09196.
- [78] Richard Cyganiak, David Wood, and Markus Lanthaler. *RDF 1.1 Concepts and Abstract Syntax*. 2014.
- [79] Deborah L McGuinness and Frank van Harmelen. *OWL Web Ontology Language Overview*. <https://www.w3.org/TR/owl-features/>. 2004.
- [80] Julie A. McMurry et al. "Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data". *PLoS Biology* 15.6 (June 2017), e2001414. DOI: 10.1371/journal.pbio.2001414.
- [81] Mark D. Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific Data* 3.1 (Mar. 2016), p. 160018. DOI: 10.1038/sdata.2016.18.
- [82] Annika Jacobsen et al. "FAIR Principles: Interpretations and Implementation Considerations". *Data intelligence* 2.4 (Jan. 2020), pp. 10–29. DOI: 10.1162/dint_r_00024.
- [83] Barry Hardy et al. "Toxicology ontology perspectives". *ALTEX* 29.2 (May 2012), pp. 139–156. DOI: 10.14573/altex.2012.2.139.
- [84] Olga Tcheremenskaia et al. "OpenTox predictive toxicology framework: Toxicological ontology and semantic media wiki-based OpenToxipedia". *Journal of Biomedical Semantics* 3.1 (Apr. 2012), pp. 1–17. DOI: 10.1186/2041-1480-3-s1-s7.
- [85] M. Cases, M. Pastor, and F. Sanz. "The eTOX Library of Public Resources for in Silico Toxicity Prediction". *Molecular Informatics* 32.1 (Jan. 2013), pp. 24–35. DOI: 10.1002/MINF.201200099.
- [86] Nina Jeliazkova et al. "The eNanoMapper database for nanomaterial safety information". *Beilstein Journal of Nanotechnology* 6.1 (2015), pp. 1609–1634. DOI: 10.3762/bjnano.6.165.
- [87] Janna Hastings et al. "eNanoMapper: Harnessing ontologies to enable data integration for nanomaterial risk assessment". *Journal of Biomedical Semantics* 6.1 (Mar. 2015), p. 10. DOI: 10.1186/s13326-015-0005-5.

-
- [88] Nikolay Kochev et al. "Your Spreadsheets Can Be FAIR: A Tool and FAIRification Workflow for the eNanoMapper Database". *Nanomaterials* 2020, Vol. 10, Page 1908 10.10 (Sept. 2020), p. 1908. DOI: 10.3390/NANO10101908.
- [89] Egon L. Willighagen et al. "Computational toxicology using the OpenTox application programming interface and Bioclipse". *BMC Research Notes* 4.1 (Nov. 2011), pp. 1–9. DOI: 10.1186/1756-0500-4-487.
- [90] Nina Jeliaskova and Vedrin Jeliaskov. "AMBIT RESTful web services: an implementation of the OpenTox application programming interface". *Journal of Cheminformatics* 3.1 (Dec. 2011), p. 18. DOI: 10.1186/1758-2946-3-18.
- [91] Cataia Ives et al. "Creating a Structured AOP Knowledgebase via Ontology-Based Annotations." *Applied in vitro toxicology* 3.4 (Dec. 2017), pp. 298–311. DOI: 10.1089/aivt.2017.0017.
- [92] Lyle D. Burgoon. "The AOPontology: A semantic artificial intelligence tool for predictive toxicology". *Applied In Vitro Toxicology* 3.3 (Sept. 2017), pp. 278–281. DOI: 10.1089/aivt.2017.0012.
- [93] Emma Arnesdotter et al. "Derivation, characterisation and analysis of an adverse outcome pathway network for human hepatotoxicity". *Toxicology* 459 (July 2021), p. 152856. DOI: 10.1016/J.TOX.2021.152856.
- [94] Nicoleta Spinu et al. "Development and analysis of an adverse outcome pathway network for human neurotoxicity". *Archives of Toxicology* 93.10 (Oct. 2019), pp. 2759–2772. DOI: 10.1007/S00204-019-02551-1.

2

WikiPathways: connecting communities

Adapted from: Marvin Martens et al. “WikiPathways: connecting communities”. *Nucleic Acids Research* 49.D1 (Jan. 2021), pp. D613–D621. DOI: [10.1093/nar/gkaa1024](https://doi.org/10.1093/nar/gkaa1024).

Abstract

WikiPathways (wikipathways.org) is a biological pathway database known for its collaborative nature and open science approaches. With the core idea of the scientific community developing and curating biological knowledge in pathway models, WikiPathways lowers all barriers for accessing and using its content. Increasingly more content creators, initiatives, projects and tools have started using WikiPathways. Central in this growth and increased use of WikiPathways are the various communities that focus on particular subsets of molecular pathways such as for rare diseases and lipid metabolism. Knowledge from published pathway figures helps prioritize pathway development, using optical character and named entity recognition. We show the growth of WikiPathways over the last three years, highlight the new communities and collaborations of pathway authors and curators, and describe various technologies to connect to external resources and initiatives. The road toward a sustainable, community-driven pathway database goes through integration with other resources such as Wikidata and allowing more use, curation and redistribution of WikiPathways content.

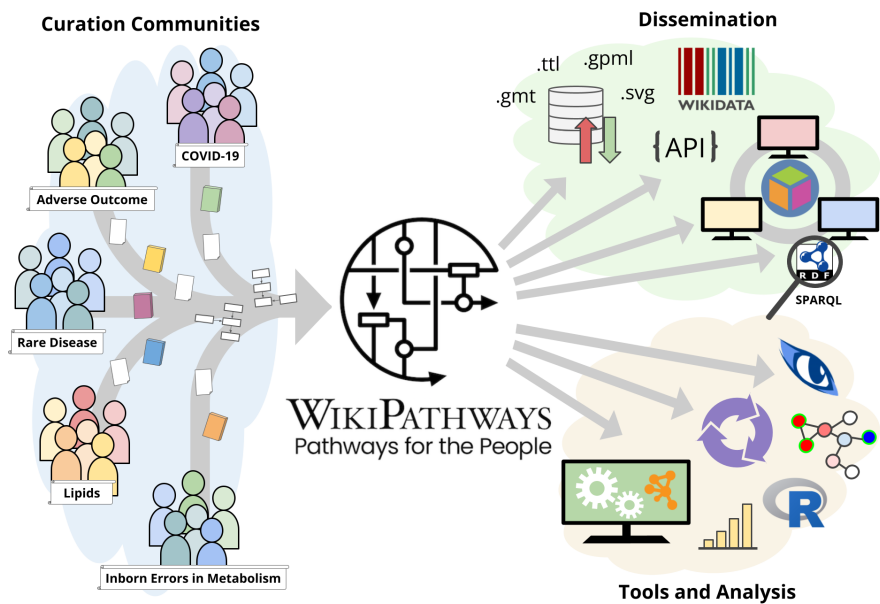


Figure 2.1: WikiPathways Graphical Abstract. WikiPathways enables research communities to collaborate on molecular pathway curation to create reusable, machine-readable pathway models. The pathway collections are freely available and integrated with many analysis tools and resources.

2.1 Introduction

The WikiPathways project was founded in 2007 upon the idea that everyone should be able to participate in the collection and curation of scientific knowledge [1]. No single research team can match the diversity and depth of expertise represented by the greater scientific community. Putting the tools of content creators and database maintainers into the hands of content consumers completes a virtuous cycle that powers growth and quality control which scales with the acquisition of new knowledge. Our approach is reflected in open science and FAIR principles [2].

WikiPathways is a database of biological pathway models collected and curated by the research community. Anyone at any time can contribute their pathway knowledge using freely available pathway editing tools. All edits are attributed to a registered author and screened by at least one other curator by means of organized and distributed community curation. This approach allows WikiPathways to grow at the scale of new discoveries and with input from diverse sources of pathway knowledge.

As previously reported, WikiPathways relies on communities of pathway authors and curators, pathway users, and developers to assemble, update and distribute content for myriad research applications [1, 3–7] In this update, we highlight unparalleled growth in content with more than seventy new pathways and thousands of revisions each year. We also present several research communities that we have collaborated with and empowered, including the COVID-19 Disease Map project [8], LIPID MAPS [9] and the rare disease community. Furthermore, to strengthen community building and curation, we started organizing monthly *Curation Cafe* events focused on selected topics, e.g. improving the quality of existing or creating novel pathways. We also detail some of the latest infrastructure updates, tool development and dissemination work which improve the free exchange of pathway information across platforms and within common analytical workflows.

2.2 Content and general updates

In the three years since our last update [7], over 70 new pathways per year (on average) were added to our data release (releases.wikipathways.org). In this section, we report on the content updates between the data releases on 10 September 2017 and 10 September 2020. Overall, WikiPathways currently contains a total of 2,857 pathways for all species, out of which 1,777 are included in the species-specific analysis collections. In the last three years, the content at WikiPathways has seen 10,079 user contributions, and 122 new contributors joined our community (Figure 2.2A). Although the content at WikiPathways represents human biology to a large extent, a total of 29 species are supported including vertebrates, invertebrates, plants, eukaryotic microorganisms and bacteria. Our human pathway collection has been extended consistently with 242 pathways (Figure 2.2B), and 9,014 genes and proteins, of which 12% are new to the database. Furthermore, of the 1,886 metabolites added, 69% were new, as a result of a concerted effort on metabolite curation. These datanodes are connected by 46,105 interactions, of which there are currently 4,026 more than in September 2017 (Figure 2.2C).

Based on data from Google Analytics in the last three years, the main WikiPathways website has recorded on average 700 visitors per day an international audience (33% from North America, 32% from Asia, 25% from Europe). Additionally, our REST webservice API recorded 27 million requests during this three-year time span. Importantly, in an effort to produce a more accessible and sustainable resource, we regularly disseminate pathway content to third-party tools and databases (see section “Connections to other initiatives”), which generate secondary usage statistics not reported here.

2.2.1 Pathway lifecycle

New biological knowledge is published every day, and as part of the curation process, pathway models get revised over time with this new information, in addition to other updates and corrections. Quality as-

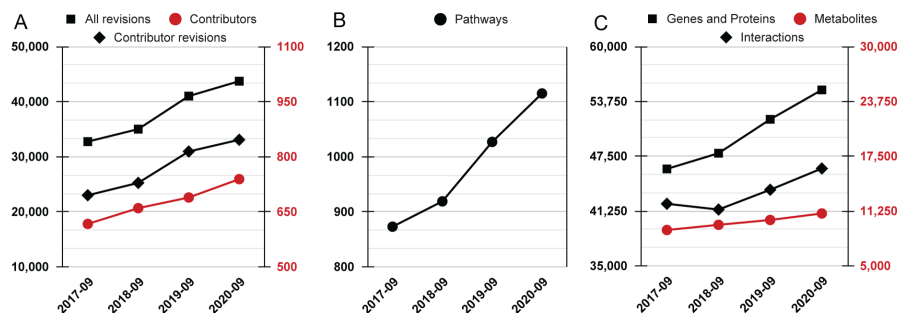


Figure 2.2: Recent growth of WikiPathways. The y-axes represent data for the human pathways accumulated in the database (approved content via rdf.wikipathways.org), focusing on the last three years (x-axes). A. Total counts for all revisions, including contributor and automated revisions (black squares), the subset of revisions made by contributors (black diamonds) and individual contributors (red circles). B. Total human pathway count. C. Total counts for genes and proteins (black squares), interactions (black diamonds) and metabolites (red circles). Data colors match corresponding y-axes.

surance of the WikiPathways content is accomplished continuously, by a combination of a weekly manual curation by a member of an organized team of curators, computer-assisted curation processes [7] and monthly curation cafes. Our manual curation protocol is designed as an interactive set of tasks which cover a wide range of topics, from recent edits and additions, assessing redundancy and overlap, to problematic content. It has been used successfully for the past three years and has greatly streamlined and standardized the process. The ease-of-use has also made it easy to bring in new contributors. Additionally, the computer-assisted curation tools are used effectively and its repertoire of tests has been expanded to better support the ongoing curation efforts and challenges in the WikiPathways database (github.com/BiGCAT-UM/WikiPathwaysCurator).

Each edit in a wiki-based system is recorded as a revision in the pathway history. These revisions are a measure for community activity and engagement. Pathway edits cover adding new biological knowledge, annotating the pathways with metadata (description, ontology

tags), improving the layout of a pathway diagram, and any combination thereof. Figure 2.3 shows that pathways from the human pathway analysis collection are updated regularly. While only 6 out of 639 pathways have not been updated yet, 14 pathways have more than 100 revisions including the "Integrated Breast Cancer Pathway" (WP1984, 445 revisions, 14 curators, [10]), the "Aryl Hydrocarbon Receptor" pathway from NetPath (WP2586, 256 revisions, 9 curators, [11, 12]), and the "Selenium Micronutrient Network" (WP15, 208 revisions, 15 curators, [13]), displaying that the collaborative nature of WikiPathways is a clear asset to pathway curation.

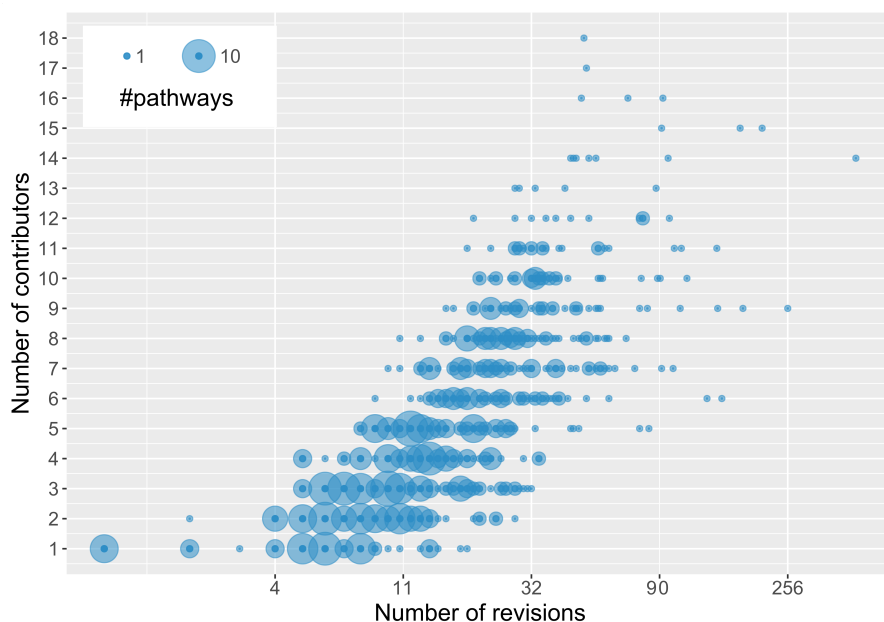


Figure 2.3: The number of revisions and contributors for all pathways in the human pathway analysis collection. The x-axis depicts how often pathways were updated (revised) on a logarithmic scale. The y-axis depicts the number of contributors who have worked on a pathway. The size of the dots expresses the number of pathways for that combination of contributors and revisions.

2.2.2 Pathway figures in published literature

Even with the continued growth at WikiPathways, approximately only 50% of protein-coding human genes are represented in pathway models. Despite the availability of free pathway modelling tools, such as PathVisio [14], CellDesigner [15] or Newt [16], the vast majority of pathway content is still shared as static images in published figures. Each month, an estimated 1,000 figures representing pathway content are published and collected at PubMed Central (PMC) [17]. Hence, the WikiPathways project initiated an analysis to convert published pathway figures into pathway models, using a pipeline beginning with a PMC image search, followed by machine learning, optical character recognition and named entity recognition. We identified 64,643 pathway figures published over the past 25 years and extracted 1,112,551 human genes, representing 13,464 unique genes [18]. These include over 3,600 genes not previously included in WikiPathways nor Reactome collections (as of January 2020). Based on enrichment analysis of disease-annotated gene sets against these pathway figures, the genes represent a wide range of diseases, including various types of cancer, cardiomyopathy, and diabetes. Prioritizing novel genes and rare diseases, we are using these published pathway figures as starting points for collaborative curation events. We have made all of the pathway figure content available via an interactive web interface (gladstone-bioinformatics.shinyapps.io/shiny-25years).

2.3 Pathway curation communities

The collaboration with various communities is an essential part of WikiPathways, where portals serve as a functional framework for communities with focused pathway interests (portals.wikipathways.org). Portal maintenance instructions are provided to enable communities to design and maintain portals themselves, with assistance from the WikiPathways team if needed. Here, we highlight recent community efforts in collaboration with

WikiPathways. All pathways receive a tag specifically for their community, allowing for automated downloads of these pathway collections through the REST API (webservice.wikipathways.org) with the "getCurationTagsByName" function, and our RDF format with SPARQL queries (rdf.wikipathways.org, [6]).

2.3.1 COVID-19

The COVID-19 Disease Map project aims to understand biological processes relevant to the COVID-19 pandemic [8]. From the start of this international effort, WikiPathways has been committed to contributing to this initiative by building, curating and sharing pathway models with a liberal license (CC0) and under FAIR community standards. Currently, the WikiPathways COVID-19 portal (covid.wikipathways.org, see Figure 2.4) contains a collection of eleven molecular pathways on SARS-CoV-2 itself, nine on other coronaviruses from earlier outbreaks, and several known processes involving ACE2, the main target membrane enzyme of SARS-CoV-2 for entering host cells. Identifiers and cross-references for coronavirus genes and proteins are provided through a Wikidata project [19]. Our pathway models are regularly updated and integrated into the COVID-19 Disease Map. For this initiative, we are currently adapting the data model to allow better support for multi-species pathways, annotation of evidence information and annotation of complexes.

2.3.2 Rare diseases

Rare diseases affect relatively few people, with the exact definition varying between 5 and 80 individuals per 100,000 for a given rare disease. However, it is estimated that up to 5.9% of the general population is affected by a rare disease. The majority of these disorders are genetic, 4,440 of 6,172 in total counted by ORPHANET [20]. For many disease-causing genes, there is little known about the normal gene function, and this knowledge is scattered over scientific publications and databases. Within WikiPathways there is a specialized por-

Portal:COVID-19

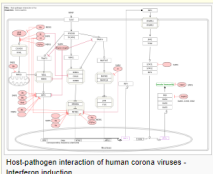
Welcome to the COVID-19 Pathways Portal on WikiPathways	Featured Pathway
<p>This special subset of disease pathways is being highlighted during the current COVID-19 crisis. This content is released under a CC0 waiver to be freely used, re-used and distributed. Let us know if you add a new pathway or want to recommend one for this collection.</p> <p>Collaborators</p> <p>Authors</p> <p>The current list of authors and contributors of content of this portal (in alphabetical order): Matthew Conroy, Rex D A B, Jordi Dojten, Lauren Dupuis, Friederike Ehrhart, Chris Evelo, Kristina Hanspers, Amber Koning, Martina Kutzon, Marvin Mariens, Penny Nyman, Alex Pico, Denise Slinger, Caroline Thorn, the Reactome project, and many more.</p> <p>Publications</p> <p>A paper about this project was published in <i>Scientific Data</i> 9, Ostaszewski, et al. "COVID-19 Disease Map, building a computational repository of SARS-CoV-2 virus-host interaction mechanisms." <i>Scientific data</i> 7.1 (2020): 1-4.</p>	<p>Host-pathogen interaction of human corona viruses - Interferon induction (Homo sapiens)</p>  <p>Host-pathogen interaction of human corona viruses - Interferon induction</p> <p>View all Featured Pathways for this Portal</p>

Figure 2.4: The COVID-19 Portal on WikiPathways (covid.wikipathways.org). The portal contains relevant information for COVID-19-related research, including all molecular pathways, contributors, and publications.

tal for exploring, curating, and expanding the collection of rare disease pathways (raredisease.wikipathways.org), partnered by EJP-RD (European Joint Programme on Rare Diseases), ELIXIR (European Bioinformatics infrastructure programme) and the Dutch Rett expertise centre (Maastricht University Medical Centre). The portal is used to capture knowledge from literature and data to gain a better understanding of these complex disorders. The pathways are created and curated in collaboration with disease experts, currently covering over 60 rare diseases including very different types of diseases, e.g. laminopathies, ciliopathies, disorders of sexual development and fertility, and copy number variation syndromes.

2.3.3 Inborn Errors of Metabolism (IEM)

Inborn errors of metabolism are a subsection of the rare disease field, which are captured in the "IEM portal" (iem.wikipathways.org), containing molecular pathways connecting clinical biomarkers to disorders. We have started a collaboration with the authors of the book "Physician's Guide to the Diagnosis, Treatment, and Follow-Up of Inherited Metabolic Diseases" [21] and are currently processing all included pathways, as well as integrating these

in the IEMbase [22]. The portal currently covers 19 chapters, 23 approved pathways, over 350 diseases linked to OMIM identifiers and 68 unique Disease Ontology terms, and is expected to be expanding with the coverage of additional chapters. Examples of data analysis for these (and other) pathways can be found at bigcat-um.github.io/PathwayAnalysisBlauBook. The disease nodes are currently represented as Labels with hyperlinks in the pathway models. We are planning to extend the data model with a non-molecular data node (e.g. Annotation / Phenotype) that will not be used for data analysis but can be annotated with a proper identifier for computational processing of the information.

2.3.4 Lipids

Lipids are a fascinating class of chemical compounds that serve several roles within organisms and are difficult to measure in a wet lab setting. The LIPID MAPS team has initiated a collaboration [9] with the WikiPathways community to maintain and extend their lipid pathway content, leading to the addition of nine highly curated lipid pathways for mouse, the original pathways are available at lipidmaps.org/resources/pathways/vanted.php. These pathways have been homology converted to their human counterpart and are now part of the Lipids Portal (lipids.wikipathways.org). Annotating individual lipids instead of lipid classes can be quite complicated; this phenomenon is in most cases due to a lack of biological knowledge on individual lipids. Furthermore, several cases are known where homology mapping between different species for proteins is hampered, e.g. for stearoyl-CoA desaturase-1 having four isoforms in mice compared to only two in humans [23].

2.3.5 Adverse Outcome Pathways

Since the introduction of Adverse Outcome Pathways (AOPs) to support regulatory decision making for risk assessment of chemicals [24], the primary focus of AOP research groups has been

capturing mechanistic data in written format. However, since the majority of biological processes described in AOPs are biological pathways that exist as pathway models on WikiPathways, the AOP Portal (aop.wikipathways.org) has been created to capture all pathways relevant to toxicological assessments [25]. These molecular AOPs contribute to the understanding of AOPs and facilitate the use of various omics approaches in risk assessments [26]. One challenge lies in the unique rationale behind molecular AOPs, where biological processes are connected in a chain of Key Events (KEs) that make an AOP, rather than presenting one molecular pathway. Second, KEs often describe disturbances or adverse responses already captured in molecular pathways in WikiPathways [25] and therefore AOPs are modelled as meta-pathways. These combine pathway nodes and KE nodes in one data model, which is linked to the AOP-Wiki, (aopwiki.org).

2.4 Connections to other initiatives

WikiPathways enables anyone to freely share, redistribute, use, and adapt pathway content in the database, by removing any barriers for people to decide to contribute to or use WikiPathways. Furthermore, WikiPathways provides a variety of options to access the data for use, through downloads in various formats for individual pathways or pathway collections, from the pathway editor and analysis software PathVisio [14], through the WikiPathways REST API and rWikiPathways R package, or through the WikiPathways SPARQL endpoint. These aspects make WikiPathways content easy to implement in services, tools, workflows or distributions.

2.4.1 BridgeDb

Managing molecular pathways requires robust use of database identifiers for all pathway components (genes, proteins, metabolites, complexes, diseases, interactions). Recently, mappings to the EBI Complex Portal [27] and IUPHAR/BPS Guide to PHARMACOLOGY [28] have

been added to the identifier mapping framework BridgeDb [29] which is integrated with WikiPathways.

2.4.2 Wikimedia Toolforge

The pathway viewer widget was updated and moved to Wikimedia Toolforge allowing users to integrate the pathways from WikiPathways into their own website with more ease (widget.wikipathways.org). The widget unifies identifiers from the data source originally specified by the pathway author, to provide several commonly used data sources for additional exploration. Metabolite identifiers are unified by BridgeDb to include ChEBI, HMDB and Wikidata identifiers. For gene products, the Wikidata API is used for mapping of NCBI Gene, Ensembl, HGNC and Wikidata identifiers. In the future, the mapping for gene products will also be done using BridgeDb.

2.4.3 SPARQL explorer

To make access to the WikiPathways RDF more user-friendly, we have introduced Wikipathways SNORQL (github.com/wikipathways/snorql-extended), a new extended implementation of the SNORQL user interface (UI), to be our go-to semantic web browser (sparql.wikipathways.org). WikiPathways SNORQL is a query editor that offers syntax highlighting for writing and executing SPARQL queries directly on our existing SPARQL endpoint (sparql.wikipathways.org/sparql). Furthermore, the user interface provides a query examples panel which is auto-populated with SPARQL queries from a customizable GitHub repository (github.com/wikipathways/SPARQLQueries). This repository stores queries in folders divided over particular topics, communities, collaborations functionality, or external data sources for federated queries, allowing new users to navigate through our example queries panel with more ease (Figure 2.5). Overall, the new UI allows collecting, storing, exploring and reusing SPARQL queries

on the WikiPathways data.

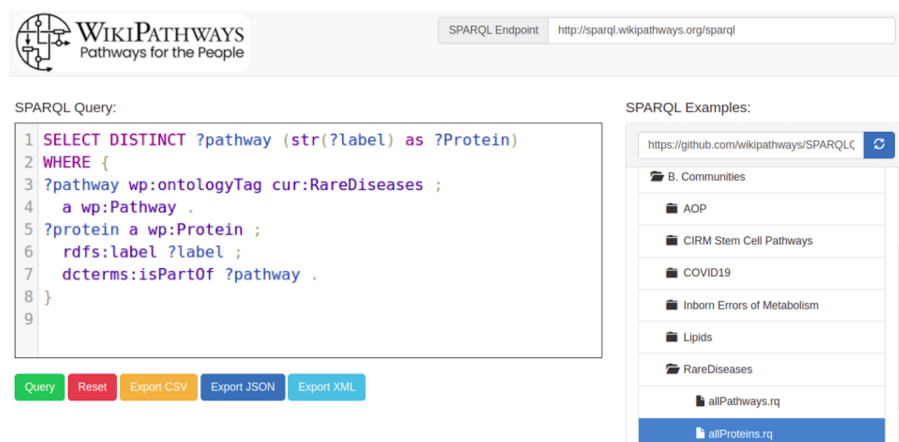


Figure 2.5: The WikiPathways SPARQL user interface (sparql.wikipathways.org). The WikiPathways SPARQL semantic web browser allows for user-friendly access to the WikiPathways RDF through providing example SPARQL queries (right panel).

One example of semantic interoperability using federated SPARQL queries is our collaboration with neXtProt. The neXtProt [30] knowledge resource of the Swiss Institute of Bioinformatics aims to document inter- and intra-individual diversity of human proteins by integrating information from a variety of protein resources. To extend the knowledge on proteins towards systems and biological pathways, federated SPARQL queries have been developed in collaboration, harnessing the semantic web capabilities to connect neXtProt with WikiPathways knowledge.

2.4.4 Wikidata

The CC0 license of WikiPathways also enabled interoperability through Wikidata, the linked data repository of Wikipedia [31]. Like Wikipedia, Wikidata is a knowledge-sharing platform open to all (humans and software). In collaboration with the Gene Wiki

and Reactome teams, we developed bots to add information about the curated pathways to Wikidata [19, 32]. The WikiPathways bot creates Wikidata items for each pathway and the content thereof in WikiPathways and aligns those with the Wikidata items on associated genes, proteins, metabolites, literature citations, and ontology annotations (e.g., [wikidata.org/wiki/Q28031254](https://www.wikidata.org/wiki/Q28031254)). The WikiPathways content of the human pathway analysis collection in Wikidata gets updated after each monthly release.

2.4.5 Scholia

Scholia is a graphical user interface that aggregates information from Wikidata around topics [33, 34], such as genes, proteins, metabolites, pathways, authors, articles, and organizations. In collaboration with the Scholia team, we developed Scholia templates for WikiPathways pathways and all pathways in Wikidata are now also addressable as Scholia topic pages scholia.toolforge.org/wikipathways/WP111. These pages show the participants of the pathway (genes, protein, metabolites), the literature cited by the pathway, and articles citing the pathway. Moreover, similarly to linking to PubMed and Europe PMC, the literature section now links to Scholia pages for the cited articles.

2.4.6 Nanopub

Nanopublications for WikiPathways have been released for several years now in collaboration with their international community [35]. The nanopublications are created using a combination of `nanopub-java` library [36] and SPARQL queries against the WikiPathways RDF (see github.com/wikipathways/nanopublications). This results in three types of nanopublications, for three types of facts in WikiPathways: complexes, interactions, and general participation in pathways. Nanopublications are currently only generated if the complex, interaction, or participation is linked to a specific literature reference, identified by a PubMed identifier, which is used as part of the

provenance of the nanopublication. Nanopublications are findable using semantic web identifiers for genes and proteins [37], but using the pathway identifier we can also find all nanopublications that originate from the corresponding pathway, such as WP15, for example with

```
curl -X GET "http://grlc.np.dumontierlab.  
com/api/local/local/find_nanopubs_with_uri  
?ref=http://identifiers.org/wikipathways/  
WP15_r107118" -H "accept: text/csv"  
with a command line.
```

2.4.7 Europe PMC and other PubMed interoperability

The PubMed identifier is still the primary, global identifier used by WikiPathways to identify the literature cited. In 2018 we started contributing links between PubMed articles and pathways in WikiPathways (excluding the Reactome-synced pathways [5]) to Europe PMC via the External links service (europepmc.org/LabsLink) functionality [38]. This allows Europe PMC to show the pathways from WikiPathways that mention that article in their database. The WikiPathways website now also links to Europe PMC for cited articles in the literature section on a pathway page, making the integration bi-directional.

2.4.8 Enrichment analysis tools

Functional enrichment analysis is a popular approach for characterizing differentially expressed genes based on Gene Ontology terms, pathways and other annotated gene sets. We release a standard Gene Matrix Transposed (GMT) file each month that includes the latest set of curated pathways approved for enrichment use cases. Using this file, WikiPathways content can be added to any protocol supporting the GMT standard. A number of R packages and online tools that perform enrichment analysis have incorporated WikiPathways into their methods and vignette examples, including g:Profiler [39], clusterProfiler [40], rSEA [41], Enrichr [42], IMPaLa [43] and WebGestalt [44].

Human and mouse WikiPathways gene sets are now also available in the Molecular Signatures Database (MSigDb, [45]) for the GSEA software [46].

2.4.9 MINERVA

The interoperability between the MINERVA platform [47] and the WikiPathways GPML got a boost with the COVID-19 Disease Map project [8]. In this large international effort, it is crucial to be able to communicate between the different resources to share, collect and unify the content. The MINERVA software can now import and export GPML files from WikiPathways. This facilitates the integration of WikiPathways pathways in the larger disease map but also allows export of other models to GPML enabling distribution of the content in RDF format in the future.

2.4.10 BEL Ecosystem

Several researchers involved in the Biological Expression Language (BEL) project [48] are harmonizing the information of different pathway databases including WikiPathways. Bio2BEL converts the content from several pathway databases into BEL and stores causal and correlative relations between biological entities across multiple modes and scales as a biological network [49]. ComPath aims to evaluate the coverage, agreements, and discrepancies between the pathway databases in terms of gene content [50]. PathMe provides normalizations between these pathway databases for other content (e.g. protein-protein interactions, complexes) [51]. This information is integrated into PathwayForte [52] and the feedback from these analyses led to additional curation efforts on WikiPathways including renaming of pathways.

2.4.11 Network Data Exchange - NDEx

The Network Data Exchange (NDEx) is a public resource for publishing and sharing biological networks and gene sets [53], and

WikiPathways is closely collaborating with the NDEx team. First, we regularly deposit the WikiPathways human pathway analysis collection into a dedicated collection at NDEx (ndexbio.org/#/user/363f49e0-4cf0-11e9-9f06-0ac135e8bacf), with over 600 pathways included. Second, as part of the Clinical Proteomic Tumor Analysis Consortium (CPTAC) initiative, we have led the curation of 28 cancer-specific pathway models and the organization of 87 cancer-related pathways, all of which we have merged into 11 network models representing common cancer hallmark categories (cptac.wikipathways.org). We regularly deposit these network models at NDEx. Third, we have extracted gene sets from over 32,000 pathway figures with 10 or more genes and have deposited them at NDEx [53] for manual download and programmatic access, as well as facilitated Cytoscape workflows and enrichment analysis via NDEx Integrated Query (iquery.ndexbio.org).

2.5 Future work

With the ever-growing collection of curated and published pathway information, the WikiPathways team is working towards a more sustainable and scalable infrastructure for all pathway knowledge. This work will involve the development of new tools and services, continued integration into community-run resources like Wikidata, and close coordination with other pathway databases and biocuration teams. For example, we are building a pathway knowledge management system using git version control, with automated diff and merge capabilities to synchronize curation efforts across multiple sites. Pathway edits made at Wikidata, NDEx, WikiPathways or Reactome would essentially open a request to merge and redistribute the new information. Utilizing gene, metabolite and disease content extracted from published pathway figures, we will organize focused curation efforts aimed at converting content as pathway models. This process will be facilitated by the addition of a portal highlighting and organizing this content.

We are also integrating pathway information into new platforms for biomedical research and discovery. This work includes the continued support for third-party pathway analysis tools online and via R and Python packages, as well as new knowledge bases that connect pathway information to other genomic and phenotypic models of biology. For example, we are contributing curated and published pathway information to the NCATS' Biomedical Data Translator program [54] using the BioThingsExplorer (`biothings-explorer.readthedocs.io`) and Smart API [55].

2.6 Conclusion

The content of WikiPathways increases every day due to the combined effort of multiple communities, including curators of pathway models and authors of pathway figures. A wide variety of third-party analytical tools utilizes the content distributed by WikiPathways, creating an immeasurable user base. WikiPathways is defined by the people authoring, curating and utilizing pathway knowledge. We invite all researchers interested in pathways to directly participate in the WikiPathways project.

2.7 Data availability

All WikiPathways data, including older data releases, are stored on data.wikipathways.org. Scripts and SPARQL queries used to generate the data published here can be found on GitHub [56].

References

- [1] Alexander R. Pico et al. "WikiPathways: Pathway Editing for the People". *PLoS Biology* 6.7 (July 2008), e184. DOI: 10.1371/journal.pbio.0060184.
- [2] Mark D. Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific Data* 3.1 (Mar. 2016), p. 160018. DOI: 10.1038/sdata.2016.18.

- [3] Thomas Kelder et al. "WikiPathways: building research communities on biological pathways". *Nucleic Acids Research* 40.D1 (Jan. 2012), pp. D1301–D1307. DOI: 10.1093/nar/gkr1074.
- [4] Martina Kutmon et al. "WikiPathways: Capturing the full diversity of pathway knowledge". *Nucleic Acids Research* 44.D1 (2016), pp. D488–D494. DOI: 10.1093/nar/gkv1024.
- [5] Anwesha Bohler et al. "Reactome from a WikiPathways Perspective". *PLoS Computational Biology* 12.5 (May 2016). DOI: 10.1371/journal.pcbi.1004941.
- [6] Andra Waagmeester et al. "Using the Semantic Web for Rapid Integration of WikiPathways with Other Biological Online Data Resources". *PLoS Computational Biology* 12.6 (June 2016). Ed. by Christos A. Ouzounis, e1004989. DOI: 10.1371/journal.pcbi.1004989.
- [7] Denise N. Slenter et al. "WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research". *Nucleic Acids Research* 46.D1 (Nov. 2018), pp. D661–D667. DOI: 10.1093/nar/gkx1064.
- [8] Marek Ostaszewski et al. "COVID-19 Disease Map, building a computational repository of SARS-CoV-2 virus-host interaction mechanisms". *Scientific Data* 7.1 (Dec. 2020), p. 136. DOI: 10.1038/s41597-020-0477-8.
- [9] Valerie B. O'Donnell et al. "LIPID MAPS: Serving the next generation of lipid researchers with tools, resources, data, and training". *Science Signaling* 12.563 (Jan. 2019), eaaw2964. DOI: 10.1126/scisignal.aaw2964.
- [10] S Ibrahim et al. *Integrated Breast Cancer Pathway (Homo sapiens)*. [wikipathways.org/instance/WP1984_r110683](https://www.wikipathways.org/instance/WP1984_r110683). 2020.
- [11] P Gupta et al. *Aryl Hydrocarbon Receptor Netpath (Homo sapiens)*. [wikipathways.org/instance/WP2586_r107439](https://www.wikipathways.org/instance/WP2586_r107439). 2019.
- [12] Kumaran Kandasamy et al. "NetPath: a public resource of curated signal transduction pathways". *Genome Biology* 11.1 (Jan. 2010), R3. DOI: 10.1186/gb-2010-11-1-r3.
- [13] User egoyenechea et al. *Selenium Micronutrient Network (Homo sapiens)*. [wikipathways.org/instance/WP15_r107118](https://www.wikipathways.org/instance/WP15_r107118). 2019.
- [14] Martina Kutmon et al. "PathVisio 3: An Extendable Pathway Analysis Toolbox". *PLoS Computational Biology* 11.2 (Feb. 2015). Ed. by Robert F. Murphy, e1004085. DOI: 10.1371/journal.pcbi.1004085.
- [15] Akira Funahashi et al. "CellDesigner: a process diagram editor for gene-regulatory and biochemical networks". *BIOSILICO* 1.5 (Nov. 2003), pp. 159–162. DOI: 10.1016/S1478-5382(03)02370-9.
- [16] Hasan Balci et al. "Newt: a comprehensive web-based tool for viewing, constructing and analyzing biological maps". *Bioinformatics* 37.10 (June 2021). Ed. by Wren Jonathan, pp. 1475–1477. DOI: 10.1093/bioinformatics/btaa850.
- [17] Anders Riutta, Kristina Hanspers, and Alexander R. Pico. "Identifying Genes in Published Pathway Figure Images". *bioRxiv* (July 2018), p. 379446. DOI: 10.1101/379446.
- [18] Kristina Hanspers et al. "25 Years of Pathway Figures". *bioRxiv* (May 2020), p. 2020.05.29.124503. DOI: 10.1101/2020.05.29.124503.
- [19] Andra Waagmeester et al. "A protocol for adding knowledge to Wikidata, a case report". *bioRxiv* (June 2020), p. 2020.04.05.026336. DOI: 10.1101/2020.04.05.026336.

-
- [20] Stéphanie Nguengang Wakap et al. "Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database". *European Journal of Human Genetics* 28.2 (Feb. 2020), pp. 165–173. DOI: 10.1038/s41431-019-0508-0.
- [21] Nenad Blau et al. *Physician's Guide to the Diagnosis, Treatment, and Follow-Up of Inherited Metabolic Diseases*. Ed. by Nenad Blau et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014. DOI: 10.1007/978-3-642-40337-8.
- [22] Jessica J.Y. Y Lee et al. "Knowledge base and mini-expert platform for the diagnosis of inborn errors of metabolism". *Genetics in Medicine* 20.1 (Jan. 2018), pp. 151–158. DOI: 10.1038/gim.2017.108.
- [23] Jian Wang et al. "Characterization of HSCD5, a novel human stearyl-CoA desaturase unique to primates". *Biochemical and Biophysical Research Communications* 332.3 (July 2005), pp. 735–742. DOI: 10.1016/j.bbrc.2005.05.013.
- [24] Gerald T. Ankley et al. "Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment". *Environmental Toxicology and Chemistry* 29.3 (Mar. 2010), pp. 730–741. DOI: 10.1002/etc.34.
- [25] Marvin Martens et al. "Introducing WikiPathways as a Data-Source to Support Adverse Outcome Pathways for Regulatory Risk Assessment of Chemicals and Nanomaterials". *Frontiers in Genetics* 9 (Dec. 2018), p. 661. DOI: 10.3389/fgene.2018.00661.
- [26] Mathieu Vinken. "Omics-based input and output in the development and use of adverse outcome pathways". *Current Opinion in Toxicology* 18 (Dec. 2019), pp. 8–12. DOI: 10.1016/j.cotox.2019.02.006.
- [27] Birgit H M Meldal et al. "Complex Portal 2018: extended content and enhanced visualization tools for macromolecular complexes". *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D550–D558. DOI: 10.1093/nar/gky1001.
- [28] Jane F. Armstrong et al. "The IUPHAR/BPS Guide to PHARMACOLOGY in 2020: Extending immunopharmacology content and introducing the IUPHAR/MMV Guide to MALARIA PHARMACOLOGY". *Nucleic Acids Research* 48.D1 (Jan. 2020), pp. D1006–D1021. DOI: 10.1093/nar/gkz951.
- [29] Martijn P van Iersel et al. "The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services". *BMC Bioinformatics* 11.1 (2010), p. 5. DOI: 10.1186/1471-2105-11-5.
- [30] Monique Zahn-Zabal et al. "The neXtProt knowledgebase in 2020: Data, tools and usability improvements". *Nucleic Acids Research* 48.D1 (Jan. 2020), pp. D328–D334. DOI: 10.1093/nar/gkz995.
- [31] I Trestian, K Hugueninz, L Su, et al. *Proceedings of the 21st International Conference Companion on World Wide Web*. ACM Press, 2012. DOI: 10.1145/2187836.
- [32] Andra Waagmeester et al. "Wikidata as a knowledge graph for the life sciences". *eLife* 9 (Mar. 2020). DOI: 10.7554/eLife.52614.
- [33] Finn Årup Nielsen, Daniel Mietchen, and Egon Willighagen. "Scholia, Scientometrics and Wikidata". *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 10577 LNCS. 2017, pp. 237–259. DOI: 10.1007/978-3-319-70407-4_36.
- [34] Lane Rasberry et al. "Robustifying Scholia: paving the way for knowledge discovery and research assessment through Wikidata". *Research Ideas and Outcomes* 5 (May 2019). DOI: 10.3897/rio.5.e35820.

- [35] Tobias Kuhn et al. "Nanopublications: A Growing Resource of Provenance-Centric Scientific Linked Data". *2018 IEEE 14th International Conference on e-Science (e-Science)*. IEEE, Oct. 2018, pp. 83–92. DOI: 10.1109/eScience.2018.00024.
- [36] Tobias Kuhn. "Nanopub-Java: A Java library for nanopublications". *CEUR Workshop Proceedings*. Vol. 1572. 2016, pp. 19–25.
- [37] Egon Willighagen. "Increasing the nanopublication recall with a BridgeDb Identifier Mapping Service". *Semantic Web Applications and Tools for Health Care and Life Sciences* 2275 (2018), pp. 1–6.
- [38] Maria Levchenko et al. "Europe PMC in 2017". *Nucleic Acids Research* 46.D1 (Jan. 2018), pp. D1254–D1260. DOI: 10.1093/nar/gkx1005.
- [39] Uku Raudvere et al. "G:Profiler: A web server for functional enrichment analysis and conversions of gene lists (2019 update)". *Nucleic Acids Research* 47.W1 (July 2019), W191–W198. DOI: 10.1093/nar/gkz369.
- [40] Guangchuang Yu et al. "ClusterProfiler: An R package for comparing biological themes among gene clusters". *OMICS A Journal of Integrative Biology* 16.5 (May 2012), pp. 284–287. DOI: 10.1089/omi.2011.0118.
- [41] Mitra Ebrahimipoor et al. "Simultaneous enrichment analysis of all possible gene-sets: Unifying self-contained and competitive methods". *Briefings in bioinformatics* 21.4 (July 2020), pp. 1302–1312. DOI: 10.1093/bib/bbz074.
- [42] Maxim V Kuleshov et al. "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update". *Nucleic Acids Research* 44.W1 (July 2016), W90–W97. DOI: 10.1093/nar/gkw377.
- [43] Atanas Kamburov et al. "Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA". *Bioinformatics* 27.20 (Oct. 2011), pp. 2917–2918. DOI: 10.1093/bioinformatics/btr499.
- [44] Yuxing Liao et al. "WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs". *Nucleic Acids Research* 47.W1 (July 2019). DOI: 10.1093/nar/gkz401.
- [45] Arthur Liberzon. "A description of the molecular signatures database (MSigDB) web site". *Methods in Molecular Biology* 1150 (2014), pp. 153–160. DOI: 10.1007/978-1-4939-0512-6_9.
- [46] Aravind Subramanian et al. "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles". *Proceedings of the National Academy of Sciences of the United States of America* 102.43 (Oct. 2005), pp. 15545–15550. DOI: 10.1073/pnas.0506580102.
- [47] David Hoksza et al. "MINERVA API and plugins: Opening molecular network analysis and visualization to the community". *Bioinformatics* 35.21 (2019), pp. 4496–4498. DOI: 10.1093/bioinformatics/btz286.
- [48] Charles Tapley Hoyt, Andrej Konotopez, and Christian Ebeling. "PyBEL: A computational framework for Biological Expression Language". *Bioinformatics* 34.4 (Feb. 2018), pp. 703–704. DOI: 10.1093/bioinformatics/btx660.
- [49] Charles Tapley Hoyt et al. "Integration of Structured Biological Data Sources using Biological Expression Language". *bioRxiv* 7 (May 2019), p. 631812. DOI: 10.1101/631812.
- [50] Daniel Domingo-Fernández et al. "PathMe: Merging and exploring mechanistic pathway knowledge". *BMC Bioinformatics* 20.1 (May 2019). DOI: 10.1186/s12859-019-2863-9.

-
- [51] Daniel Domingo-Fernández et al. "ComPath: an ecosystem for exploring, analyzing, and curating mappings across pathway databases". *npj Systems Biology and Applications* 4.1 (Dec. 2018), p. 43. DOI: 10.1038/s41540-018-0078-8.
- [52] Sarah Mubeen et al. "The Impact of Pathway Database Choice on Statistical Enrichment Analysis and Predictive Modeling". *Frontiers in Genetics* 10 (Nov. 2019). DOI: 10.3389/fgene.2019.01203.
- [53] Rudolf T. Pillich et al. "NDEx: A community resource for sharing and publishing of biological networks". *Protein Bioinformatics*. Vol. 1558. 2017, pp. 271-301. DOI: 10.1007/978-1-4939-6783-4_13.
- [54] Biomedical Data Translator Consortium and The Biomedical Data Translator Consortium. "The Biomedical Data Translator Program: Conception, Culture, and Community". *Clinical and Translational Science* 12.2 (Mar. 2019), pp. 91-94. DOI: 10.1111/cts.12592.
- [55] Amrapali Zaveri et al. "SmartAPI: Towards a more intelligent network of web APIs". *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 10250 LNCS. Springer Verlag, May 2017, pp. 154-169. DOI: 10.1007/978-3-319-58451-5_11.
- [56] Martina Summer-Kutmon, Marvin Martens, and Denise Slenter. *wikipathways/Scripts_NAR2021: NAR 2021 Submission*. Sept. 2020. DOI: 10.5281/zenodo.4031648.

3

Introducing WikiPathways as a Data-Source to Support Adverse Outcome Pathways for Regulatory Risk Assessment of Chemicals and Nanomaterials

Adapted from: Marvin Martens et al. "Introducing WikiPathways as a Data-Source to Support Adverse Outcome Pathways for Regulatory Risk Assessment of Chemicals and Nanomaterials". *Frontiers in Genetics* 9 (Dec. 2018), p. 661. DOI: 10.3389/fgene.2018.00661.

Abstract

A paradigm shift is taking place in risk assessment to replace animal models, reduce the number of economic resources, and refine the methodologies to test the growing number of chemicals and nanomaterials. Therefore, approaches such as transcriptomics, proteomics, and metabolomics have become valuable tools in toxicological research, and are finding their way into regulatory toxicity. One promising framework to bridge the gap between the molecular-level measurements and risk assessment is the concept of adverse outcome pathways (AOPs). These pathways comprise mechanistic knowledge and connect biological events from a molecular level toward an adverse effect outcome after exposure to a chemical. However, the implementation of omics-based approaches in the AOPs and their acceptance by the risk assessment community is still a challenge. Because the existing modules in the main repository for AOPs, the AOP Knowledge Base (AOP-KB), do not currently allow the integration of omics technologies, additional tools are required for omics-based data analysis and visualization. Here we show how WikiPathways can serve as a supportive tool to make omics data interoperable with the AOP-Wiki, part of the AOP-KB. Manual matching of key events (KEs) indicated that 67% could be linked with molecular pathways. Automatic connection through linkage of identifiers between the databases showed that only 30% of AOP-Wiki chemicals were found on WikiPathways. More loose linkage through gene names in KE and Key Event Relationships descriptions gave an overlap of 70 and 71%, respectively. This shows many opportunities to create more direct connections, for example with extended ontology annotations, improving its interoperability. This interoperability allows the needed integration of omics data linked to the molecular pathways with AOPs. A new AOP Portal on WikiPathways is presented to allow the community of AOP developers to collaborate and populate the molecular pathways that underlie the KEs of AOP-Wiki. We conclude that the integration of WikiPathways and AOP-Wiki will improve risk assessment because

omics data will be linked directly to KEs and therefore allow the comprehensive understanding and description of AOPs. To make this assessment reproducible and valid, major changes are needed in both WikiPathways and AOP-Wiki.

3.1 Introduction

The last decades have seen many developments in risk assessment strategies for an ever-growing number of chemicals and nanomaterials, aiming to reduce the use of animals and cost of risk assessment and to increase the predictive value. In parallel to these changes, experimental approaches in regular toxicology research have also made major steps setting up novel high-throughput technologies for generating large-scale (omics) datasets such as transcriptomics, metabolomics, and proteomics. However, these technologies are not consistently implemented in regulatory risk assessment and there is a need for proper integration of knowledge, testing systems, and analysis tools for these approaches to be of added value over existing methodologies in risk assessment.

To support the paradigm shift toward animal-free, cheap and more effective risk assessments of chemicals, the concept of adverse outcome pathways (AOPs) emerged [1], which integrate mechanistic knowledge of the toxicological effects of chemical compounds and nanomaterials and thereby assist integrated approaches to testing and assessment strategies. AOPs are structured as logical sequences of causally linked and measurable biological events [key events (KEs)] that occur after exposure to a stressor that triggers a biological perturbation, called the molecular initiating event (MIE). These KEs are connected by Key Event Relationships (KERs) and describe the downstream effects on increasing levels of biological organization, from molecular, cellular, tissue, organ, individual, and population responses toward an adverse outcome (AO) [2–4].

The Organisation for Economic Co-operation and Development (OECD) was the first organization to embrace AOPs by launching the AOP Development Programme in 2012 for the establishment of AOPs in a qualitative way and provide guidance material for standardized, structured development of AOPs [5, 6]. With that, the AOP Knowledge Base (AOP-KB (aopkb.oecd.org/index.html)) emerged in 2014 as a collective platform of various tools to assist in

the development of AOPs. Its main components are the AOP-Wiki (aopwiki.org), Effectopedia (effectopedia.org) and the AOPXplorer Cytoscape application.

The AOP-Wiki is the result of collaboration between the European Commission's Joint Research Center (JRC) and the United States Environmental Protection Agency (US EPA). It is developed to be a central knowledge-sharing platform which facilitates cooperative development of AOPs and strictly follows the OECD's guidance materials for AOP development. Nowadays, it is the most actively used module of the AOP-KB and with the recent efforts on annotation with ontology tags, it has been aiming for semantic interoperability. This started with the development of the AOP Ontology [7] and recently, the addition of various other ontologies to match the various domains described in AOPs, from Gene Ontology for biology annotation toward the Population and Community Ontology for annotation of events on the population level [8].

Effectopedia [9] is another tool from AOP-KB, developed by OECD, dedicated to the collaborative development of quantitative AOPs. The AOP diagram is the focal point of its user interface providing visual means for adding new and navigation through existing AOP elements, offering easy access to their description. In addition to KE and KER, Effectopedia also has an explicit representation of test methods, collected data and executable models. The integration of response data in KER allows the system to predict downstream KEs using measurements or models for upstream KEs that can be measured using in chemico, high throughput and or in vitro methods. The goal of fully quantified AOPs is to allow the prediction of an adverse outcome in time and magnitude using a minimum number of experimental measurements for KE responses that cannot be adequately modeled by other means.

The third is AOPXplorer, a Cytoscape application, meant for building networks of KEs, forming AOP Networks (AOPNs) and allow data visualization of various types on top of the AOPNs. The goal of AOPXplorer is to help investigators and risk assessors understand

how chemical exposures result in information flow throughout the AOPN, allowing them to make defensible stories and inferences about potential adverse outcomes.

It has been postulated that omics technologies can be used for various goals in regulatory toxicology, such as biological read-across based on molecular events to prioritize chemicals for testing, cross-species extrapolation to link to evolutionary biology and the identification of KEs [10]. Although omics approaches have already been used in toxicology to define specific modes of action [11] or identifying biomarkers [12], they have not found their way into regulatory acceptance for assessment of chemicals and nanomaterials [13, 14]. There is a need for well-established experimental protocols for data generation, storage, processing, analysis, and interpretation to reach regulatory acceptance. Besides, an integration framework for data interpretation to identify relevant molecular changes and pathways is required, as well as the filling of knowledge gaps that keep risk assessors from causally linking molecular events to an adverse outcome at a higher level of biological organization [13, 15–18]. Taken together, the level of uncertainties and inconsistencies in experimental design should be minimized to allow omics approaches in risk assessment and AOPs. So far, various ideas have emerged to introduce omics data to the concept of the AOPs, such as a pipeline for KE enrichment [19], workflow for computationally predicted AOPs from public data [20] and the Transcriptomics Reporting Framework [21].

There is a demand for a consistent, well-defined protocol to analyze and integrate the data in order to describe the molecular effects downstream of an MIE [15]. Molecular pathway databases and tools exist to analyze omics datasets through pathway analysis, which happens through probability scoring of pathways containing differently expressed genes and thereby reducing the number of dimensions of omics datasets to the number of biological pathways. Various molecular pathway databases exist which could be viable tools for the integration of omics approaches in regulatory risk assessment, such as KEGG [22], Reactome [23] and WikiPathways [24].

In this paper, we describe how WikiPathways (wikipathways.org) [24] an open-science molecular pathway database which captures mechanistic knowledge in pathway diagrams, can be a supportive database for AOPs and the analysis and interpretation of omics datasets through pathway analysis. WikiPathways has similar levels of coverage of genes and metabolites as Reactome and KEGG [24, 25] and performs better in covering signaling pathways [26]. This can be done with PathVisio [27], a pathway diagram drawing tool that is connected to WikiPathways, in which omics data can be visualized and pathway analysis can be performed. Also, WikiPathways exists as a Cytoscape application, which allows the same pathways to be used for network analysis [28].

Thanks to the adaptability and accessibility of WikiPathways, communities can collaborate on creating, assessing and improving the understanding of molecular pathways [29]. Therefore, WikiPathways could be a valuable tool for the risk assessment community. It can provide improved molecular descriptions of early KEs which support biological plausibility. At the same time, it can serve as empirical support to KERs and allow the integration of omics technologies in the concept of AOPs in a systematic manner. As illustrated in Figure 3.1, ideally, all KEs in AOP-Wiki are linked by at least one molecular pathway, which can be highlighted by omics analysis and thereby revealing KEs. However, WikiPathways needs to be integrated with the existing modules in the AOP-KB. Here, we focus on the AOP-Wiki by describing its current implementation of semantic annotations and we will show how we can connect the AOP-Wiki with WikiPathways through identifiers for genes, proteins and metabolites, and ontologies [30], which are predefined vocabularies used to describe knowledge and assist in the integration of data sources. Furthermore, we will propose a strategy for future work on connecting the two databases, describing the planned work on WikiPathways and suggestions for improving the AOP-Wiki and its contents to allow linkage of databases.

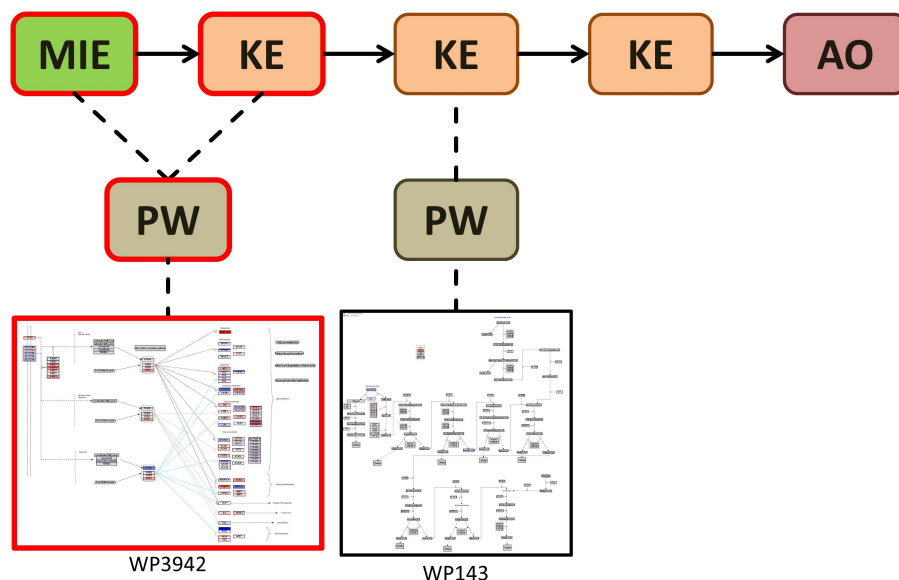


Figure 3.1: Illustrative description of the linkage of KEs of an AOP with molecular pathways described in WikiPathways and the practical application of transcriptomics. Transcriptomics and pathway enrichment analysis are commonly used to elucidate molecular pathways affected after exposure to a chemical or stress signal. In this illustration, gene expression levels in WP3942 [31] are significantly changed (red and blue nodes in the pathway diagram, for up- and downregulation). Because this pathway is linked to the MIE and first KE, these are hypothetically affected by the chemical, highlighted with red borders and require validation. WP143 [32] is not affected by the exposure of this chemical at the same time and dose, and the KE that is linked to this biological pathway is not considered to be affected but could follow later or at a higher dose. AO, adverse outcome; KE, key event; MIE, molecular initiating event; PW, pathway; WP, WikiPathways.

3.2 Materials and Methods

3.2.1 Retrieval of AOP-Wiki Data

The AOP-Wiki allows the use of their data for publication purposes, by storing permanent quarterly downloads on the website (aopwiki.org/downloads). For this paper, we used the AOP-Wiki XML file of April 1st, 2018, containing all AOP-Wiki content.

3.2.2 Parsing the AOP-Wiki XML

The AOP-Wiki XML was parsed with Python 3.5 [33] and the Element-Tree XML API with the “.parse”-function which resulted in an ElementTree wrapper class that represents an entire element hierarchy. The information, that was required for the experiments, was extracted included stressor information, ontology annotations, and information on KEs and KERs. The source code, as well as a brief tutorial on the execution of it, are available on GitHub [34].

3.2.3 BridgeDb Identifier Mapping in R

In order to perform identifier mapping for the chemicals that are stored on AOP-Wiki with CAS Registration Numbers (CAS numbers), we used the BridgeDb, an identifier mapping framework [35]. The CAS numbers from the AOP-Wiki were saved as plain text file and imported in RStudio (version 1.1.447; R version 3.4.4) [36, 37], in which the R-package BridgeDbR [38] was utilized to map the CAS numbers to ChEBI identifiers with the BridgeDb metabolite identifier mapping dataset [24]. The R code used for the identifier mapping is available on GitHub along with a tutorial to execute the script [34].

3.2.4 WikiPathways Data

Information from WikiPathways was retrieved using the WikiPathways SPARQL endpoint (`sparql.wikipathways.org`) [39], version 20180610. SPARQL is a query language to select specific subsets of data from a collection of RDF, a standard framework for knowledge descriptions. For this manuscript, various queries were performed to request information about WikiPathways’ use of ontologies and to retrieve pathways for lists of genes related to KEs.

3.2.5 Textual Identifier Mapping for Genes and Proteins

In order to perform identifier mapping on the free-text descriptions of AOP-Wiki, we downloaded a human gene identifier dataset from

the HUGO Gene Nomenclature Committee (HGNC) [40] in May 2018 via genenames.org, a curated online repository for HGNC-approved gene nomenclature, gene families and associated resources [41]. A custom download was performed in which we requested HGNC IDs, approved symbols, approved names, previous symbols, synonyms, and Ensembl IDs. These identifiers were loaded in Python and used to filter the descriptions of KEs for genes, which are filtered for KEs on the molecular, cellular, tissue, and organ level of biological organization. Also, the KERs that connect these KEs were parsed and identifiers were mapped on their descriptions and texts on biological plausibility and empirical support.

3.2.6 Manual Matching of AOP-Wiki KEs to Molecular Pathways on WikiPathways

All AOP-Wiki KE IDs on the molecular, cellular, tissue, and organ level were extracted and their corresponding web pages were opened on aopwiki.org. From the KE titles and descriptive text, pathway names were selected and queried on wikipathways.org via the search-bar for molecular pathways. If results showed up for this initial search, the KE was considered present in WikiPathways. If the KEs did not contain a direct mention of a pathway, the genes and proteins were noted and were queried for their presence in pathways via the WikiPathways SPARQL endpoint. For KEs at the cellular level, at least the majority of the genes and proteins should be present in at least one pathway. However, for molecular KEs that describe only an interaction between two molecules, only the presence of the target molecule in WikiPathways was necessary to consider the KE covered by WikiPathways. This method was meant to give a rough overview of the overlap between the AOP-Wiki and WikiPathways databases. Because it does not include synonyms or ontological similarity, this overview is expected to underestimate the overlap.

3.3 Results

For hard linkage of the two databases, meaning explicit identifier matching, we looked at the usage of ontology annotations of the AOP-Wiki and WikiPathways. For the AOP-Wiki we extracted ontology annotations from KEs on the molecular, cellular, tissue and organ level and identified which ontology sources were currently in use for biological processes, biological objects, cell-terms, and organ-terms. As shown in Figure 3.2, a large amount of KEs are not yet annotated with ontology tags. When looking more in detail, one can notice that biological processes are mostly described with Gene Ontology (GO) tags, especially at the molecular and cellular KEs whereas the biological objects are mostly annotated with tags from ChEBI and Protein Ontology (PR). Although AOP-Wiki contains various ontology sources, WikiPathways only uses three: Pathway Ontology (PW), Cell Ontology (CL), and the Disease Ontology (DO) (Figure 3.3). However, apart from the CL for a contextual description of the process, WikiPathways and AOP-Wiki do not share ontologies for other biological elements.

Although no direct mappings through ontologies are possible at the moment of writing this paper, an alternative approach for hard linkage is the mapping of chemicals, metabolites, and genes to WikiPathways. Although we do not expect to find many of the AOP-Wiki stressor chemicals in WikiPathways, we wanted to identify the existing overlap of chemicals between the two databases nevertheless. First, we found all 306 stressors, describing 207 chemicals, which were annotated with 205 CAS numbers. We mapped these CAS numbers to ChEBI IDs in R with BridgeDbR and created a SPARQL query to find all pathways that have any of the metabolites included. This resulted in a total 194 out of 205 CAS numbers mapped to 298 ChEBI IDs, of which 48 mapped to a total of 133 WikiPathways.

As opposed to the hard linkage of the two databases, we also investigated a soft linkage, which entails the indirect linking of these databases through a text-based identifier mapping approach

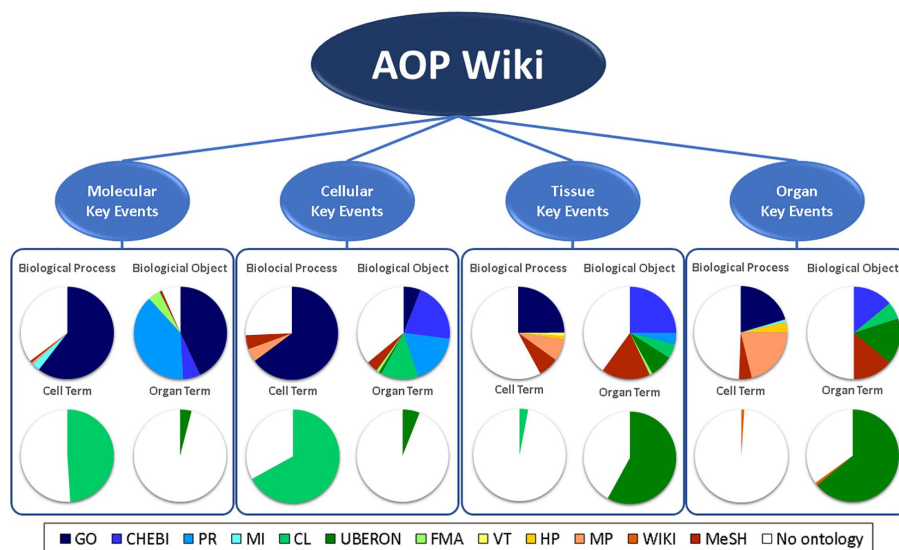


Figure 3.2: Ontology usage of AOP-Wiki for KEs on the molecular, cellular, tissue, and organ level of biological organization. GO, gene Ontology; CHEBI, chemical entities of biological interest; PRO, protein ontology; MI, molecular interactions; CL, cell ontology; UBERON, uber anatomy ontology; FMA, foundational model of anatomy; VT, vertebrate trait; HP, human phenotype ontology; MP, mammalian phenotype ontology; WIKI, AOP-Wiki; MeSH, medical subject headings.

of human genes and performed a similar SPARQL query as for the metabolites (Figure 3.4). After extracting all KE descriptions from the AOP-Wiki, we mapped gene identifiers, symbols, alternative names, and previous names from HGNC to each description, leading to the identification of 523 genes in a total of 234 KE descriptions out of 787 KEs. In total, 70% of these genes were found in the molecular pathways of WikiPathways. Also, identifier mapping was performed on all 874 KERs that connect the KEs on the molecular, cellular, tissue and organ level. This was done on all texts for KER descriptions, biological plausibility, and empirical support, when available, and resulted in the identification of 417 genes, of which 296 are present in pathways on WikiPathways, which is 71%.

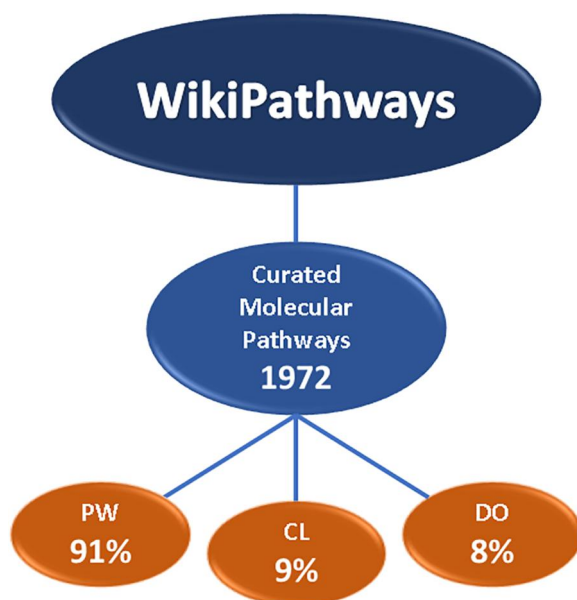


Figure 3.3: WikiPathways statistics. The total number of molecular pathways in the WikiPathways database, and the level of pathway annotations with ontology tags. PW, pathway ontology; CL, cell ontology; DO, disease ontology.

Furthermore, to benchmark the hard and soft connections between the AOP-Wiki and WikiPathways through ontologies and identifiers, we performed a full-scale manual check for all KEs on the molecular, cellular, tissue, and organ level of biological organization. This showed us that at least 2/3rd of all KEs can be mapped to molecular pathways on WikiPathways.

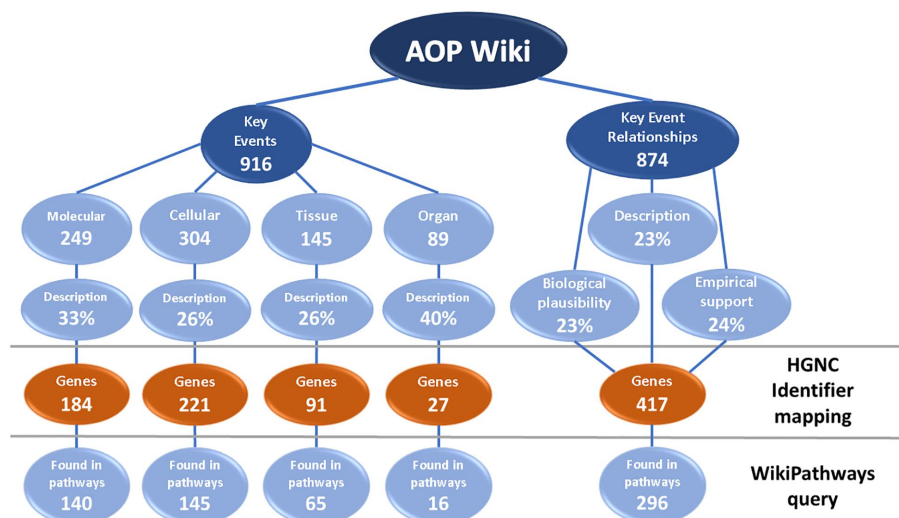


Figure 3.4: AOP-Wiki statistics on KEs and KERs, identifier mapping with HGNC identifiers and links to molecular pathways in WikiPathways. The KEs on the molecular, cellular, tissue, and organ level of biological organization and the KERs that connect them were parsed for texts of descriptions, on the biological plausibility and on the empirical support. HGNC Identifier mapping was performed to find all human genes described in the key event descriptions, after which these genes were queried on WikiPathways to find pathways that contain these genes.

3.4 Discussion

In this paper we explored possibilities for the integration of WikiPathways in the AOP-KB through ontologies, identifiers and manual judgment, to support AOPs and become a valuable tool in regulatory risk assessment. We looked at hard and soft linkages between the AOP-Wiki, the most actively used AOP module of the AOP-KB, and WikiPathways. We did this by extracting different types of information from the AOP-Wiki, such as chemical CAS numbers, KE and KER descriptions, and ontology annotations, and we performed a manual judgment of the linkage.

We found that the AOP-Wiki uses various ontologies to describe the different elements of KEs. To link the underlying molecular pathways

to these KEs, we are mainly interested in the Biological Process that is annotated in the KEs, which describe the biology of the KEs. However, the ontologies currently used in the AOP-Wiki do not directly connect with the ontologies that describe the molecular pathways of WikiPathways. Consequently, manual effort is currently required to make this mapping, which negatively impacts the scalability.

Furthermore, we focused on the metabolites and genes/proteins described on the AOP-Wiki. For the metabolites, we parsed all CAS numbers, mapped these to ChEBI identifiers, and found that only 16% of these are found in WikiPathways. This is not unexpected, because most toxicological effects are caused by exogenous compounds, whereas WikiPathways mostly stores biological pathways containing endogenous metabolites. In fact, most WikiPathways that contain such a stressor do so because the pathway described the biotransformation of the toxic compound.

On the other hand, gene/protein identifiers that we obtained through mapping with an HGNC dataset did show high coverage by WikiPathways (70%). However, with the gene/protein identifier mapping, we only focused on human variants, although KE descriptions on the AOP-Wiki cover a variety of species. The taxonomic information is absent in most KEs and if it is available, the taxonomy identifiers are inconsistent, so we were not able to take this into account in our experiment of identifier mapping. Although species specification with ontologies does exist on the AOP-Wiki, the number of annotations and the consistency in reporting should increase for it to become a useful piece of data.

Apart from the automated linkages, we performed a manual check, which indicated that the majority of the processes in the AOP-Wiki KEs are covered by the WikiPathways database, either completely, as a part of a pathway or, in case of molecular interactions, the target molecule is part of a molecular pathway. This indicates us that there is potential in the interoperability of AOP-Wiki and WikiPathways to describe KEs. However, there is no one-to-one mapping of biological

pathways possible. For example, molecular-level KEs currently often describe a single interaction between a list of stressors and a molecule, which would only be a part of a biological pathway on WikiPathways, besides the downstream cascade of molecular effects. Also, KEs on the tissue- and organ-level of biological organization are often non-specific. This could lead to the mapping of multiple molecular pathways to a single AOP-Wiki KE, even with the current WikiPathways content.

Besides the identification of connections between the AOP-Wiki and WikiPathways for improved descriptions of KEs, we aim for the possibility to introduce omics data analysis in the concept of AOPs. However, one concern mentioned in literature in the implementation of transcriptomics data in the concept of AOPs is the difference in the causal and reactive pathways [42]. Transcriptomics studies, for example, do not differentiate in its measurements between these two types of pathways, and by focusing on gene expression fold changes, pathway enrichment may highlight the reactive pathways. However, KEs may describe a causal event or pathway. Therefore, AOP-Wiki KE descriptions would not necessarily overlap with the results from pathway analysis with omics data. This should be taken into account in the descriptions of the molecular responses of KEs as this might impact the usability of omics approaches and their connections to KEs on the AOP-Wiki.

It is expected that omics approaches have great potential in the field of regulatory toxicology [13, 15]. However, there is a demand for well-described protocols and tools for omics data analysis and interpretation. The integration of WikiPathways in the AOP-KB as a data source and as omics data analysis tool allows more detailed descriptions of KEs and consistency in analysis and interpretation of omics data in the concept of AOPs. For that, you would ideally have molecular mechanistic descriptions for all AOP events in WikiPathways. The current analysis shows that useful connections already exist. To prepare for the integration of molecular pathways in the concept of AOPs, we created an AOP Portal on WikiPathways (aop.wikipathways.org), in

which all molecular pathways that are linked to AOP-Wiki KEs will be gathered and stored. This portal is meant to bridge the molecular knowledge and expertise of biologists and toxicologists to the framework of AOPs and allows the whole community to contribute to the collection of molecular pathways. This collection will be available for pathway analysis and network analysis with omics data for large-scale hypothesis generation for AOPs in response to a stressor or for biological read-across on the AOP level [15]. That would allow a more consistent, standardized approach for the integration of omics approaches in AOPs, and thus for regulatory use.

A variety of molecular pathway databases could fill this role as an omics analysis and interpretation tool for toxicological effects, such as KEGG and Reactome. However, molecular pathways can vary across pathway databases due to differences in pathway annotations by focusing on specific cellular contexts, such as diseases or specific cell types [43]. Moreover, Reactome and KEGG cannot be tailored like WikiPathways for specific communities or purposes such as described in this paper [29, 44]. Besides, the accessibility of WikiPathways, being a community-driven, free-to-use molecular pathway database, fits with the existing AOP-KB modules and meets the requirements identified by the OECD: open access, standardized representation of data, and consistency in reporting [8, 45]. Because the AOP-KB is driven by a scientific community to develop, share and discuss AOPs, this community can also describe the molecular processes underlying the AOPs and contribute to WikiPathways and expand the AOP Portal.

Other work on the linkage of data related to the AOP-Wiki is the development of the AOP-DataBase (AOP-DB) [46]. This database will soon be publicly available and will contain various types of information linked to gene IDs that is useful for AOPs to provide a standardized, systematic structure for AOP development. Among a large amount of data, biological pathways from databases such as KEGG, Reactome, and ConsensusDB are included based on GO annotations of KEs in AOP-Wiki [46]. While the AOP-DB connects pathway databases based on the ontology annotations to of existing AOPs and assisting the iden-

tification of putative AOPs, we think that a direct link between KEs and molecular pathways would be valuable and more reliable.

In order to make a connection between AOP-Wiki and WikiPathways, we recommend a couple of improvements in terms of annotations and accessibility of the data. Since January 2018, the AOP-Wiki made available full XML files containing all data, which are stored as permanent downloads, as well as nightly exports of the full database. These files need to be parsed to retrieve the data, as described in this paper. This could be improved by developing an RDF version of the AOP-Wiki, allowing federated SPARQL queries to request all data, enable automatic information sharing, and has the use of ontologies as a core feature.

Furthermore, the current implementation of annotations with ontologies could be improved by annotating more specific elements of the KEs, as the existing KE components describe the KEs in general. More detailed annotations could be performed for many elements. For example, key genes, proteins, and metabolites should be annotated, as well as detection methods and biological assays, which can be annotated with ontologies such as the Chemical Methods Ontology or BioAssay Ontology. Also, when biological pathways are described in a KE, annotations with the Pathway Ontology would allow a direct connection to the WikiPathways database including all genes, proteins, and metabolites involved, which are annotated with various databases through BridgeDb in the WikiPathways diagrams.

Besides the ontology annotations, the only molecules annotated on the AOP-Wiki are the chemicals related to stressors, which are identified with CAS numbers. However, not all of these CAS numbers are linked to open structure data that is incorporated in the BridgeDb mapping that we performed. It is essential that these CAS numbers are included in public databases, such as WikiData [47] or that public database identifiers are used, such as from ChEBI or even Wikidata as an outside database for chemical information. Besides chemicals, nanomaterials, which are extensively investigated for toxicity, also require annotations, for example with the eNanoMapper ontology [48].

Also, the free-text descriptions of KEs that describe the biological process can also be improved by more consistent reporting, such as a fixed vocabulary for all genes, proteins, and metabolites involved in the biological processes. For example, listing the most important molecules by HGNC symbols or ChEBI IDs for a KE would improve machine-readability and the automated discovery of new connections between KEs.

On the other hand, WikiPathways will also need to undergo updates to fit the connection as described, with a specific category of KE-related molecular pathways and the need for so-called meta-pathways to create an AOP Network. Also, the AOP Portal will be populated with pathways in a case-study approach, proving the usefulness of the database. Other improvements related to toxicity research is the linkage to kinetics databases, more info on post-translational modifications of proteins, and improved semantic annotations of localizations, for example, specific organelles, cells, or tissues.

Taken together, we claim that a tight integration of WikiPathways and AOP-KB will improve risk assessment because we can link omics data directly to KEs and therefore AOPs. However, to make assessment reproducible and valid, major changes are needed.

References

- [1] Gerald T. Ankley et al. "Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment". *Environmental Toxicology and Chemistry* 29.3 (Mar. 2010), pp. 730–741. DOI: 10.1002/etc.34.
- [2] Daniel L. Villeneuve et al. "Adverse Outcome Pathway Development II: Best Practices". *Toxicological Sciences* 142.2 (Dec. 2014), pp. 321–330. DOI: 10.1093/TOXSCI/KFU200.
- [3] Marcel Leist et al. "Adverse outcome pathways: opportunities, limitations and open questions". *Archives of Toxicology* 91.11 (Nov. 2017), pp. 3477–3505. DOI: 10.1007/s00204-017-2045-3.
- [4] Mathieu Vinken et al. "Adverse outcome pathways: a concise introduction for toxicologists". *Archives of Toxicology* 91.11 (Nov. 2017), pp. 3697–3707. DOI: 10.1007/s00204-017-2020-z.

- [5] Mathieu Vinken. "The adverse outcome pathway concept: A pragmatic tool in toxicology". *Toxicology* 312.1 (Oct. 2013), pp. 158–165. DOI: 10.1016/j.tox.2013.08.011.
- [6] OECD. "Revised Guidance Document on Developing and Assessing Adverse Outcome Pathways". *OECD Series on Testing and Assessment* No. 184. (2017).
- [7] Lyle D. Burgoon. "The AOPontology: A semantic artificial intelligence tool for predictive toxicology". *Applied In Vitro Toxicology* 3.3 (Sept. 2017), pp. 278–281. DOI: 10.1089/aivt.2017.0012.
- [8] Cataia Ives et al. "Creating a Structured AOP Knowledgebase via Ontology-Based Annotations." *Applied in vitro toxicology* 3.4 (Dec. 2017), pp. 298–311. DOI: 10.1089/aivt.2017.0017.
- [9] Karen H. Watanabe-Sailor et al. "Big Data Integration and Inference". *Issues in Toxicology* 2020-January.41 (Dec. 2019), pp. 264–306. DOI: 10.1039/9781782623656-00264.
- [10] Thomas Hartung. "Making big sense from big data in toxicology by read-across". *Altex* 33.2 (2016), pp. 83–93. DOI: 10.14573/ALTEX.1603091.
- [11] Stephen W. Edwards and R. Julian Preston. "Systems biology and mode of action based risk assessment". *Toxicological Sciences* 106.2 (2008), pp. 312–318. DOI: 10.1093/TOXSCI/KFN190.
- [12] Roland C. Grafström et al. "Toward the Replacement of Animal Experiments through the Bioinformatics-driven Analysis of 'Omics' Data from Human Cell Cultures". *Alternatives to laboratory animals : ATLA* 43.5 (Nov. 2015), pp. 325–332. DOI: 10.1177/026119291504300506.
- [13] Roland Buesen et al. "Applying 'omics technologies in chemicals risk assessment: Report of an ECETOC workshop". *Regulatory Toxicology and Pharmacology*. Vol. 91. Academic Press, Dec. 2017, S3–S13. DOI: 10.1016/j.yrtph.2017.09.002.
- [14] Bennard van Ravenzwaay, Ursula G. Sauer, and Olivier de Matos. "Editorial: Applying 'omics technologies in chemicals risk assessment". *Regulatory Toxicology and Pharmacology* 91 (Dec. 2017), S1–S2. DOI: 10.1016/J.YRTPH.2017.11.017.
- [15] Erica K. Brockmeier et al. "The Role of Omics in the Application of Adverse Outcome Pathways for Chemical Risk Assessment". *Toxicological Sciences* 158.2 (Aug. 2017), pp. 252–262. DOI: 10.1093/toxsci/kfx097.
- [16] Ursula G. Sauer et al. "The challenge of the application of 'omics technologies in chemicals risk assessment: Background and outlook". *Regulatory Toxicology and Pharmacology* 91 (Dec. 2017), S14–S26. DOI: 10.1016/j.yrtph.2017.09.020.
- [17] Julien Vachon et al. "Barriers to the use of toxicogenomics data in human health risk assessment: A survey of Canadian risk assessors". *Regulatory Toxicology and Pharmacology* 85 (Apr. 2017), pp. 119–123. DOI: 10.1016/J.YRTPH.2017.01.008.
- [18] Bruno Campos and John K. Colbourne. "How omics technologies can enhance chemical safety regulation: perspectives from academia, government, and industry: The Perspectives column is a regular series designed to discuss and evaluate potentially competing viewpoints and research findings on current environmental issues". *Environmental Toxicology and Chemistry* 37.5 (May 2018), pp. 1252–1259. DOI: 10.1002/ETC.4079.

-
- [19] Penny Nymark et al. "A Data Fusion Pipeline for Generating and Enriching Adverse Outcome Pathway Descriptions". *Toxicological Sciences* 162.1 (Mar. 2018), pp. 264–275. DOI: 10.1093/toxsci/kfx252.
- [20] Shannon M. Bell et al. "Integrating publicly available data to generate computationally predicted adverse outcome pathways for fatty liver". *Toxicological Sciences* 150.2 (Apr. 2016), pp. 510–520. DOI: 10.1093/toxsci/kfw017.
- [21] Timothy W. Gant et al. "A generic Transcriptomics Reporting Framework (TRF) for 'omics data processing and analysis". *Regulatory Toxicology and Pharmacology* 91 (Dec. 2017), S36–S45. DOI: 10.1016/J.YRTPH.2017.11.001.
- [22] M. Kanehisa. "KEGG: Kyoto Encyclopedia of Genes and Genomes". *Nucleic Acids Research* 28.1 (Jan. 2000), pp. 27–30. DOI: 10.1093/nar/28.1.27.
- [23] Antonio Fabregat et al. "The Reactome Pathway Knowledgebase". *Nucleic Acids Research* 46.D1 (Jan. 2018), pp. D649–D655. DOI: 10.1093/nar/gkx1132.
- [24] Denise N. Slenter et al. "WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research". *Nucleic Acids Research* 46.D1 (Nov. 2018), pp. D661–D667. DOI: 10.1093/nar/gkx1064.
- [25] Martina Kutmon et al. "WikiPathways: Capturing the full diversity of pathway knowledge". *Nucleic Acids Research* 44.D1 (2016), pp. D488–D494. DOI: 10.1093/nar/gkv1024.
- [26] A. K.M. Azad, Alfons Lawen, and Jonathan M. Keith. "Bayesian model of signal rewiring reveals mechanisms of gene dysregulation in acquired drug resistance in breast cancer". *PLoS ONE* 12.3 (Mar. 2017). DOI: 10.1371/JOURNAL.PONE.0173331.
- [27] Martina Kutmon et al. "PathVisio 3: An Extendable Pathway Analysis Toolbox". *PLOS Computational Biology* 11.2 (Feb. 2015). Ed. by Robert F. Murphy, e1004085. DOI: 10.1371/journal.pcbi.1004085.
- [28] Martina Kutmon et al. "WikiPathways App for Cytoscape: Making biological pathways amenable to network analysis and visualization". *F1000Research* 3 (Sept. 2014), p. 152. DOI: 10.12688/f1000research.4254.2.
- [29] Alexander R. Pico et al. "WikiPathways: Pathway Editing for the People". *PLoS Biology* 6.7 (July 2008), e184. DOI: 10.1371/journal.pbio.0060184.
- [30] Jonathan B.L. Bard and Seung Y. Rhee. "Ontologies in biology: Design, applications and future challenges". *Nature Reviews Genetics* 5.3 (Mar. 2004), pp. 213–222. DOI: 10.1038/NRG1295.
- [31] Michiel Adriaens et al. *Fatty acid beta-oxidation (Homo sapiens) - WikiPathways*. wikipathways.org/instance/WP143_r98914. 2018.
- [32] Kristina Hanspers and Denise N Slenter. *PPAR signaling pathway (Homo sapiens)*. 2017.
- [33] Python Software Foundation. *Python Language Reference, version 3.5*. 2010.
- [34] Marvin Martens. "marvinm2/AOPWikiXMLparsing: Version 1.0" (July 2018). DOI: 10.5281/ZENODO.1306408.
- [35] Martijn P van Iersel et al. "The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services". *BMC Bioinformatics* 11.1 (2010), p. 5. DOI: 10.1186/1471-2105-11-5.
- [36] RStudio Team. *RStudio: integrated development for R*. 2015.
- [37] R Core Team. *R: A language and environment for statistical computing*. 2013.
- [38] C Leemans et al. *BridgeDbR: Code for Using BridgeDb Identifier Mapping Framework From Within R*. 2018.

- [39] Andra Waagmeester et al. "Using the Semantic Web for Rapid Integration of WikiPathways with Other Biological Online Data Resources". *PLOS Computational Biology* 12.6 (June 2016). Ed. by Christos A. Ouzounis, e1004989. DOI: 10.1371/journal.pcbi.1004989.
- [40] HGNC Database et al. *HUGO Gene Nomenclature Committee (HGNC)*. 2018.
- [41] Bethan Yates et al. "Genenames.org: The HGNC and VGNC resources in 2017". *Nucleic Acids Research* 45.D1 (Jan. 2017), pp. D619–D625. DOI: 10.1093/NAR/GKW1033.
- [42] Shana J. Sturla et al. "Systems toxicology: From basic research to risk assessment". *Chemical Research in Toxicology* 27.3 (Mar. 2014), pp. 314–329. DOI: 10.1021/TX400410S.
- [43] Ralf Herwig et al. "Analyzing and interpreting genome data at the network level with ConsensusPathDB". *Nature Protocols* 11.10 (Oct. 2016), pp. 1889–1907. DOI: 10.1038/nprot.2016.117.
- [44] Mamatha Hanumappa et al. "WikiPathways for plants: A community pathway curation portal and a case study in rice and arabidopsis seed development networks". *Rice* 6.1 (2013), pp. 1–10. DOI: 10.1186/1939-8433-6-14.
- [45] Dirk Pilat and Yukiko Fukasaku. "OECD Principles and Guidelines for Access to Research Data from Public Funding". *Data Science Journal* 6 (2007), OD4–OD11. DOI: 10.2481/DSJ.6.OD4.
- [46] Maureen E. Pittman et al. "AOP-DB: A database resource for the exploration of Adverse Outcome Pathways through integrated association networks". *Toxicology and Applied Pharmacology* 343 (Mar. 2018), pp. 71–83. DOI: 10.1016/j.taap.2018.02.006.
- [47] Daniel Mitchen et al. "Enabling Open Science: Wikidata for Research (Wiki4R)". *Research Ideas and Outcomes* 1 (Dec. 2015), e7573. DOI: 10.3897/RIO.1.E7573.
- [48] Janna Hastings et al. "eNanoMapper: Harnessing ontologies to enable data integration for nanomaterial risk assessment". *Journal of Biomedical Semantics* 6.1 (Mar. 2015), p. 10. DOI: 10.1186/s13326-015-0005-5.

4

Providing adverse outcome pathways from the AOP-Wiki in a semantic web format to increase usability and accessibility of the content.

Adapted from: Marvin Martens, Chris T. Evelo, and Egon L. Wil-
lighagen. "Providing Adverse Outcome Pathways from the AOP-Wiki
in a Semantic Web Format to Increase Usability and Accessibility of
the Content". *Applied in vitro toxicology* 8.1 (Mar. 2022), pp. 2–13. DOI:
10.1089/AIVT.2021.0010.

Abstract

The AOP-Wiki is the main platform for the development and storage of Adverse Outcome Pathways. These Adverse Outcome Pathways describe mechanistic information about toxicodynamic processes and can be used to develop effective risk assessment strategies. However, it is challenging to automatically and systematically parse, filter, and use its contents. We explored solutions to better structure the AOP-Wiki content and to link it with chemical and biological resources. Together this allows more detailed exploration which can be automated.

We converted the complete AOP-Wiki content into Resource Description Framework (RDF) as triples. We used over twenty ontologies for the semantic annotation of property-object relations, including the Chemical Information Ontology, Dublin Core, and the Adverse Outcome Pathway Ontology. The latter was used over 8,000 times. Furthermore, over 3,500 link-outs were added to twelve chemical databases and over 7,500 link-outs to four gene and protein databases.

The AOP-Wiki RDF has been made available at aopwiki.rdf.bigcat-bioinformatics.org where SPARQL queries can be used to answer biological and toxicological questions, such as listing measurement methods for all Key Events leading to an Adverse Outcome of interest. The full power that the use of this new resource provides becomes apparent when combining the content with external databases using federated queries. Overall, the AOP-Wiki RDF allows new ways to explore the rapidly growing Adverse Outcome Pathway knowledge and makes the integration of this database in automated workflows possible.

4.1 Introduction

Since its establishment in 2010, the Adverse Outcome Pathway (AOP) concept has become a prominent tool for the risk assessment community [1, 2]. AOPs are a chain of biological processes, called Key Events (KEs), starting from a molecular perturbation with a stressor towards an Adverse Outcome (AO), connected by Key Event Relationships (KERs). AOPs exist to capture all mechanistic toxicological knowledge from literature and data, to direct future studies to fill gaps of existing knowledge, and to drive Integrated Approaches to Testing and Assessment (IATA) development [1, 3]. This was demonstrated with the AOP-based IATA for skin sensitization, resulting in various IATA with combinations of *in vitro* and *in silico* assays outperforming animal tests [4].

The majority of the AOPs are developed and stored in the AOP-Wiki (aopwiki.org), which is part of the AOP Knowledge Base, released in 2014 as a result of the AOP development program initiated by the Organisation for Economic and Collaborative Development (OECD) [5]. This wiki is designed to facilitate collaborative development of qualitative AOP descriptions, and thereby promote their incorporation into risk assessments and stimulate effective reuse of mechanistic toxicological knowledge [6, 7].

The resulting AOPs describe much of the biological context surrounding toxicological processes, most of the information on genes, chemicals, biological pathways, and phenotypes, among other things, are already captured in specialised databases or ontologies outside of AOP-Wiki [8]. However, the AOP-Wiki has limited possibilities for linking of external information and data, mostly consisting of free-text descriptions and links to the US CompTox Chemistry Dashboard [9] and to NCBI for taxonomic applicability [10]. An initiative to make the reporting more consistent was the introduction of Key Event Components [11] for the annotation of Biological Processes, Biological Objects and Biological Actions for KEs, and annotations of cell types and organs in which KEs can occur.

Since the AOP-Wiki is the central repository for AOPs and therefore a key player in the shift towards animal-free testing strategies, it is essential that its contents can be queried and utilized effectively to answer biological questions and to reuse existing knowledge. However, accessing the data computationally or linking with other resources is hardly possible when only downloadable eXtensible Markup Language (XML) data dumps are provided that consist mostly of free text. Because of these aspects, parsing and querying the continuously growing amount of information in the AOP-Wiki is a complex, time-consuming task. This is a problem because it prevents the integration of AOP knowledge with other data and resources.

This could be resolved by applying Linked Open Data solutions, such as structuring the data in a Resource Description Framework (RDF) model [12], introducing persistent identifiers and semantic annotations, and implementing Application Programming Interfaces (APIs) for accessing the data. RDF represents knowledge as semantic triples, in which a subject, predicate and object, together define a statement and assist in the meaningful representation of knowledge in a machine-readable manner.

These concepts are generally in line with the FAIR principles [13] for data and knowledge management, developed to enhance the Findability, Accessibility, Interoperability, and Reusability of data and allow computational support of data usage. For example, such as the solutions applied by the Swiss Institute of Bioinformatics with the development of neXtProt Linked Data by implementing RDF annotations for easier exploration and retrieval of data through web services [14, 15].

Also, the use of ontologies and vocabularies for semantic annotations allows for the integration of data between resources, such as the direct linking of chemical or protein databases with WikiPathways [16, 17].

In this paper we show how using RDF makes the AOP-Wiki content more usable for automated exploration in combination with other existing semantic web based information sources. We describe our implementation of Linked Open Data solutions for the AOP-Wiki to in-

roduce new, effective ways of accessing and using the data. These solutions will enhance the usefulness of the AOP-Wiki to risk assessors, developers, and modelers, and facilitate answering complex research questions, also across databases or as part of automated workflows. We hypothesise that with the implementation of RDF, with the use of standard ontologies for semantic modelling of information captured in AOPs, the data can be better exploited [18]. Furthermore, the domain-specific AOP Ontology (AOPO), in combination with other relevant ontologies, can be used to link various pieces of mechanistic toxicological information and thereby facilitate knowledge-based hazard identification using AOPs [19]. The use of persistent, unique and resolvable identifiers allows the interoperability with other related data sources. When combined with computational tools that can access experimental data these approaches can make AOP information a core element for predictive modelling [20].

4.2 Methods

4.2.1 Registering AOP-Wiki identifiers in Identifiers.org

Prior to the development of the AOP-Wiki RDF, we registered the identifiers for the AOP, KE, KER, and stressor in the Minimum Information Required In the Annotation of Models (MIRIAM) Registry [21] to allow Identifiers.org to resolve Internationalized Resource Identifiers (IRIs). In order to make all identifiers in the AOP-Wiki resolvable and linking to their corresponding database webpages, these IRIs, along with a variety of chemical and gene database identifier types, were implemented in the AOP-Wiki RDF.

4.2.2 XML-to-RDF conversion code

The code for the XML-to-RDF conversion was written as a Jupyter notebook using Python version 3.7.3 in JupyterLab version 0.35.5, and is stored in GitHub (github.com/marvinm2/AOPWikiRDF) [22].

Downloading and parsing the AOP-Wiki XML

It downloads the AOP-Wiki XML quarterly download file of January 1st, 2021 from aopwiki.org/downloads and parses the file with the ElementTree XML API Python library. Next, the Jupyter notebook stores all the AOP-Wiki content in a Python nested dictionary data model, one for each of the main components which form the basis of the existing AOP-Wiki. These are the AOPs, KEs, KERs, stressors, chemicals, taxonomy, cell-terms, organ-terms, and the KE Components, which comprise of Biological Processes, Biological Objects and Biological Actions.

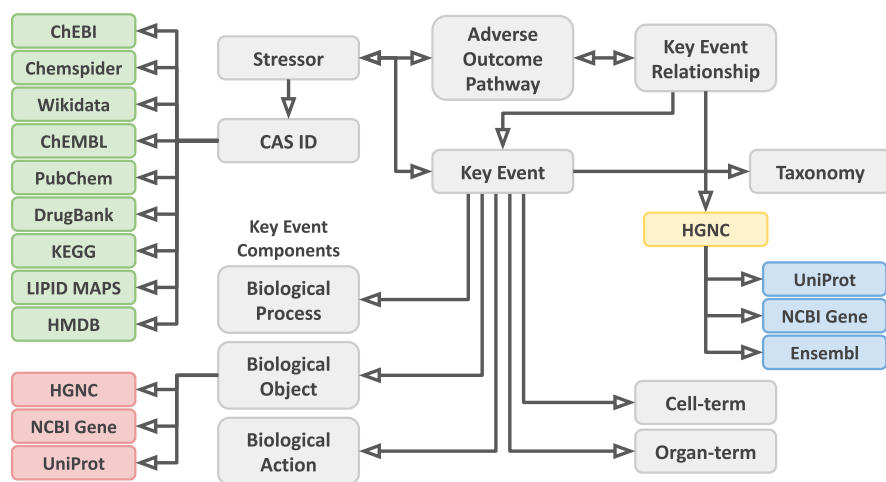


Figure 4.1: General overview of the AOP-Wiki RDF scheme. Arrows show the directional relationships described in the RDF. Grey boxes are the basic elements of the AOP-Wiki. Green boxes indicate added chemical IDs using BridgeDb. Red boxes indicate added gene/protein IDs using Protein Ontology mapping. The yellow box indicates the text-mapped gene IDs and the blue boxes indicate the added gene/protein IDs mapped from the text-mapped gene IDs using BridgeDb.

Semantic annotation in the RDF

Terms from common biomedical terminologies and standard meta-data vocabularies were used as predicates. These terms were retrieved from BioPortal [23] or in the corresponding Web Ontology Language (OWL) [24] files stored in GitHub. These ontologies include Dublin Core [25], DCMi Metadata Terms [26], RDF Schema [27], Friend Of A Friend [28], Adverse Outcome Pathway Ontology [19], Phenotypic Quality Ontology [29], Chemical Information ontology [30], NCI Thesaurus [31], Measurement Method Ontology [32], Simple Knowledge Organization System [33], National Center for Biotechnology Information Organismal Classification [10], Gene Ontology [34], EDAM bioinformatics operations, types of data, data formats, identifiers, and topics [35], Provenance, Authoring and Versioning [36], Vocabulary of Interlinked Datasets [37], Data Catalog Vocabulary [38]. Table S4.1 provides an overview of these, including their prefixes and IRI patterns. Furthermore, the IRIs were completed for annotations that already exist in the AOP-Wiki such as the KE Components, cell-terms and organ-terms. These annotations include terms of the Cell Ontology [39], Uber-anatomy ontology [40], Gene Ontology [34], Molecular Interactions Controlled Vocabulary [41], Mammalian Phenotype Ontology [42], Medical Subject Headings [43], Human Phenotype Ontology [44], Population and Community Ontology [45, 46], Neuro Behavior Ontology [47], Vertebrate trait ontology [48], PRotein Ontology [49], Chemical Entities of Biological Interest [50], and Foundational Model of Anatomy Ontology [51]. These ontologies are listed in Table S4.2 (Annex) together with their prefixes and IRI patterns.

Addition of gene and protein identifiers

In order to increase the number of annotations and add more types of gene and protein identifiers for improved linking of data and repositories, the XML-to-RDF conversion includes two methods of mapping to gene and protein identifiers.

Table 4.1: Ontologies and vocabularies used in the RDF.

Ontology name	Prefix in RDF	Base IRI
Dublin Core [25]	dc	http://purl.org/dc/elements/1.1/
DCMI Metadata Terms [26]	dcterms	http://purl.org/dc/terms/
RDF Schema [27]	rdfs	http://www.w3.org/2000/01/rdf-schema#
Friend Of A Friend [28]	foaf	http://xmlns.com/foaf/0.1/
Adverse Outcome Pathway Ontology [19]	aopo	http://aopkb.org/aop_ontology#
Phenotypic Quality Ontology [29]	pato	http://purl.obolibrary.org/obo/PATO_
Chemical Information ontology [30]	cheminf	http://semanticscience.org/resource/CHEMINF
NCI Thesaurus [31]	nci	http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#
Measurement Method Ontology [32]	mmo	http://purl.obolibrary.org/obo/MMO_
Simple Knowledge Organization System [33]	skos	http://www.w3.org/2004/02/skos/core#
National Center for Biotechnology Information Organismal Classification [10]	ncbitaxon	http://purl.bioontology.org/ontology/NCBITAXON/
Gene Ontology [34]	go	http://purl.obolibrary.org/obo/GO_
EDAM bioinformatics operations, types of data, data formats, identifiers, and topics [35]	edam	http://edamontology.org/
Provenance, Authoring and Versioning [36]	pav	http://purl.org/pav/
Vocabulary of Interlinked Datasets [37]	void	http://rdfs.org/ns/void#
Data Catalog Vocabulary [38]	dcat	http://www.w3.org/ns/dcat#

The first of which is based on existing Biological Object annotations with PProtein ontology (PR) terms [49] in the AOP-Wiki, which were mapped to identifiers from NCBI Gene [52] and UniProt [53], and symbols from the HUGO Gene Nomenclature Committee (HGNC) [54] with the PR mapping file, promapping.txt, downloaded from proconsortium.org/download/current on May 10th, 2020.

The second method involved textual gene identifier mapping for KEs and KERs, for which we extracted approved symbols, names, and synonyms for all human genes from the HGNC (downloaded from genenames.org [54] in January 2020). After loading in the HGNC file, a symbol dictionary was created which included all official gene symbols, names and alternative gene names, and the library has been extended with variants of textual separators surrounding the symbols to avoid partial word overlaps. Next, text matching was done for each KE and KER description, the MIE- and AO-specific section of KEs, and biological plausibility and empirical support sections of KERs. For each perfect match, the matching HGNC identifier was added to the KE and KER dictionaries to store KE-gene information.

BridgeDb identifier mapping

On top of the chemicals already present in the AOP-Wiki and the genes and proteins IDs added to the RDF with the textual mapping of HGNC symbols, we extended the coverage of external molecular databases using BridgeDb, an identifier mapping service for chemicals, genes, proteins, and interactions [55]. The “requests” Python library (version 2.22.0) was used for calling BridgeDb’s “xref” function to perform identifier mapping for chemicals and HGNC IDs that resulted from the textual mapping. The BridgeDb service was loaded with the Metabolite BridgeDb ID Mapping Database (version HMDB4.0.20190116-CHEBI193-WIKIDATA20201104, released on 4 November 2020) [56] and the Gene/Protein BridgeDb ID Mapping Database (version 91, released on 9 May 2018) [57]. For chemicals, the CAS IDs from the AOP-Wiki XML were used as input to retrieve identifiers from ChEBI [50], ChemSpider [58], Wikidata [59, 60], ChEMBL [61], PubChem [62], Drugbank [63], KEGG [64], LIPID MAPS [65], and HMDB [66]. For genes, the HGNC IDs were used to request matching identifiers for NCBI Gene, UniProt and Ensembl [67].

File creation

All AOP-Wiki content, persistent identifiers, ontology annotations, and additional information for chemicals, genes and proteins, were stored into three RDF files using Turtle (ttl) syntax. While the central AOP-Wiki RDF file (AOP-Wiki.ttl) contains all existing AOP-Wiki components plus added chemical identifiers and identifiers mapped from PR terms in Biological Objects, the second file (AOP-Wiki-genes.ttl) contains all the text-mapped gene IDs and matching identifiers added by BridgeDb (Figure 4.1). These files are accompanied by a metadata file (AOP-Wiki-void.ttl) which describes the datasets, code, and provenance, using standard vocabularies for semantic annotations of metadata, such as Dublin Core [25, 26], Data Catalog Vocabulary [38], Friend of a Friend [28], and Vocabulary of

Interlinked Datasets [37].

4.2.3 Validation and testing of the RDF

The RDF files were validated with the IDLab Turtle validator [68], an open-source RDF validator for Turtle syntax and XSD datatype errors.

SPARQL Query:

```

1 #This query takes all Stessor chemicals in the AOP-Wiki and their mappings to ChEBI IDs
2 #Then, these ChEBI IDs are used to look up human molecular pathways in WikiPathways
3
4 PREFIX wp: <http://vocabularies.wikipathways.org/wp#>
5
6 SELECT DISTINCT ?ChemicalName ?mappedid ?LinkedStressor (group_concat(?LinkedAOP;separator=
7 (STR(?PathwayTitle) AS ?PathwayName) ?PathwayURI
8
9 WHERE{
10 ?cheLook a cheminf:000000; dc:title ?ChemicalName; dcterms:isPartOf ?LinkedStressor;
11 skos:exactMatch ?mappedid .
12 ?LinkedStressor dcterms:isPartOf ?LinkedAOPURI .
13 ?LinkedAOPURI a aopo:AdverseOutcomePathway; rdfs:label ?LinkedAOP.
14 ?mappedid a cheminf:000407; cheminf:000407 ?ChEBI.
15 SERVICE <http://sparql.wikipathways.org/sparql>{
16 ?metabolite wp:bbChEBI ?mappedid; dcterms:isPartOf ?PathwayURI.
17 ?PathwayURI a wp:Pathway; dcterms:identifier ?PathwayID;
18 dc:title ?PathwayTitle; wp:organismName "Homo sapiens"^^xsd:string .}}
19 ORDER BY DESC (?ChemicalName)

```

SPARQL Examples:

https://github.com/marvinm2/AOPWikiSNC

Type part of the query file name to search for...

Search Clear

- A. Metadata
- B. Datapump
- C. Search
- D. Simpleconversions
- E. Chemical-centered
- F. Federated

1. PathwayswithChem.rq

SPARQL results (211 results)

ChemicalName	mappedid	LinkedStressor	AOPs	Pathway_ID	PathwayName	PathwayURI
tert-Butyl hydroperoxide	https://identifiers.org/chebi/CHEBI:64090	https://identifiers.org/aop.stressor/457	AOP 296	WP4008	NO/cGMP/PKG mediated neuroprotection	https://identifiers.org/
Valproic acid	https://identifiers.org/chebi/CHEBI:39867	https://identifiers.org/aop.stressor/401	AOP 212 AOP	WP3871	Valproic acid pathway	https://identifiers.org/

Powered by AOP-Wiki and AOP-Wiki SNORQL - Cookie Policy

BigCat See also Martens et al. (2021)

Figure 4.2: The AOP-Wiki SNORQL User Interface. The AOP-Wiki SNORQL User Interface allows for user-friendly access to the AOP-Wiki RDF by syntax highlighting and through providing a SPARQL Examples panel (right panel).

Loading and testing the RDF

After validation of the RDF, the AOP-Wiki RDF was loaded in a public SPARQL endpoint (`aopwiki.rdf.bigcat-bioinformatics.org/sparql`) and is accessible through the developed SNORQL User Interface (`aopwiki.rdf.bigcat-bioinformatics.org`, Figure 4.2).

The data was tested with a Jupyter notebook that executes SPARQL queries through the SPARQL endpoint. These SPARQL queries retrieve metadata and the statistics for types of subjects, frequency of ontology usage, and the number of link-outs to the various databases [22]. All SPARQL queries used for the testing of the RDF are available in the SPARQL Examples panel in the AOP-Wiki SNORQL User Interface.

Validation of the SPARQL endpoint

On January 17th, 2021, the AOP-Wiki SPARQL endpoint was registered in YummyData [69], which monitors compliance with Linked Data standards and scoring each SPARQL endpoint on availability, freshness, operation, usefulness, validity and performance to calculate the Umaka Score. The scoring is done on a daily basis by performing SPARQL queries and HTTP requests related to the various measures for each aspect. YummyData also provides feedback on how to improve the score, which was used for improving the AOP-Wiki RDF and SPARQL endpoint.

4.3 Results

The main result of this project is an RDF schema and scripts that lead to the production of RDF content for all AOP-Wiki content with additional semantic annotations, persistent identifiers and extended identifiers for genes and chemicals, and consists of 122,576 unique triples consisting of 15,132 unique subjects, 158 unique predicates, and 53,087

unique objects (Figure 4.1). The semantic annotation was done using eight standard metadata vocabularies and seventeen domain-specific ontologies and vocabularies. We here detail these results.

The metadata vocabulary we used most is Dublin Core, of which terms are present in 49,710 triples in the AOP-Wiki RDF. Its original set of terms was used to relate various subjects to their identifier, title, description, source, and creator, and the extended set of terms was used to describe the alternative name, abstract, creation and modification date, and relational information to other subjects with 'dcterms:isPartOf'. Other standard vocabularies we extensively used are the RDF vocabulary to describe the type of subjects with the 'rdf:type' term which we used 21,541 times, and the RDF Schema vocabulary to describe the label of subjects with the 'rdfs:label' was used 6,617 times. Furthermore, the Friend Of A Friend vocabulary is used to define the webpage URLs of AOPs, KEs, KERs and stressors with 'foaf:page' a total of 3,335 times, and the term 'skos:exactMatch' was used 8,765 times to map chemical and gene/protein identifiers to other database identifiers. The NCI Thesaurus is used a total of 523 times for objects and 2,975 times for describing seven distinct properties of AOPs, KEs, KERs, such as overall assessment and applications of AOPs, biological plausibility and uncertainties of KERs, among others.

Developed for the AOP domain of research and facilitate consistent reporting, the AOPO has been used for the semantic annotations for AOP-specific elements. Terms of the AOPO were used to provide relational information for AOPs, KEs, KERs, and life-stage applicability, stressors, chemicals, and cell- and organ-terms. In total, the AOPO is used 8,913 times for predicate annotations, and 2,937 times as object annotations.

4.3.1 Adverse Outcome Pathways

The 316 AOP subjects have 26 different types of predicates to create triples (Figure 4.3). The overall most used vocabulary for

predicates is Dublin Core and its extended set of terms, creating triples for the identifier, title, alternative title, creator, abstract, description, source, access rights, creation and modification date for AOPs. The majority of these predicates also exist for other subjects. The AOPO is used to connect AOPs to KE subjects with the predicates 'has_key_event', 'has_molecular_initiating_event' and 'has_adverse_outcome', and connect with KER subjects with the predicate 'has_key_event_relationship'. Other terms of the AOPO were used for describing the overall applicability, life stage applicability, and weight of evidence. Furthermore, AOP subjects are unique to contain information on the quantitative considerations of the AOP, which is linked with 'edam:operation_3799'. The link with stressors, overall assessment description, KE essentiality, and the potential applications of the AOP were annotated using the NCI Thesaurus terms 'nci:C54571', 'nci:C25217', 'nci:C48192', and 'nci:C25725', respectively. Finally, the sex applicability of AOPs is annotated with 'pato:0000047' which stands for biological sex (Figure 4.3).

4.3.2 Key Events and Key Event Relationships

Whereas the majority of triples for KEs and KERs have predicates identical to ones for AOPs, there are properties that are unique to the 1131 KEs and 1363 KERs (Figure 4.4). For KEs, these properties include measurement methods, level of biological organization, and structured information on cell-terms, organ-terms, Biological Processes, Objects, and Actions (Figure 4.4). The measurement methods are coupled to 350 KEs with the predicate 'mmo:0000000' from the Measurement Method Ontology, which stands for measurement method, and level of biological organization is linked with the 'nci:C25664' to all KEs. Cell terms and Organ terms are described with the AOPO terms 'aopo:CellTypeContext' and 'aopo:OrganCotext', respectively. The Biological Process triples have the predicate 'go:008150' which stands for biological process, and the Biological Objects and Actions are linked with 'pato:0001241'

Adverse Outcome Pathway	Predicate	Object	Object example
	a	aopo:AdverseOutcomePathway	
	dc:identifier	Adverse Outcome Pathway (IRI)	aop:38
	rdfs:label	Label (literal)	"AOP 38"
	dc:title	Title (literal)	"Protein Alkylation leading to Liver Fibrosis"
	dcterms:alternative	Alternative title (literal)	"Protein Alkylation to Liver Fibrosis"
	dc:creator	Author (literal)	""Brigitte Landesmann...
	dcterms:abstract	Abstract (literal)	""Hepatotoxicity in general is of special interest...
	nci:C54571	Stressor (IRI)*	aop.stressor:9,aop.stressor:13,aop.stressor:60,...
	aopo:has_key_event	Key Event (IRI)*	aop.events:55,aop.events:1492,aop.events:1493,...
	aopo:has_molecular_initiating_event	Key Event (IRI)*	aop.events:244
	aopo:has_adverse_outcome	Key Event (IRI)*	aop.events:344
	aopo:has_key_event_relationship	Key Event Relationship (IRI)*	aop.relationships:269,aop.relationships:1718,...
	dc:description	Description (literal)	""Two prototypical chemicals acting via protein alkylation are...
	pato:0000047	Sex applicability (literal)	"Unspecific"
	aopo:LifeStageContext	Life stage applicability (literal)	"Not Otherwise Specified"
	aopo:AopContext	Applicability (literal)	""The described AOP is valid for both sexes and any life stage...
	edam:operation_3799	Quantitative considerations (literal)	""More advanced in vitro models systems are needed...
	aopo:has_evidence	Weight of Evidence (literal)	""Support for Essentiality of KEs...
	nci:C25725	Potential applications (literal)	""This systematic and coherent display of currently available...
	nci:C25217	Overall assessment (literal)	""Assessment of the Weight-of-Evidence supporting the AOP...
	nci:C48192	Key Event essentiality (literal)	""The essentiality of each of the KEs for this AOP...
	dc:accessRights	AOP status (literal)	"Open for citation & comment"
	foaf:page	Webpage (URL)	<https://identifiers.org/aop/38>
	dcterms:created	Date of creation (literal)	"2016-11-29T18:41:16"
	dcterms:modified	Date of latest modification (literal)	"2019-04-30T12:53:51"
	dc:source	Source (literal)	"AOPWiki"

Figure 4.3: Adverse Outcome Pathways and their properties in RDF. From left to right, the columns indicate predicates, objects and an example of the object taken from the RDF. Asterisks indicate the object IRIs that connect to other subjects in the RDF.

and 'pato:0000001', respectively. These ontological annotations are connected through IRIs of other subjects in the RDF. A shared predicate with AOPs is the term 'nci:C54571' for MIEs that have links to stressors in the AOP-Wiki (Figure 4.4).

Properties that are specific to the 1363 KERs are the biological plausibility, empirical support, uncertainties, which we linked with the predicates 'nci:C80263', 'edam:data_2042' and 'nci:71478'. These stand for the rationale, evidence, and uncertainty, respectively. Also, the RDF connects the upstream and downstream KEs of KERs with the terms 'aopo:has_upstream_key_event' and 'aopo:has_downstream_key_event' from the AOPO (Figure 4.4B).

A

	Predicate	Object	Object example
Key Event	a	aopo:KeyEvent	
	dc:identifier	Key Event (IRI)	aop.events:1502
	rdfs:label	Label (literal)	"KE 1502"
	dc:title	Title (literal)	"Histone deacetylase inhibition"
	dcterms:alternative	Alternative title (literal)	"Histone deacetylase inhibition"
	nci:C25664	Level of biological organization (literal)	""Molecular""
	dc:description	Description (literal)	""The inhibition of HDAC by HDIs is well conserved...
	edam:data_1025	HGNC ID (IRI)*	hgnc:HDAC9,hgnc:MAA,hgnc:PRDX2
	mmo:0000000	Measurement method (literal)	""The measurement of HDAC inhibition monitors changes...
	nci:C54571	Stressor (IRI)*	aop.stressor:340,aop.stressor:341,aop.stressor:342,...
	aopo:CellTypeContext	Cell-term (IRI)*	cl:0000000
	aopo:OrganContext	Organ-term (IRI)*	uberon:0000062
	go:0008150	Biological process (IRI)*	go:0004857
	pato:0001241	Biological object (IRI)*	pr:000008478
	pato:0000001	Biological action (IRI)*	"Wiki:2"
	ncbitaxon:131567	Taxonomy (IRI or literal)*	ncbitaxon:10116,"WCS_9606",ncbitaxon:10090
	pato:0000047	Sex applicability (literal)	"Unspecific"
	aopo:LifeStageContext	Life stage applicability (literal)	"All life stages"
	foaf:page	Webpage (URL)	<https://identifiers.org/aop.events/1502>
	dcterms:isPartOf	Adverse Outcome Pathway (IRI)*	aop:212,aop:274,aop:275
	dc:source	Source (literal)	"AOPWiki"

B

	Predicate	Object	Object example
Key Event Relationship	a	aopo:KeyEventRelationship	
	dc:identifier	Key Event Relationship (URI)	aop.relationships:865
	rdfs:label	Label (literal)	"KER 865"
	aopo:has_upstream_key_event	Key Event (URI)*	aop.events:844
	aopo:has_downstream_key_event	Key Event (URI)*	aop.events:845
	dc:description	Description (literal)	""One of the oxidation products of uroporphyrinogen...
	nci:C80263	Biological plausibility (literal)	""Reduced UROD enzyme activity, not protein levels...
	edam:data_2042	Empirical support (literal)	""Include consideration of temporal concordance...
	nci:C71478	Uncertainties or inconsistencies (literal)	""The precise mechanism of UROD inhibition has yet...
	edam:data_1025	HGNC ID (IRI)*	hgnc:UROD
	ncbitaxon:131567	Taxonomy (IRI or literal)*	ncbitaxon:10090,ncbitaxon:10116,"WCS_9606"
	pato:0000047	Sex applicability (literal)	"Unspecific"
	aopo:LifeStageContext	Life stage applicability (literal)	"All life stages","Adult","Juvenile"
	foaf:page	Webpage (URL)	<https://identifiers.org/aop.relationships/865>
	dcterms:created	Date of creation (literal)	2016-11-29T18:41:35"
	dcterms:modified	Date of latest modification (literal)	"2018-05-30T10:58:18"
	dcterms:isPartOf	Adverse Outcome Pathway (URI)*	aop:131

Figure 4.4: Key Events and their properties in RDF. From left to right, the columns indicate predicates, objects and an example of the object taken from the RDF. Asterisks indicate the object IRIs that connect to other subjects in the RDF.

Similar to AOPs, triples describing the applicability of KEs and KERs exist for life stage and sex. However, unique to KEs and KERs is the

taxonomic applicability which is linked with 'ncbitaxon:131567'. Also, KE and KER triples describe relational information to AOPs with 'dc-terms:isPartOf'.

4.3.3 Stressors and Chemicals

The RDF contains general stressor information such as descriptions and identifiers, and stressor triples describe their connections to MIEs and AOPs with 'dcterms:isPartOf'. 63% of the 523 stressors are also linked to chemicals with the predicate 'aopo:has_chemical_entity' (Figure 4.5A). The 329 chemical subjects are annotated with CAS and CompTox identifiers, and 320 also have InChIKeys, all of which we annotated with the Chemical Information Ontology (Figure 4.5B). Furthermore, the chemicals have predicate 'skos:exactMatch' to link to all mapped chemical subjects present in the RDF, providing link-outs to nine additional external databases (Figure 4.6A). We annotated these with 'rdf:type' and terms from the Chemical Information Ontology. In total, there are 3,904 link-outs to twelve different chemical databases, allowing users to explore the AOP-Wiki by using their preferred type of chemical identifiers.

4.3.4 Ontological annotations

Since the taxonomies, cell terms, organ terms, and the KE components all already have ontological annotations in the AOP-Wiki, they have the same properties that describe their type (4.6B), identifier, title and source. These titles are based on the user-provided entries in the AOP-Wiki. Unique for the biological objects annotated with the Protein Ontology is the inclusion of the 'skos:exactMatch' predicate linking to 576 matching identifiers from UniProt, HGNC and NCBI Gene (Figure 4.6C) to 126 Protein Ontology tags which are used in 166 KEs.

A	Stressor	Predicate	Object	Object example
		a	nci:C54571	
		dc:identifier	Stressor (IRI)	aop.stressor:208
		rdfs:label	Label (literal)	"Stressor 208"
		dc:title	Title (literal)	"Gemfibrozil"
		dc:description	Description (literal)	""Fibrate drug""
		aopo:has_chemical_entity	Chemical identifier (IRI)*	cas:25812-30-0
		foaf:page	Webpage (URL)	<https://identifiers.org/aop.stressor/208>
		dcterms:created	Date of creation (literal)	"2016-11-29T18:42:27"
		dcterms:modified	Date of latest modification (literal)	"2020-03-31T10:24:40"
B	Chemical	Predicate	Object	Object example
		a	cheminf:000000 cheminf:000446	
		dc:identifier	CAS identifier (IRI)	cas:103-90-2
		cheminf:000446	CAS identifier (literal)	"103-90-2"
		dc:title	Title (literal)	"Acetaminophen"
		dcterms:alternative	Synonyms (literal)	"4-Acetamidophenol", "Paracetamol", ...
		cheminf:000059	InChIKey (IRI)	inchikey:RZVAJINKPMORJF-UHFFFAOYSA-N
		cheminf:000568	CompTox identifier (IRI)	comptox:DTXSID2020006
		skos:exactMatch	Matched identifier (IRI)*	chebi:46195,chemspider:1906,wikidata:Q57055,...
		dcterms:isPartOf	Stressor (IRI)*	aop.stressor:57

Figure 4.5: Stressors and chemicals and their properties in RDF. From left to right, the columns indicate predicates, objects and an example of the object taken from the RDF. Asterisks indicate the object IRIs that connect to other subjects in the RDF.

4.3.5 Gene and protein identifiers

Extending the links of KEs and KERs with genes and proteins, RDF triples of 846 unique text-mapped gene identifiers on KEs and KERs are stored in a separate file. These make triples of KE and KER subjects to link to the mapped HGNC identifiers with the predicate 'edam:data_1025', which stands for Gene identifier. These HGNC identifiers are subjects themselves, and have the 'skos:exactMatch' predicate to link to matching identifiers from UniProt, NCBI Gene, and Ensembl, providing a total of 6,001 link-outs using this method

A			B	
Database	rdfs:type	#	Subject	rdfs:type
CAS	cheminf:000446	329	Cell-term	aopo:CellTypeContext
ChEBI	cheminf:000407	803	Organ-term	aopo:OrganContext
ChemSpider	cheminf:000405	343	Taxonomy	ncbitaxon:131567
ChEMBL compound	cheminf:000412	287	Biological Process	go:0008150
CompTox	cheminf:000568	329	Biological object	pato:0001241
Drugbank	cheminf:000406	161		Matched identifier (IRI)*
HMDB	cheminf:000408	363	Biological action	pato:0000001
InChIKey	cheminf:000059	320		
KEGG compound	cheminf:000409	264		
Lipid maps	cheminf:000564	30		
PubChem compound	cheminf:000140	338		
Wikidata	cheminf:000567	328		

C				
Database	rdfs:type	#1	#2	
HGNC	edam:data_2298	97	846	
Ensembl	edam:data_1033	-	813	
Entrez Gene	edam:data_1027	32	804	
UniProt	edam:data_2291	447	3653	

Figure 4.6: **Ontology annotations and molecular identifiers.** A. Cell-terms, organ-terms, taxonomies, Key Event Components and type annotation in the RDF. The asterisk indicates the matching identifiers to other subjects in the RDF. B. Gene and protein databases, their type annotation, and the number of identifiers present in the RDF. Values in #1 are based on Protein Ontology mappings, and values in #2 are based on textual mapping with HGNC symbols. C. Chemical databases, their type annotation and the number of identifiers present in the RDF.

(Figure 4.6C).

4.3.6 Federated SPARQL query example

The addition of external identifiers facilitates the execution of federated SPARQL queries to combine resources, such as the example shown in Figure 4.2, which is located in the SPARQL example panel under the folder 'F. Federated'. The SPARQL query looks up all entities defined as a Chemical in the AOP-Wiki RDF with the predicate

and subject 'a cheminf:000000', and extracts their names, mapped identifiers, and linked stressors. Next, the query looks for AOPs that mention the stressor. The next line in the query restricts the results by explicitly adding the type of 'a aopo:AdverseOutcomePathway'. Similarly, the query makes sure that the mapped identifier of the chemical is from the ChEBI database by defining the mapped identifier as type 'cheminf:000407'. This is followed by the 'SERVICE' statement which defines the external SPARQL endpoint and defines the federated part of the query. In this example, the external SPARQL endpoint is WikiPathways (sparql.wikipathways.org/sparql), where the ChEBI identifier is used to match relevant pathways using 'dcterms:isPartOf'. The last part of the SPARQL query retrieves information about the pathway and filters for human pathways. This SPARQL query results in a table of 211 rows with each chemical, their ChEBI identifier, related stressor and AOP, and the molecular pathways that the chemical is involved in. These linked pathways provide additional insights in the general functions of chemicals and their involvement in cellular processes and responses.

4.3.7 Validation by YummyData

As an external validation of the developed SPARQL endpoint and RDF, YummyData indexes and ranks the AOP-Wiki SPARQL endpoint based on an array of tests on a daily basis (yummydata.org/endpoint/142). As of February 2021, the AOP-Wiki SPARQL endpoint is considered A rank with a Umaka score above 80, consistently scoring above average on all aspects. Incidentally the Umaka score drops slightly below 80, giving it a B rank.

4.4 Discussion

The work described in this paper has led to the creation of AOP-Wiki RDF based on the existing AOP-Wiki XML, combined with a variety of ontologies and enriched with persistent identifiers. Besides, the

data is extended with additional identifiers for chemicals, proteins and genes. The RDF has been validated, loaded on a SPARQL endpoint, and tested using a Jupyter notebook, and is indexed in YummyData for external validation of the SPARQL endpoint. These developments made AOP-Wiki content ready for use in risk assessment workflows, through coding environments, or in federated SPARQL queries.

Because the AOPO is developed for consistent reporting in the domain of AOPs and allowing the integration of data and tools [19], it was an obvious choice to implement the AOPO for semantic annotations in the AOP-Wiki RDF. The ontology includes a variety of AOP-specific definitions for properties and classes which were directly applicable to AOP-Wiki content in the RDF. However, these do not fully cover all types of entities and relationships that exist in the AOP-Wiki. For example, while it has terms to describe the connections between AOPs, KEs and KERs, there is no annotation for the link with stressors. Similarly, terms are lacking for sex and taxonomic applicability, KE components, KER-specific information, and AOP assessment sections such as KE essentiality and quantitative considerations, among others.

For the terms missing in the AOPO, we selected definitions from a wide range of other ontologies and vocabularies for the semantification of predicates and subjects, including NCI Thesaurus [31], NCBI of Organismal Classification, Gene Ontology, and Measurement Method Ontology, among others. Whereas the majority of the AOP-Wiki contents are generic and can be described with well-established metadata ontologies, some properties of AOPs, KEs and KERs could not be found, leading to the selection of more general terms, lacking detail that would be preferred. With the conversion to RDF, most of the necessary terms have been uncovered and documented, and these will be added to the AOPO.

Because the realm of AOPs includes many types of data, knowledge, repositories and services, the development and implementation of a central, community-wide vocabulary would facilitate their integration. Since the AOPO has been developed to fill that purpose, it could be ex-

tended to include descriptions of classes and properties for all ontology terms used in the AOP-Wiki RDF. Having a central, field-wide ontology for AOPs helps maintaining a high quality vocabulary through continuous development and involvement of the community. Such an ontology would facilitate the annotation and integration of data, resources and tools, as is done with the eNanoMapper ontology in the nanotoxicology community [70].

With the increased importance of consistent use of identifiers to integrate knowledge and data, our implementation of persistent identifiers for AOPs, KEs, KERs, Stressors, chemicals, proteins and genes will benefit the integration of the AOP-Wiki with other resources, data, and tools [71]. These persistent identifiers stored in the MIRIAM registry are stable, unique, resolvable, documented, and directly link to the corresponding entries in the databases [21, 72]. Furthermore, our efforts have introduced additional content to the AOP-Wiki RDF through ID mappings and text-mapping for chemicals and genes, providing more ways of extrapolating the data and linking with other resources and data.

While we have added molecular identifiers to increase the number of link-outs and improve the usefulness of the database, our addition of genes through textual ID mapping does introduce errors to the AOP-Wiki RDF. The automated process on free-text content assumes good practice in writing gene symbols and names according to the HGNC guidelines [73]. Although HGNC strives for stable gene symbols and makes justified changes for problematic ones [54], some gene symbols still overlap with free-text abbreviations in the AOP-Wiki and are therefore falsely recognised.

Opportunities exist to improve the AOP-Wiki machine-readability by having more structured text and annotations for molecular entities, pathways, organs, species and other biological concepts that are relevant for AOPs and not yet covered by the KE Components. Text-mining tools, such as ProMiner [74], ContentMine [75] and PolySearch2 [76], could be implemented for extracting biological

concepts and understanding associations to add more structured information in the AOP-Wiki and facilitate the integration with other databases and tools. Once such concepts are recognised and extracted, the RDF could be extended and increase the interoperability of the AOP-Wiki with external databases, such as pathway databases.

The AOP-Wiki RDF allows for new and efficient ways of accessing the data, and using it to answer questions. By loading the RDF in a SPARQL endpoint, SPARQL queries can be used to access the data and extract all necessary information. It allows complex queries across the complete AOP-Wiki database, optional filtering for any variable, and requesting an outputs suitable for answering the research question. Furthermore, these SPARQL queries can be executed from most coding environments as part of larger workflows or data pipelines. It also facilitates direct linkage of databases through federated SPARQL queries, which returns information across databases with a single query. Any database with a SPARQL endpoint can be used for such questions across databases, such as WikiPathways [16, 17], Wikidata [59, 60], neXtProt [15], UniProt [53], ChEMBL-RDF [77], DisGeNET [78, 79], Rhea [80], Pathway Commons [81], among others. Examples of federated queries are stored in the SPARQL Examples panel in the AOP-Wiki SNORQL User Interface `aopwiki.rdf.bigcat-bioinformatics.org` (Figure 4.2).

Another way of using the RDF to extract AOP-Wiki content is through a web service such as the git repository linked data API constructor (grlc) [82], which can build a Web API on top of a SPARQL endpoint with predefined SPARQL queries. While more straight-forward than SPARQL queries, the API is limited to the predefined SPARQL queries and variables implemented in these.

An advantage of creating RDF for the AOP-Wiki is the ability to link and expand AOP-Wiki content with information from other databases. For example, the AOP-DB combines knowledge from the AOP-Wiki with annotations of genes, chemicals, diseases, tissues, pathways, on-

tologies, and ToxCast data [83, 84]. Future work should focus on integrating such efforts by developing RDF and thereby allow full integration of their data and tools [85] with the AOP-Wiki RDF and other databases.

In terms of compliance with Linked Data standards according to the YummyData registry, the AOP-Wiki SPARQL endpoint consistently scores above average in availability, freshness, operation, usefulness, validity, and performance [69]. With a consistent A rank, the AOP-Wiki SPARQL endpoint is placed among the 10% best-scoring of the 70+ SPARQL endpoints registered in YummyData. However, incidentally the Umaka score dropped due to slow server response or when the service has been down for maintenance or data loading. Based on the feedback given by YummyData, a point for improvement would be supporting more response formats of the SPARQL endpoint to increase the usefulness score.

The implementation of compact identifiers and development of a formal, machine-readable RDF schema makes the content of the AOP-Wiki more findable and interoperable to other components of the database, and by allowing SPARQL queries and API to explore the data, the AOP-Wiki database was made accessible through new methods. Furthermore, the addition of link-outs to various chemical, gene and protein databases, as well as the data storage in an RDF format and implementing Linked Open Data standards, has made the data more interoperable with other databases and tools. Furthermore, the AOP-Wiki has recently introduced licenses on its content, and the code for the creation and validation of the AOP-Wiki RDF are available under MIT license. These provide clear statements and terms of using, sharing and modifying the content. Taken together with the addition of metadata and semantic information represented by ontology annotations, the content of the AOP-Wiki has been made more accessible and reusable. Therefore, the development of the AOP-Wiki RDF addresses all major FAIR principles [13].

Overall, the AOP-Wiki RDF allows for new ways of exploring the data,

using it in automated workflows, from coding environments, or directly through a SPARQL endpoint. With the implementation also comes the possibility to execute federated queries to combine data of multiple resources and answer more elaborate questions. For example, Key Events can be linked to the results of ToxCast assays that measure the activity of the protein described, or molecular pathways can be explored for a more detailed description of mechanistic processes.

4.5 Data links

All data and code used in this manuscript are publicly available. The main conversion code, statistics code and the created Turtle files can be found on github.com/marvinm2/AOPWikiRDF. The AOP-Wiki XML can be downloaded on aopwiki.org/downloads. The HGNC mapping file can be downloaded via genenames.org and the Protein Ontology mapping file can be downloaded with proconsortium.org/download/current.

References

- [1] Gerald T. Ankley et al. "Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment". *Environmental Toxicology and Chemistry* 29.3 (Mar. 2010), pp. 730–741. DOI: 10.1002/etc.34.
- [2] Daniel Krewski et al. "Toxicity Testing in the 21st Century: Implications for Human Health Risk Assessment". *Risk Analysis* 29.4 (Apr. 2009), pp. 474–479. DOI: 10.1111/j.1539-6924.2008.01150.x.
- [3] Marcel Leist et al. "Adverse outcome pathways: opportunities, limitations and open questions". *Archives of Toxicology* 91.11 (Nov. 2017), pp. 3477–3505. DOI: 10.1007/s00204-017-2045-3.
- [4] Catherine Willett. "The Use of Adverse Outcome Pathways (AOPs) to Support Chemical Safety Decisions Within the Context of Integrated Approaches to Testing and Assessment (IATA)". *Alternatives to Animal Testing* (2019), pp. 83–90. DOI: 10.1007/978-981-13-2447-5_11.
- [5] Mathieu Vinken et al. "Adverse outcome pathways: a concise introduction for toxicologists". *Archives of Toxicology* 91.11 (Nov. 2017), pp. 3697–3707. DOI: 10.1007/s00204-017-2020-z.
- [6] Daniel L. Villeneuve et al. "Adverse outcome pathway (AOP) development I: strategies and principles." *Toxicological Sciences* 142.2 (Dec. 2014), pp. 312–320. DOI: 10.1093/toxsci/kfu199.

-
- [7] Jaeseong Jeong and Jinhee Choi. "Use of adverse outcome pathways in chemical toxicity testing: potential advantages and limitations". *Environmental Health and Toxicology* 33.1 (Dec. 2017), e2018002. DOI: 10.5620/eh.t.e2018002.
- [8] Noffisat O. Oki and Stephen W. Edwards. "An integrative data mining approach to identifying adverse outcome pathway signatures". *Toxicology* 350-352 (Mar. 2016), pp. 49–61. DOI: 10.1016/j.tox.2016.04.004.
- [9] Antony J. Williams et al. "The CompTox Chemistry Dashboard: A community data resource for environmental chemistry". *Journal of Cheminformatics* 9.1 (Nov. 2017), p. 61. DOI: 10.1186/s13321-017-0247-6.
- [10] Scott Federhen. "The NCBI Taxonomy database". *Nucleic Acids Research* 40.D1 (Jan. 2012), pp. D136–D143. DOI: 10.1093/nar/gkr1178.
- [11] Cataia Ives et al. "Creating a Structured AOP Knowledgebase via Ontology-Based Annotations." *Applied in vitro toxicology* 3.4 (Dec. 2017), pp. 298–311. DOI: 10.1089/aivt.2017.0017.
- [12] Richard Cyganiak, David Wood, and Markus Lanthaler. *RDF 1.1 Concepts and Abstract Syntax*. 2014.
- [13] Mark D. Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific Data* 3.1 (Mar. 2016), p. 160018. DOI: 10.1038/sdata.2016.18.
- [14] Christine Chichester et al. "Converting neXtProt into Linked Data and nanopublications". *Semantic Web* 6.2 (Jan. 2015), pp. 147–153. DOI: 10.3233/SW-140149.
- [15] Monique Zahn-Zabal et al. "The neXtProt knowledgebase in 2020: Data, tools and usability improvements". *Nucleic Acids Research* 48.D1 (Jan. 2020), pp. D328–D334. DOI: 10.1093/nar/gkz995.
- [16] Andra Waagmeester et al. "Using the Semantic Web for Rapid Integration of WikiPathways with Other Biological Online Data Resources". *PLOS Computational Biology* 12.6 (June 2016). Ed. by Christos A. Ouzounis, e1004989. DOI: 10.1371/journal.pcbi.1004989.
- [17] Marvin Martens et al. "WikiPathways: connecting communities". *Nucleic Acids Research* 49.D1 (Jan. 2021), pp. D613–D621. DOI: 10.1093/nar/gkaa1024.
- [18] Fiona Sewell et al. "The future trajectory of adverse outcome pathways: a commentary". *Archives of Toxicology* 92.4 (2018), pp. 1657–1661. DOI: 10.1007/s00204-018-2183-2.
- [19] Lyle D. Burgoon. "The AOPontology: A semantic artificial intelligence tool for predictive toxicology". *Applied In Vitro Toxicology* 3.3 (Sept. 2017), pp. 278–281. DOI: 10.1089/aivt.2017.0012.
- [20] Clemens Wittwehr et al. "How adverse outcome pathways can aid the development and use of computational prediction models for regulatory toxicology". *Toxicological Sciences* 155.2 (Feb. 2017), pp. 326–336. DOI: 10.1093/toxsci/kfw207.
- [21] Nick Juty, Nicolas Le Novère, and Camille Laibe. "Identifiers.org and MIRIAM Registry: community resources to provide persistent identification". *Nucleic Acids Research* 40.D1 (Dec. 2011), pp. D580–D586. DOI: 10.1093/nar/gkr1097.
- [22] Marvin Martens. *marvinm2/AOPWikiRDF: Finished notebooks*. Nov. 2020. DOI: 10.5281/ZENODO.4292485.
- [23] Patricia L. Whetzel et al. "BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use on-

- tologies in software applications." *Nucleic Acids Research* 39.suppl_2 (July 2011), W541–W545. DOI: 10.1093/nar/gkr469.
- [24] Deborah L McGuinness and Frank van Harmelen. *OWL Web Ontology Language Overview*. <https://www.w3.org/TR/owl-features/>. 2004.
 - [25] DCMI Usage Board. *DCMI: Dublin Core™ Metadata Element Set, Version 1.1: Reference Description*. 2012.
 - [26] DCMI Usage Board. *DCMI: DCMI Metadata Terms*. 2020.
 - [27] W3C, Dan Brickley, and R V Guha. *RDF Schema 1.1*. 2014.
 - [28] Dan Brickley and Libby Miller. *FOAF Vocabulary Specification*. 2014.
 - [29] Chris Mungall et al. *pato-ontology/pato: 2018-11-12 release*. Nov. 2018. DOI: 10.5281/ZENODO.1484533.
 - [30] Janna Hastings et al. "The Chemical Information Ontology: Provenance and Disambiguation for Chemical Data on the Biological Semantic Web". *PLoS ONE* 6.10 (Oct. 2011). Ed. by Franca Fraternali, e25513. DOI: 10.1371/journal.pone.0025513.
 - [31] Nicholas Sioutos et al. "NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information". *Journal of Biomedical Informatics* 40.1 (Feb. 2007), pp. 30–43. DOI: 10.1016/j.jbi.2006.02.013.
 - [32] Jennifer R. Smith et al. "The clinical measurement, measurement method and experimental condition ontologies: Expansion, improvements and new applications". *Journal of Biomedical Semantics* 4.1 (Oct. 2013), p. 26. DOI: 10.1186/2041-1480-4-26.
 - [33] Alistair Miles and Sean Bechhofer. *SKOS Simple Knowledge Organization System Reference*. 2009.
 - [34] Michael Ashburner et al. "Gene Ontology: tool for the unification of biology". *Nature Genetics* 25.1 (May 2000), pp. 25–29. DOI: 10.1038/75556.
 - [35] Jon Ison et al. "EDAM: An ontology of bioinformatics operations, types of data and identifiers, topics and formats". *Bioinformatics* 29.10 (May 2013), pp. 1325–1332. DOI: 10.1093/bioinformatics/btt113.
 - [36] Paolo Ciccarese et al. "PAV ontology: provenance, authoring and versioning". *Journal of Biomedical Semantics* 4.1 (2013), p. 37. DOI: 10.1186/2041-1480-4-37.
 - [37] Keith Alexander et al. *Describing Linked Datasets with the VoID Vocabulary*. <https://www.w3.org/TR/void/>. 2011.
 - [38] Simon Cox et al. *Data Catalog Vocabulary (DCAT) - Version 2*. 2020.
 - [39] Jonathan Bard, Seung Y. Rhee, and Michael Ashburner. "An ontology for cell types". *Genome biology* 6.2 (2005), R21. DOI: 10.1186/gb-2005-6-2-r21.
 - [40] Christopher J. Mungall et al. "Uberon, an integrative multi-species anatomy ontology". *Genome Biology* 13.1 (Jan. 2012), R5. DOI: 10.1186/gb-2012-13-1-r5.
 - [41] Henning Hermjakob et al. "The HUPO PSI's Molecular Interaction format - A community standard for the representation of protein interaction data". *Nature Biotechnology* 22.2 (Feb. 2004), pp. 177–183. DOI: 10.1038/nbt926.
 - [42] Cynthia L. Smith and Janan T. Eppig. "The mammalian phenotype ontology: Enabling robust annotation and comparative analysis". *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 1.3 (Nov. 2009), pp. 390–399. DOI: 10.1002/wsbm.44.

-
- [43] N. Baumann. "How to use the medical subject headings (MeSH)". *International Journal of Clinical Practice* 70.2 (Feb. 2016), pp. 171–174. DOI: 10.1111/ijcp.12767.
- [44] Sebastian Köhler et al. "Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources". *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D1018–D1027. DOI: 10.1093/nar/gky1105.
- [45] Ramona L. Walls et al. "Meeting report: advancing practical applications of biodiversity ontologies". *Standards in Genomic Sciences* 9.1 (Dec. 2014), p. 17. DOI: 10.1186/1944-3277-9-17.
- [46] Ramona L. Walls et al. "Semantics in support of biodiversity knowledge discovery: an introduction to the biological collections ontology and related ontologies." *PloS one* 9.3 (Mar. 2014), e89606. DOI: 10.1371/journal.pone.0089606.
- [47] Georgios V. Gkoutos, Paul N. Schofield, and Robert Hoehndorf. "The Neurobehavior Ontology. An Ontology for Annotation and Integration of Behavior and Behavioral Phenotypes." *International Review of Neurobiology*. Vol. 103. Academic Press Inc., Jan. 2012, pp. 69–87. DOI: 10.1016/B978-0-12-388408-4.00004-6.
- [48] Carissa A. Park et al. "The Vertebrate Trait Ontology: A controlled vocabulary for the annotation of trait data across species". *Journal of Biomedical Semantics* 4.1 (Aug. 2013), p. 13. DOI: 10.1186/2041-1480-4-13.
- [49] Darren A Natale et al. "The Protein Ontology: a structured representation of protein forms and complexes." *Nucleic Acids Research* 39.Database issue (Jan. 2011). DOI: 10.1093/nar/gkq907.
- [50] Kirill Degtyarenko et al. "ChEBI: a database and ontology for chemical entities of biological interest." *Nucleic Acids Research* 36 (Jan. 2008), pp. D344–D350. DOI: 10.1093/nar/gkm791.
- [51] Cornelius Rosse and José L.V. Mejino. "A reference ontology for biomedical informatics: the Foundational Model of Anatomy". *Journal of Biomedical Informatics* 36.6 (Dec. 2003), pp. 478–500. DOI: 10.1016/j.jbi.2003.11.007.
- [52] Donna Maglott. "Entrez Gene: gene-centered information at NCBI". *Nucleic Acids Research* 33.Database issue (Dec. 2004), pp. D54–D58. DOI: 10.1093/nar/gki031.
- [53] Alex Bateman and UniProt Consortium. "UniProt: a worldwide hub of protein knowledge". *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D506–D515. DOI: 10.1093/nar/gky1049.
- [54] Bryony Braschi et al. "Genenames.org: the HGNC and VGNC resources in 2019". *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D786–D792. DOI: 10.1093/nar/gky930.
- [55] Martijn P van Iersel et al. "The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services". *BMC Bioinformatics* 11.1 (2010), p. 5. DOI: 10.1186/1471-2105-11-5.
- [56] De Sl. *Metabolite BridgeDb ID Mapping Database (20201104)*. 2020. DOI: 10.6084/m9.figshare.13187585.v1.
- [57] BiGCaT. *Gene/Protein BridgeDb ID Mapping Database (Ensembl 91)*. 2020. DOI: 10.5281/zenodo.3667670.
- [58] Harry E. Pence and Antony Williams. "Chemspider: An online chemical information resource". *Journal of Chemical Education* 87.11 (Nov. 2010), pp. 1123–1124. DOI: 10.1021/ed100697w.

- [59] Andra Waagmeester et al. "A protocol for adding knowledge to Wikidata, a case report". *bioRxiv* (June 2020), p. 2020.04.05.026336. DOI: 10.1101/2020.04.05.026336.
- [60] Fredo Erxleben et al. "Introducing Wikidata to the Linked Data Web". *Mika P. et al. (eds) The Semantic Web – ISWC 2014. ISWC 2014. Lecture Notes in Computer Science*. Vol. 8796. Springer, Cham, Oct. 2014, pp. 50–65. DOI: 10.1007/978-3-319-11964-9_4.
- [61] Anna Gaulton et al. "ChEMBL: a large-scale bioactivity database for drug discovery." *Nucleic Acids Research* 40.Database issue (Jan. 2012), pp. D1100–D1107. DOI: 10.1093/nar/gkr777.
- [62] Sunghwan Kim et al. "PubChem substance and compound databases". *Nucleic Acids Research* 44.D1 (2016). DOI: 10.1093/nar/gkv951.
- [63] David. S. Wishart. "DrugBank: a comprehensive resource for in silico drug discovery and exploration". *Nucleic Acids Research* 34.90001 (Jan. 2006). DOI: 10.1093/nar/gkj067.
- [64] M. Kanehisa. "KEGG: Kyoto Encyclopedia of Genes and Genomes". *Nucleic Acids Research* 28.1 (Jan. 2000), pp. 27–30. DOI: 10.1093/nar/28.1.27.
- [65] Eoin Fahy et al. "Update of the LIPID MAPS comprehensive classification system for lipids". *Journal of Lipid Research* 50.SUPPL. (Apr. 2009), S9–S14. DOI: 10.1194/jlr.R800095-JLR200.
- [66] David S. Wishart et al. "HMDB: Database Statistics". *Nucleic Acids Research* 35 (2007), pp. D521–D526. DOI: 10.1093/nar/gkl923.
- [67] T. Hubbard. "The Ensembl genome database project". *Nucleic Acids Research* 30.1 (2002), pp. 38–41. DOI: 10.1093/nar/30.1.38.
- [68] IDLab - Ghent University. *IDLabResearch/TurtleValidator: A Turtle validator on command line and in browser*. 2020.
- [69] Yasunori Yamamoto, Atsuko Yamaguchi, and Andrea Splendiani. "Yummy-Data: providing high-quality open life science data". *Database* 2018.2018 (Jan. 2018), p. 22. DOI: 10.1093/database/bay022.
- [70] Janna Hastings et al. "eNanoMapper: Harnessing ontologies to enable data integration for nanomaterial risk assessment". *Journal of Biomedical Semantics* 6.1 (Mar. 2015), p. 10. DOI: 10.1186/s13326-015-0005-5.
- [71] Sarala M Wimalaratne et al. "Uniform resolution of compact identifiers for biomedical data". *Scientific Data* 5.180029 (). DOI: 10.1038/sdata.2018.29.
- [72] Julie A. McMurphy et al. "Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data". *PLoS Biology* 15.6 (June 2017), e2001414. DOI: 10.1371/journal.pbio.2001414.
- [73] Hester M. Wain et al. "Guidelines for human gene nomenclature". *Genomics* 79.4 (Apr. 2002), pp. 464–470. DOI: 10.1006/geno.2002.6748.
- [74] Daniel Hanisch et al. "ProMiner: Rule-based protein and gene entity recognition". *BMC Bioinformatics* 6 (May 2005), S14. DOI: 10.1186/1471-2105-6-S1-S14.
- [75] Richard Smith-Unna and Peter Murray-Rust. "The ContentMine Scraping Stack: Literature-scale Content Mining with Community-maintained Collections of Declarative Scrapers". *D-Lib Magazine* 20.11/12 (Nov. 2014). DOI: 10.1045/november14-smith-unna.
- [76] Yifeng Liu, Yongjie Liang, and David Wishart. "PolySearch2: a significantly improved text-mining system for discovering associations between human dis-

-
- eases, genes, drugs, metabolites, toxins and more". *Nucleic Acids Research* 43.W1 (2015), pp. 535–542. DOI: 10.1093/nar/gkv383.
- [77] Egon L Willighagen et al. "The ChEMBL database as linked open data". *Journal of Cheminformatics* 5.5 (Dec. 2013), p. 23. DOI: 10.1186/1758-2946-5-23.
- [78] Janet Piñero et al. "The DisGeNET knowledge platform for disease genomics: 2019 update". *Nucleic Acids Research* 48.D1 (2020), pp. D845–D855. DOI: 10.1093/nar/gkz1021.
- [79] Núria Queralt-Rosinach et al. "DisGeNET-RDF: Harnessing the innovative power of the Semantic Web to explore the genetic basis of diseases". *Bioinformatics* 32.14 (2016), pp. 2236–2238. DOI: 10.1093/bioinformatics/btw214.
- [80] Thierry Lombardot et al. "Updates in Rhea: SPARQLing biochemical reaction data". *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D596–D600. DOI: 10.1093/nar/gky876.
- [81] Ethan G Cerami et al. "Pathway Commons, a web resource for biological pathway data". *Nucleic Acids Research* 39.Database (Jan. 2011), pp. D685–D690. DOI: 10.1093/nar/gkq1039.
- [82] Albert Meroño-Peñuela et al. "CLARIAH/grlc: January 2020 patch" (Jan. 2020). DOI: 10.5281/ZENODO.3606813.
- [83] Holly M. Mortensen et al. "Leveraging human genetic and adverse outcome pathway (AOP) data to inform susceptibility in human health risk assessment". *Mammalian Genome* 29.1-2 (Feb. 2018), pp. 190–204. DOI: 10.1007/s00335-018-9738-7.
- [84] Maureen E. Pittman et al. "AOP-DB: A database resource for the exploration of Adverse Outcome Pathways through integrated association networks". *Toxicology and Applied Pharmacology* 343 (Mar. 2018), pp. 71–83. DOI: 10.1016/j.taap.2018.02.006.
- [85] Paul S. Price, Annie M. Jarabek, and Lyle D. Burgoon. "Organizing mechanism-related information on chemical interactions using a framework based on the aggregate exposure and adverse outcome pathways". *Environment International* 138 (May 2020), p. 105673. DOI: 10.1016/j.envint.2020.105673.

Annex

Table 4.2: Prefixes in the RDF for the Key Event Component annotations.

Ontology name	Prefix in RDF	Base IRI
Cell Ontology [39]	cl	http://purl.obolibrary.org/obo/CL_
Uber-anatomy ontology [40]	uberon	http://purl.obolibrary.org/obo/UBERON_
Gene Ontology [34]	go	http://purl.obolibrary.org/obo/GO_
Molecular Interactions Controlled Vocabulary [41]	mi	http://purl.obolibrary.org/obo/MI_
Mammalian Phenotype Ontology [42]	mp	http://purl.obolibrary.org/obo/MP_
Medical Subject Headings [43]	mesh	http://purl.bioontology.org/ontology/MESH/
Human Phenotype Ontology [44]	hp	http://purl.obolibrary.org/obo/HP_
Population and Community Ontology [45, 46]	pco	http://purl.obolibrary.org/obo/PCO_
Neuro Behavior Ontology [47]	nbo	http://purl.obolibrary.org/obo/NBO_
Vertebrate trait ontology [48]	vt	http://purl.obolibrary.org/obo/VT_
PRotein Ontology [49]	pr	http://purl.obolibrary.org/obo/PR_
Chemical Entities of Biological Interest [50]	chebio	http://purl.obolibrary.org/obo/CHEBI_
Foundational Model of Anatomy Ontology [51]	fma	http://purl.org/sig/ont/fma/fma

5

The AOP-DB RDF: Applying FAIR Principles to the Semantic Integration of AOP Data Using the Research Description Framework

Adapted from: Holly M. Mortensen et al. "The AOP-DB RDF: Applying FAIR Principles to the Semantic Integration of AOP Data Using the Research Description Framework". *Frontiers in Toxicology* 4 (Feb. 2022), pp. 1–6. DOI: 10.3389/ftox.2022.803983.

Abstract

Computational toxicology is central to the current transformation occurring in toxicology and chemical risk assessment. There is a need for more efficient use of existing data to characterize human toxicological response data for environmental chemicals in the US and Europe. The Adverse Outcome Pathway (AOP) framework helps to organize existing mechanistic information and contributes to what is currently being described as New Approach Methodologies (NAMs). AOP knowledge and data are currently submitted directly by users and stored in the AOP-Wiki (aopwiki.org). Automatic and systematic parsing of AOP-Wiki data is challenging, so we have created the EPA Adverse Outcome Pathway Database. The AOP-DB, developed by the US EPA to assist in the biological and mechanistic characterization of AOP data, provides a broad, systems-level overview of the biological context of AOPs. Here we describe the recent semantic mapping efforts for the AOP-DB, and how this process facilitates the integration of AOP-DB data with other toxicologically relevant datasets through a use case example.

5.1 Introduction

There is a need for more efficient use of existing data through improved data integration and compatibility of data structures to characterize human toxicological response data for environmental chemicals. Assessors in the US are moving towards the use of existing mechanistic data (*in vitro* and *in silico*) that provide insights into adverse outcomes in humans [1–4], and reduced animal testing [5]. The Adverse Outcome Pathway (AOP) framework helps to organize existing mechanistic information and contributes to what is currently being described as New Approach Methodologies (NAMs) [6]. The US EPA Adverse Outcome Pathway-Database (AOP-DB) is a decision support tool for risk assessors, developed by the EPA’s Center for Public Health and Environmental Assessment, which contributes to NAMs (e.g., computational toxicology tools) used for the Toxic Substances Control Act (Public Law 114–182, 2016). The AOP-DB has been made available through the Office of Science Management as a public EPA database since November 2021. Pertinent AOP-DB data is currently integrated with the CompTox Chemicals Dashboard (comptox.epa.gov/dashboard/chemical_lists/AOPSTRESSORS), which maps the Distributed Structure-Searchable Toxicity records to the most current list of AOP-DB stressors.

The AOP-DB integrates AOP content to help users characterize AOPs from the OECD-funded AOP-KB (aopkb.oecd.org/index.html) effort, where the AOP-Wiki (aopwiki.org) is the primary repository for direct user submission of AOP information to the AOP-KB. Because the AOP-Wiki data is challenging to parse in its current format [7, 8], the AOP-DB was developed to assist in automating and organizing AOP data, as well as integrating with publicly available datasets to allow biological and mechanistic characterization of AOPs and provide a systems-level overview of the biological context of AOPs [9, 10]. Recent updates to AOP-DB in version 2 [11, 12] include 280 AOPs (1,111 KEs) from the AOP-Wiki XML. The semantic mapping of AOP-DB data, described herein, extends AOP capabilities to users through the incorporation of the Research Description Framework (RDF), which

creates additional ontological linkages and improves capabilities for computational analyses (Figure 5.1). These tools are useful to AOP users trying to retrieve information for AOP development or to understand and characterize existing AOPs. Here we describe the recent semantic mapping efforts for the AOP-DB, and how this process integrates AOP-DB data with other toxicologically relevant datasets.

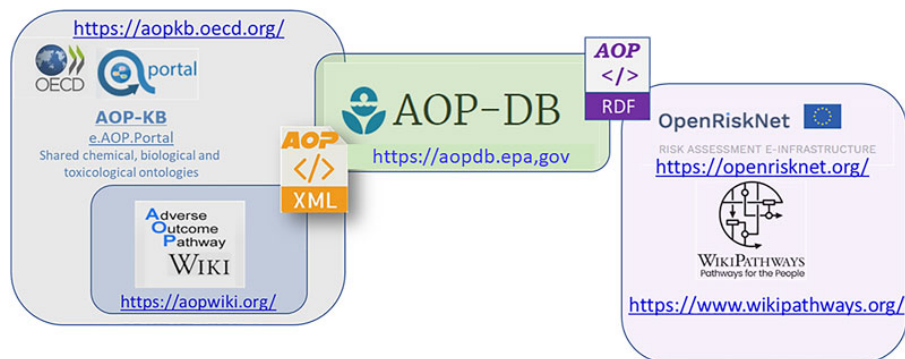


Figure 5.1: The OECD funded AOP-KB currently support the AOP-Wiki. The EPA AOP-DB, currently slated as a third-party tool for integration with the AOPKB 2.0, automatically and programmatically pulls AOP data from the AOP-KB XML, and extends AOP capabilities to users with semantic resources like WikiPathways and the OpenRiskNet e-infrastructure that incorporate the Research Description Framework (RDF). Integration of data across the AOP-KB (AOP-Wiki), AOP-DB, and expanding research frameworks through WikiPathways and the EU funded OpenRiskNet, creates additional ontological linkages and improves capabilities for computational analyses. These tools are useful to AOP users trying to retrieve information for AOP development, as well as those trying to understand and characterize existing AOPs.

As part of OpenRiskNet, a 3 years project supported by the European Commission within Horizon2020 EINFRA-22-2016 Programme, the US EPA AOP-DB was selected as an Implementation Challenge winner. The Implementation Challenge was created to select external tools for use in risk assessment to be prioritized for integration in the OpenRiskNet e-Infrastructure (openrisknet.org) and foster collaborative interaction between project partners. In contribution to this effort, US EPA and Maastricht University project partners have completed the semantic mapping of several AOP-DB data tables into RDF, which is a

standard model for data interchange [13]. The application of RDF defines relationships between data objects using triplestores that include three positional statements (subject, predicate and object). The mapping of AOP-DB data to the RDF data model stores relevant AOP information in a computer-readable format, and contributes to the identification, disambiguation, and meaningful linkage of AOP data with other data structures, following FAIR (findable, accessible, interoperable, and reusable) principles [14, 15].

5.2 Materials and Methods

We selected seven AOP-DB data tables for semantic integration, specifically the Gene Interaction, Biological Pathway, Toxcast Assay, Taxonomy, Chemical-Gene, Gene Info, and Key Event tables. In developing the AOP-DB RDF, we implemented the most recent version of the SQL AOP-DB [16] to map each table of interest into RDF triples. Each table was filtered using the R version 3.6 and Rstudio version 1.2.83 (R Core Team, 2020) to include only records involving a molecular initiating event (MIE) or key event (KE) that maps to a molecular identifier (e.g., gene, protein, cytokine). Code was developed to implement each record as input, modify and filter the AOP-DB table data, and output each modified record to an RDF triple. Additionally, subjects were created for Ensembl and UniProt identifiers. Ontology terms were referenced using BioPortal [17] in order to find the most appropriate ontology terms for each entity, in line with the AOP-Wiki RDF [18] for optimal interoperability between the two resources. Terms were selected with the most accurate description from ontologies that are relevant to the context of the field. For the development of the AOP-DB RDF, several ontologies and consistent vocabularies have been included. Furthermore, publicly available datasets included in the AOP-DB for RDF mapping are described in detail in Mortensen et al. (2021). Table 5.1 provides an overview of the included ontologies and database links, including their prefix in the RDF and their corresponding Internationalized Resource Identifier (IRI).

5.2.1 Testing the AOP-DB RDF

Using a Jupyter notebook (Jupyterlab version 3.2.5, Python version 3.8.5), the AOP-DB SPARQL endpoint has been tested by executing SPARQL queries, using the SPARQLWrapper Python library (version 1.8.5). SPARQL queries were used to extract statistics of the data, and a federated SPARQL query was constructed to explore the integrative capabilities of the AOP-DB RDF. The Jupyter notebook, SPARQL queries for extracting data counts, and instructions for setting up the AOP-DB SPARQL endpoint are available on github.com/BigCAT-UM/AOP-DB-RDF.

5.3 Results

5.3.1 The AOP-DB Semantic Mapping

The AOP-DB RDF schema developed according to the methods described above resulted in the primary and secondary table structure, as illustrated in Figure 5.2. The AOP-DB extends AOP-Wiki RDF with the inclusion of gene/protein, chemical, ToxCast, and biological pathway and taxonomy information. In total, the RDF contains 157 kEs, 376 NCBI genes linked to KEs, 93,449 Chemical-Gene Interactions (3,982 unique chemicals and 122 unique genes), 763,446 Protein-Protein Interactions, 1,143 ToxCast Assays 110,833 Biological Pathways from 10 sources, and 22 taxonomies. Also, the NCBI Gene IDs were matched to 299 Ensembl IDs and 1,026 UniProt IDs. The AOP-DB RDF data tables associate the gene and protein information of AOP genes to chemical, pathway, and assay information organized within the AOP-DB (Mortensen, 2020; Mortensen, 2021).

The Key Event subjects are linked to NCBI Genes through the 'data_1,027' term of the EDAM ontology, which in turn is linked to pathways and assays with respectively the terms 'pw:0000001' from the Pathway Ontology and 'mmo:0000441' from the Measurement Method Ontology. Furthermore, matching identifiers were linked with 'skos:exactMatch', providing IRIs of Ensembl IDs, HGNC Symbols, and UniProt IDs. On the other hand, Chemical-Gene

Table 5.1: Overview of ontologies, consistent vocabularies and databases included in the AOP-DB RDF.

Ontologies and Vocabularies		
Name	Prefix in RDF	IRI
AOP Ontology [19]	aopo	http://aopkb.org/aop_ontology
BioAssay Ontology [20]	bao	http://www.bioassayontology.org/bao
Chemical Information Ontology [21]	cheminf	http://semanticscience.org/resource/CHEMINF
Dublin Core	dc	http://purl.org/dc/elements/1.1
EDAM Ontology [22]	edam	http://edamontology.org/
Friend Of A Friend	foaf	http://xmlns.com/foaf/0.1
Logical Observation Identifier Names and Codes [23]	loinc	http://purl.bioontology.org/ontology/LNC
Molecular Interactions [24]	mi	http://purl.obolibrary.org/obo/MI
Measurement Method Ontology [25]	mmo	http://purl.obolibrary.org/obo/MMO
NCBI Taxonomy [26]	ncbitaxon	http://purl.bioontology.org/ontology/NCBITAXON
Pathway Ontology [27]	pw	http://purl.obolibrary.org/obo/PW
RDF Schema	rdfs	http://www.w3.org/2000/01/rdf-schema
Semantics Science Ontology [28]	sio	http://semanticscience.org/resource
Simple Knowledge Organization System	skos	http://www.w3.org/2004/02/skos/core
Uber Anatomy Ontology [29]	uberon	http://purl.obolibrary.org/obo/UBERON
Databases		
AOP-Wiki	aop.events	http://identifiers.org/aop.events
Comptox Dashboard [30]	assay	https://comptox.epa.gov/dashboard/assay_endpoints
CAS Common Chemistry	cas	https://identifiers.org/cas
Ensembl [31]	ensembl	http://identifiers.org/ensembl
HUGO Genome Nomenclature Committee [32]	hgnc	https://identifiers.org/hgnc
NCBI Gene	ncbigene	https://identifiers.org/ncbigene
Uniprot [33]	uniprot	https://identifiers.org/uniprot
KEGG Pathways [34]	kegg.pathway	https://identifiers.org/kegg.pathway
PharmGKB Pathways [35]	pharmgkb.pathways	https://identifiers.org/pharmgkb.pathways
Small Molecule Pathway Database [36]	smpdb	https://identifiers.org/smpdb
BioCyc [37]	biocyc	https://identifiers.org/biocyc
BioCarta Pathways	biocarta.pathway	https://identifiers.org/biocarta.pathway
Reactome [38]	reactome	https://identifiers.org/reactome
NCI Pathway Interaction Database [39]	pid.pathway	https://identifiers.org/pid.pathway
NetPath [40]	netpath	http://netpath.org/pathways?path_id=
WikiPathways [18]	wikipathways	https://identifiers.org/wikipathways
AOP-DB Chemical-Gene association	chemicalgeneassociation	http://example.org/ChemicalGeneAssociation
AOP-DB Protein Interaction	proteinInteraction	http://example.org/proteinInteraction

interactions, Protein-protein interactions, ToxCast assays, and Pathways have links to NCBI Gene subjects through the term ‘data_1,027’ from the EDAM ontology. Finally, taxonomy is referenced by ToxCast assay and pathway subjects through the term ‘ncbitaxon:131,567’ indicating cellular organism.

5.3.2 The AOP-DB SPARQL Endpoint

The AOP-DB RDF can be explored through the AOP-DB SPARQL (aopdb.rdf.bigcat-bioinformatics.org/sparql). It allows custom SPARQL queries to return output tables in a variety of formats, where it is possible to directly combine different resources with federated SPARQL queries.

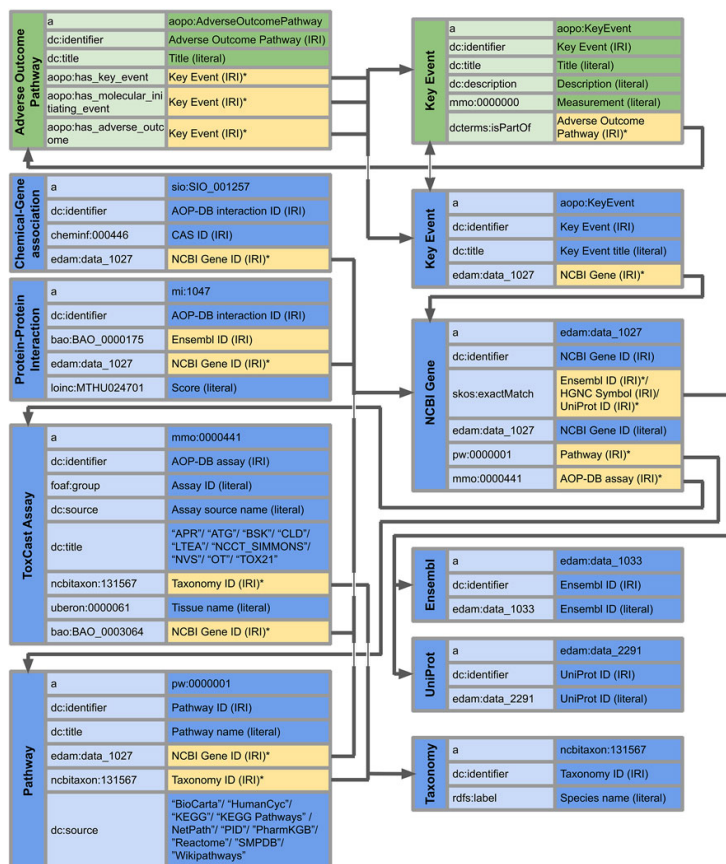


Figure 5.2: AOP-DB Semantic Mapping. Semantic mapping illustrating the predicates and objects of the nine core subject types in the AOP-DB RDF (in blue). Vertical columns show subjects, and the middle and right columns indicate predicates and objects, respectively. Where applicable, the type of entry is indicated (literal or IRI). Yellow objects with an asterisk (*) indicate the connection between their subjects and the subjects of other tables. The interaction with the AOP-Wiki RDF is highlighted at the Key Events and Adverse Outcome Pathways (in green). Forward slashes indicate the inclusion of multiple objects as part of the subject-predicate-object triple.

5.3.3 AOP-DB RDF Use Case Example

SPARQL queries can be used to query the RDF in order to answer biological and toxicological questions, such as which molecular targets (e.g. genes/proteins), chemical stressors, key events, or *in vitro* as-

says are relevant for adverse outcomes of interest. The use case examples provided herein illustrate the utility of the AOP-DB RDF content, as well as the power of integrating these data with other diverse, external databases using federated queries. Our first use case implements the AOP-DB RDF to identify AOP-relevant molecular targets that have associated ToxCast assay targets, which has previously not been possible. The automated linkage of ToxCast assays and KEs in AOP-Wiki can serve as a prioritization tool by exploring the activation of KEs by the many chemicals that have been investigated in ToxCast. The second use case shows the integration of the AOP-DB RDF with other databases that provide access to their data through SPARQL endpoints. A single SPARQL query can be executed to extract AOP IDs, KE IDs, KE titles and protein names from the AOP-Wiki RDF, extract protein descriptions from the Protein Ontology, and the names and descriptions of pathways in WikiPathways, all based on the NCBI Gene IDs captured in the AOP-DB. Through the integration of these diverse data sources, we can effectively explore the data and build automated computational workflows to address questions of toxicological concern.

5.4 Conclusion

A central goal of computational toxicology is to predict and explain how the human body responds after exposure to specific xenobiotics or other chemicals *in silico*. This effort has been hampered by several major limiting factors, including fragmented and poorly structured data, and insufficient access to computational resources and expertise. The AOP-DB RDF and SPARQL endpoint created and discussed herein allow improved access to rigorously structured AOP data and other associated data of toxicological interest. This work improves computational organization and efficiency, through improved data integration, for toxicological and related datasets, and contributes to continued progress in computational toxicology, chemical screening and the improvement of human health risk assessment.

The AOP-DB RDF will be improved with regular data updates and continued data integration with relevant datasets. Future work includes semantic integration of AOP-DB disease-gene data, tissue-specific gene interaction networks, AOP functional single nucleotide polymorphism (SNP) and population SNP frequency information and chemical-specific datasets.

References

- [1] National Research Council (NRC). *Toxicity Testing in the 21st Century: A Vision and a Strategy*. 2007.
- [2] National Research Council (NRC). *Using 21st Century Science to Improve Risk-Related Evaluations*. Washington, D.C.: National Academies Press, Feb. 2017. DOI: 10.17226/24635.
- [3] National Research Council (NRC). *Science and Decisions: Advancing Risk Assessment*. 2009.
- [4] National Research Council (NRC). *Toxicity-Pathway-Based Risk Assessment: Preparing for Paradigm Change*. 2010.
- [5] Andrew R. Wheeler. *Directive to prioritize efforts to reduce animal testing*. 2019.
- [6] Russell S. Thomas et al. "The next generation blueprint of computational toxicology at the U.S. Environmental protection agency". *Toxicological Sciences* 169.2 (June 2019), pp. 317–332. DOI: 10.1093/TOXSCI/KFZ058.
- [7] Cataia Ives et al. "Creating a Structured AOP Knowledgebase via Ontology-Based Annotations." *Applied in vitro toxicology* 3.4 (Dec. 2017), pp. 298–311. DOI: 10.1089/aivt.2017.0017.
- [8] Marvin Martens et al. "Introducing WikiPathways as a Data-Source to Support Adverse Outcome Pathways for Regulatory Risk Assessment of Chemicals and Nanomaterials". *Frontiers in Genetics* 9 (Dec. 2018), p. 661. DOI: 10.3389/fgene.2018.00661.
- [9] Holly M. Mortensen et al. "Leveraging human genetic and adverse outcome pathway (AOP) data to inform susceptibility in human health risk assessment". *Mammalian Genome* 29.1-2 (Feb. 2018), pp. 190–204. DOI: 10.1007/s00335-018-9738-7.
- [10] Maureen E. Pittman et al. "AOP-DB: A database resource for the exploration of Adverse Outcome Pathways through integrated association networks". *Toxicology and Applied Pharmacology* 343 (Mar. 2018), pp. 71–83. DOI: 10.1016/j.taap.2018.02.006.
- [11] Holly M. Mortensen et al. "The 2021 update of the EPA's adverse outcome pathway database". *Scientific Data* 8.1 (Dec. 2021), p. 169. DOI: 10.1038/s41597-021-00962-3.
- [12] Holly M Mortensen. "The EPA Adverse Outcome Pathway Database version 2.0 (AOP-DB_v2)". *US EPA Office of Research and Development (ORD)* (2021).
- [13] Richard Cyganiak, David Wood, and Markus Lanthaler. *RDF 1.1 Concepts and Abstract Syntax*. 2014.

-
- [14] Mark D. Wilkinson et al. "Evaluating FAIR maturity through a scalable, automated, community-governed framework". *Scientific Data* 6.1 (Dec. 2019). DOI: 10.1038/S41597-019-0184-5.
- [15] Mark D. Wilkinson et al. "Addendum: The FAIR Guiding Principles for scientific data management and stewardship". *Scientific data* 6.1 (Mar. 2019), p. 6. DOI: 10.1038/S41597-019-0009-6.
- [16] Holly M Mortensen et al. "Enhancing the EPA Adverse Outcome Pathway Database (AOP-DB): Recent Updates and Semantic Integration". *The Toxicologist* 174.1 (2020).
- [17] Patricia L. Whetzel et al. "BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications." *Nucleic Acids Research* 39.suppl_2 (July 2011), W541–W545. DOI: 10.1093/nar/gkr469.
- [18] Marvin Martens et al. "WikiPathways: connecting communities". *Nucleic Acids Research* 49.D1 (Jan. 2021), pp. D613–D621. DOI: 10.1093/nar/gkaa1024.
- [19] Lyle D. Burgoon. "The AOPontology: A semantic artificial intelligence tool for predictive toxicology". *Applied In Vitro Toxicology* 3.3 (Sept. 2017), pp. 278–281. DOI: 10.1089/aivt.2017.0012.
- [20] Saminda Abeyruwan et al. "Evolving BioAssay Ontology (BAO): Modularization, integration and applications". *Journal of Biomedical Semantics* 5 (2014). DOI: 10.1186/2041-1480-5-S1-S5.
- [21] Janna Hastings et al. "The Chemical Information Ontology: Provenance and Disambiguation for Chemical Data on the Biological Semantic Web". *PLoS ONE* 6.10 (Oct. 2011). Ed. by Franca Fraternali, e25513. DOI: 10.1371/journal.pone.0025513.
- [22] Jon Ison et al. "EDAM: An ontology of bioinformatics operations, types of data and identifiers, topics and formats". *Bioinformatics* 29.10 (May 2013), pp. 1325–1332. DOI: 10.1093/bioinformatics/btt113.
- [23] Clement J. McDonald et al. "LOINC, a universal standard for identifying laboratory observations: A 5-year update". *Clinical Chemistry* 49.4 (Apr. 2003), pp. 624–633. DOI: 10.1373/49.4.624.
- [24] P P Millán. *Molecular Interactions Controlled Vocabulary*. 2020.
- [25] Jennifer R. Smith et al. "The clinical measurement, measurement method and experimental condition ontologies: Expansion, improvements and new applications". *Journal of Biomedical Semantics* 4.1 (Oct. 2013), p. 26. DOI: 10.1186/2041-1480-4-26.
- [26] Olivier Bodenreider. "The Unified Medical Language System (UMLS): Integrating biomedical terminology". *Nucleic Acids Research* 32.DATABASE ISS. (Jan. 2004). DOI: 10.1093/NAR/GKH061.
- [27] Victoria Petri et al. "The pathway ontology - updates and applications". *Journal of Biomedical Semantics* 5.1 (Feb. 2014), pp. 1–12. DOI: 10.1186/2041-1480-5-7.
- [28] Michel Dumontier et al. "The semanticscience integrated ontology (SIO) for biomedical research and knowledge discovery". *Journal of Biomedical Semantics* 5.1 (Mar. 2014). DOI: 10.1186/2041-1480-5-14.
- [29] Christopher J. Mungall et al. "Uberon, an integrative multi-species anatomy ontology". *Genome Biology* 13.1 (Jan. 2012), R5. DOI: 10.1186/gb-2012-13-1-r5.

- [30] Antony J. Williams et al. "The CompTox Chemistry Dashboard: A community data resource for environmental chemistry". *Journal of Cheminformatics* 9.1 (Nov. 2017), p. 61. DOI: 10.1186/s13321-017-0247-6.
- [31] Andrew D. Yates et al. "Ensembl 2020". *Nucleic Acids Research* 48.D1 (Jan. 2020), pp. D682–D688. DOI: 10.1093/NAR/GKZ966.
- [32] Bryony Braschi et al. "Genenames.org: the HGNC and VGNC resources in 2019". *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D786–D792. DOI: 10.1093/nar/gky930.
- [33] Alex Bateman and UniProt Consortium. "UniProt: a worldwide hub of protein knowledge". *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D506–D515. DOI: 10.1093/nar/gky1049.
- [34] Minoru Kanehisa et al. "KEGG: Integrating viruses and cellular organisms". *Nucleic Acids Research* 49.D1 (Jan. 2021), pp. D545–D551. DOI: 10.1093/NAR/GKAA970.
- [35] Michelle Whirl-Carrillo et al. "An Evidence-Based Framework for Evaluating Pharmacogenomics Knowledge for Personalized Medicine". *Clinical Pharmacology and Therapeutics* 110.3 (Sept. 2021), pp. 563–572. DOI: 10.1002/CPT.2350.
- [36] Timothy Jewison et al. "SMPDB 2.0: Big improvements to the small molecule pathway database". *Nucleic Acids Research* 42.D1 (Jan. 2014). DOI: 10.1093/NAR/GKT1067.
- [37] Peter D. Karp et al. "The BioCyc collection of microbial genomes and metabolic pathways". *Briefings in Bioinformatics* 20.4 (Mar. 2018), pp. 1085–1093. DOI: 10.1093/BIB/BBX085.
- [38] Bijay Jassal et al. "The reactome pathway knowledgebase". *Nucleic Acids Research* 48.D1 (Jan. 2020), pp. D498–D503. DOI: 10.1093/NAR/GKZ1031.
- [39] Carl F. Schaefer et al. "PID: The pathway interaction database". *Nucleic Acids Research* 37.SUPPL. 1 (2009). DOI: 10.1093/NAR/GKN653.
- [40] Kumaran Kandasamy et al. "NetPath: a public resource of curated signal transduction pathways". *Genome Biology* 11.1 (Jan. 2010), R3. DOI: 10.1186/gb-2010-11-1-r3.

6

AOPLink Jupyter Notebook: Extracting and analysing data related to an AOP of interest

6.1 Introduction

In order to drive the transition toward *in vitro* and *in silico* methods for risk assessment of chemicals and nanomaterials, the project OpenRiskNet has aimed to develop virtual research environments for integrating and analysing data, creating simulations and deploy computational models. These developments were driven by case studies [1], demonstrating the capabilities to end-users of such a platform, including risk assessors, regulators, and fellow researchers, aligned with the SEURAT-9 risk assessment framework [2]. One of these case studies was AOPLink (openrisknet.org/e-infrastructure/development/case-studies/case-study-aoplink), which explored Adverse Outcome Pathways (AOPs) and supporting experimental data. In general, AOPs contain descriptions of mechanistic knowledge of toxicological processes and are based on scientific literature [3]. However, the use of AOPs for regulatory purposes also requires qualitative and quantitative validation [4], which can be found in literature and databases. The main goal of the case study is to establish links between AOPs in the AOP-Wiki and supporting experimental data, allowing the identification of AOPs based on data, and finding data related to the AOP of interest. This case study has resulted in a Jupyter notebook, combining the functionalities of a set of services and repositories that were developed or implemented during the project, including the AOP-Wiki SPARQL endpoint [5], AOP-DB SPARQL endpoint [6], BridgeDb webservice [7], EdelweissData Explorer for ToxCast and TG-GATEs data, ChemIdConverter, and WikiPathways SPARQL endpoint [8, 9]. This Jupyter notebook represents a computational workflow that starts with the selection of an AOP of interest, explores related information, and looks up supporting assay data in ToxCast and transcriptomic data in TG-GATEs, followed by pathway analysis of transcriptomics data that was retrieved. This Jupyter notebook is available on github.com/OpenRiskNet/notebooks/blob/master/AOPLink/ExtractingandanalysingdatarelatedtoanAOPofinterest.

`ipy nb`. However, some of the services have been relocated or are not functional anymore since the creation of the Jupyter notebook.

6.2 Jupyter notebook

Extracting and analysing data related to an AOP of interest

Citation: Marvin Martens, Thomas Exner, Tomaž Mohorič, Chris T Evelo, Egon L Willighagen. Workflow for extracting and analyzing data related to an AOP of interest. 2020

One of the main questions to solve in AOPLink is the finding of data that supports an AOP of interest. To answer that, we have developed this Jupyter notebook that does that by using a variety of OpenRiskNet services:

- AOP-Wiki RDF
- ChemIdConvert
- AOP-DB RDF
- BridgeDb
- EdelweissData explorer
- WikiPathways

After selecting an AOP of interest, information is extracted from the AOP-Wiki RDF, ChemIdConvert, and AOP-DB RDF, to get a better understanding of the AOP. Next, the EdelweissData explorer was used to search for fitting data sets from ToxCast and TG-GATES based on the genes and compounds linked to the AOP of interest. The final part involves pathway analysis using the transcriptomics data of TG-GATES and the molecular pathways of WikiPathways, identifying significantly affected pathways upon exposure to the chemicals of interest.

In order to execute the Jupyter notebook, a set of Python libraries are required. The following section should import, or install, all of them.

```
[1]: import sys

!{sys.executable} -m pip install --upgrade pip
!{sys.executable} -m pip install watermark
```

```
try:
    from SPARQLWrapper import SPARQLWrapper, JSON
except ImportError:
    !{sys.executable} -m pip install sparqlwrapper
    from SPARQLWrapper import SPARQLWrapper, JSON

try:
    from pyvis.network import Network
except ImportError:
    !{sys.executable} -m pip install pyvis
    from pyvis.network import Network

try:
    from IPython.display import display, HTML, IFrame
except ImportError:
    !{sys.executable} -m pip install ipython
    from IPython.display import display, HTML, IFrame

try:
    import urllib
except ImportError:
    !{sys.executable} -m pip install urllib
    import urllib

try:
    import simplejson as json
except ImportError:
    !{sys.executable} -m pip install simplejson
    import simplejson as json

try:
    import pandas as pd
except ImportError:
    !{sys.executable} -m pip install pandas
    import pandas as pd

try:
    import re
except ImportError:
    !{sys.executable} -m pip install re
    import re
```

```
try:
    import requests
except ImportError:
    !{sys.executable} -m pip install requests
    import requests

try:
    import warnings
except ImportError:
    !{sys.executable} -m pip install warnings
    import warnings

try:
    import statistics
except ImportError:
    !{sys.executable} -m pip install statistics
    import statistics

try:
    from edelweiss_data import API, QueryExpression as Q
except ImportError:
    !{sys.executable} -m pip install edelweiss_data
    from edelweiss_data import API, QueryExpression as Q

pd.set_option('display.max_colwidth', -1)
warnings.filterwarnings("ignore", category=UserWarning)
```

Define the AOP of interest

This Jupyter notebook focuses on the AOP of interest, which is based on the identifier of the AOP. The identifiers of AOPs can be found on the AOP-Wiki website.

```
[2]: AOPid = "37"
```

Set service URLs

The notebook uses a variety of external services. To keep an overview of these, their URLs are defined at the start of the notebook.

```
[3]: # SPARQL endpoint URLs
aopwikisparql = SPARQLWrapper("http://aopwiki-rdf.prod.
↳openrisknet.org/sparql/")
aopdbsparql = SPARQLWrapper("http://aopdb-rdf.prod.
↳openrisknet.org/sparql/")
wikipathwayssparql = SPARQLWrapper("http://sparql.
↳wikipathways.org/sparql/")

# ChemIdConvert URL
chemidconvert = 'https://chemidconvert.cloud.douglasconnect.
↳com/v1/'

# BridgeDB base URL
bridgedb = 'http://bridgedb.prod.openrisknet.org/'

# EdelweissData API URL
edelweiss_api_url = 'https://api.staging.kit.cloud.
↳douglasconnect.com'
```

AOP-Wiki RDF

Service description

The AOP-Wiki repository is part of the AOP Knowledge Base (AOP-KB), a joint effort of the US-Environmental Protection Agency and European Commission - Joint Research Centre. It is developed to facilitate collaborative AOP development, storage of AOPs, and therefore allow reusing toxicological knowledge for risk assessors. This Case Study has converted the AOP-Wiki XML data into an RDF schema, which has been exposed in a public SPARQL endpoint in the OpenRiskNet e-infrastructure.

Implementation

First, general information of the AOP is fetched using a variety of SPARQL queries, using predicates from the AOP-Wiki RDF schema. This is used for: - Creating an overview table of the AOP of interest - Extending the AOP network with connected AOPs

Second, stressor chemicals are retrieved and stored for further analysis and fetching of data.

Creating overview table

```
[4]: #Define all variables as ontology terms present in AOP-Wiki
      ↪RDF
title = 'dc:title'
webpage = 'foaf:page'
creator = 'dc:creator'
abstract = 'dcterms:abstract'
key_event = 'aopo:has_key_event'
molecular_initiating_event = 'aopo:
      ↪has_molecular_initiating_event'
adverse_outcome = 'aopo:has_adverse_outcome'
key_event_relationship = 'aopo:has_key_event_relationship'
stressor = 'ncit:C54571'

#Create the list of all terms of interest
listofters = [title, webpage, creator, abstract, key_event,
      ↪molecular_initiating_event, adverse_outcome,
      ↪key_event_relationship, stressor]

#Initiate the DataFrame
AOPinfo = pd.DataFrame(columns=['Properties'], index =
      ↪[list(listofters)])

#Query all terms of interest in the selected AOP
for term in listofters:
    sparqlquery = '''
        PREFIX ncit: <http://ncicb.nci.nih.gov/xml/owl/EVS/
      ↪Thesaurus.owl#>
        SELECT (group_concat(distinct ?item;separator=";") as ?
      ↪items)
        WHERE{
            ?AOP_URI a aopo:AdverseOutcomePathway;''' + term + ''' ?item.
            FILTER (?AOP_URI = aop:''' + AOPid + ''')
        }
        '''
    aopwikisparql.setQuery(sparqlquery)
    aopwikisparql.setReturnFormat(JSON)
    results = aopwikisparql.query().convert()
```

```

for result in results["results"]["bindings"]:
    if 'identifiers.org' in result["items"]["value"]:
        AOPinfo.at[term, 'Properties'] = ', '
        ↪join(result["items"]["value"].split(';'))
    else:
        AOPinfo.at[term, 'Properties'] = _
        ↪result["items"]["value"]
display(AOPinfo)

```

	Properties
dct:title	PPARAlpha-dependent liver cancer
foaf:page	http://identifiers.org/aop/37
dc:creator	J. Christopher Corton, Cancer AOP Workgroup. National Health and Environmental Effects Research Laboratory, Office of Research and Development, Integrated Systems Toxicology Division, US Environmental Protection Agency, Research Triangle Park, NC. Corresponding author for wiki entry (corton.chris@epa.gov)\n
dcterms:abstract	Several therapeutic agents ... overlapping dose levels.\n
aopo:has_key_event	http://identifiers.org/aop.events/1170 , http://identifiers.org/aop.events/1171 , http://identifiers.org/aop.events/227 , http://identifiers.org/aop.events/716 , http://identifiers.org/aop.events/719 http://identifiers.org/aop.events/227 http://identifiers.org/aop.events/719
aopo:has_molecular_initiating_event	http://identifiers.org/aop.relationships/1229 , http://identifiers.org/aop.relationships/1230 , http://identifiers.org/aop.relationships/1232 , http://identifiers.org/aop.relationships/1239
aopo:has_adverse_outcome	http://identifiers.org/aop.stressor/11 , http://identifiers.org/aop.stressor/175 , http://identifiers.org/aop.stressor/191 , http://identifiers.org/aop.stressor/205 , http://identifiers.org/aop.stressor/206 , http://identifiers.org/aop.stressor/207 , http://identifiers.org/aop.stressor/208 , http://identifiers.org/aop.stressor/210 , http://identifiers.org/aop.stressor/211
aopo:has_key_event_relationship	
ncit:C54571	

6.2.1 Generating AOP network

```

[5]: #Generate network from AOP + interlinked AOPs.
Key_Events = str(AOPinfo.iat[4, 0]).split(', ')

#From all the KEs from the selected, get the other AOPs in_
↪which they are present, and find all KEs of those other_
↪AOPs
for Key_Event in Key_Events:

```

```

sparqlquery = '''
SELECT ?MIE_ID ?KE_ID ?AO_ID ?KER_ID ?KE_Title
WHERE{
  ?KE_URI a aopo:KeyEvent; dcterms:isPartOf ?AOP_URI.
  ?AOP_URI aopo:has_key_event ?KE_URI2; aopo:
↳has_molecular_initiating_event ?MIE_URI; aopo:
↳has_adverse_outcome ?AO_URI; aopo:
↳has_key_event_relationship ?KER_URI.
  ?KE_URI2 rdfs:label ?KE_ID; dc:title ?KE_Title.
  ?MIE_URI rdfs:label ?MIE_ID.
  ?AO_URI rdfs:label ?AO_ID.
  ?KER_URI rdfs:label ?KER_ID.
  FILTER (?KE_URI = <''' +Key_Event+'''>)}
'''

aopwikisparql.setQuery(sparqlquery)
aopwikisparql.setReturnFormat(JSON)
results = aopwikisparql.query().convert()

MIEs = set([])
KEs = set([])
KETitle = {}
AOs = set([])
KERs = set([])
for result in results["results"]["bindings"]:
    MIEs.add(result["MIE_ID"]["value"])
    AOs.add(result["AO_ID"]["value"])
    KEs.add(result["KE_ID"]["value"])
    KERs.add(result["KER_ID"]["value"])
    KETitle[result["KE_ID"]["value"]] = _
↳result["KE_Title"]["value"]

#List all intermediate KEs that are not MIEs or AOs
KEsIntermediate = []
for item in KEs:
    if item not in MIEs and item not in AOs:
        KEsIntermediate.append(item)

#Initiate network figure
net= Network(height="100%", width="100%")

#Add nodes for Molecular Initiating Events, Key Events, and _
↳Adverse Outcomes to the network figure

```

```

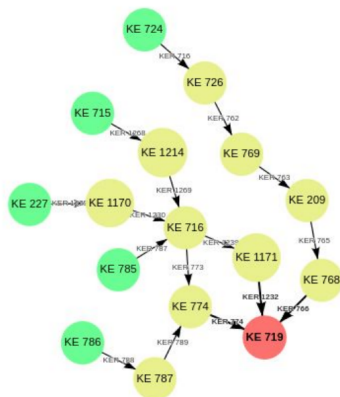
for MIE in MIEs:
    net.add_node(MIE, color = 'lightgreen', size = 50, shape =
↳ 'circle', font = '20px arial black', title =
↳ KEtittle[MIE])
for KE in KEsIntermediate:
    net.add_node(KE, color = 'khaki', size = 50, shape =
↳ 'circle', font = '20px arial black', title = KEtittle[KE])
for AO in AOs:
    net.add_node(AO, color = 'salmon', size = 50, shape =
↳ 'circle', font = '20px arial black', title = KEtittle[AO])

#Add all Key Event Relationships to the network figure after
↳ querying all KERs for all KEs in AOP-Wiki RDF
for KER in KERs:
    sparqlquery = '''
    SELECT ?KE_UP_ID ?KE_DOWN_ID
    WHERE{
        ?KER_URI a aopo:KeyEventRelationship; rdfs:label ?KER_ID;
↳ aopo:has_upstream_key_event ?KE_UP_URI; aopo:
↳ has_downstream_key_event ?KE_DOWN_URI.
        ?KE_UP_URI rdfs:label ?KE_UP_ID.
        ?KE_DOWN_URI rdfs:label ?KE_DOWN_ID.
        FILTER (?KER_ID = ''' + KER + ''')
    }
    '''
    aopwikisparql.setQuery(sparqlquery)
    aopwikisparql.setReturnFormat(JSON)
    results = aopwikisparql.query().convert()
    for result in results["results"]["bindings"]:
        net.add_edge(result["KE_UP_ID"]["value"],
↳ result["KE_DOWN_ID"]["value"], width = 2, color =
↳ 'black', label = KER, arrows = 'to')

net.show('mygraph.html')
IFrame(src='./mygraph.html', width=700, height=600)

```

[5]:



6.2.2 Query all chemicals that are part of the selected AOP

```
[6]: sparqlquery = '''
PREFIX ncit: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.
      ↪owl#>
SELECT ?CAS_ID (fn:substring(?CompTox,33) as ?CompTox_ID) ?
      ↪Chemical_name
WHERE{
?AOP_URI a aopo:AdverseOutcomePathway; ncit:C54571 ?Stressor.
?Stressor aopo:has_chemical_entity ?Chemical.
?Chemical cheminf:CHEMINF_000446 ?CAS_ID; dc:title ?
      ↪Chemical_name.
OPTIONAL {?Chemical cheminf:CHEMINF_000568 ?CompTox.}
FILTER (?AOP_URI = aop:'''+AOPid +''')
}'''

aopwikisparql.setQuery(sparqlquery)
aopwikisparql.setReturnFormat(JSON)
results = aopwikisparql.query().convert()

Chemical_names = {}
CompTox = {}

for result in results["results"]["bindings"]:
```



```

    try: CompTox[result["CAS_ID"]["value"]]
    ↪=result["CompTox_ID"]["value"]
    except: pass
for result in results["results"]["bindings"]:
    try: Chemical_names[result["CAS_ID"]["value"]]
    ↪=result["Chemical_name"]["value"]
    except: pass

Chemdata = pd.DataFrame(columns=['Chemical_name', 'CAS_ID',
    ↪'CompTox_ID'])
for CAS_ID in Chemical_names:
    Chemdata = Chemdata.append({
        'Chemical_name' : Chemical_names[CAS_ID],
        'CAS_ID'       : CAS_ID,
        'CompTox_ID'   : CompTox[CAS_ID],
    }, ignore_index=True)
display(Chemdata)

```

	Chemical_name	CAS_ID	CompTox_ID
0	Di(2-ethylhexyl) phthalate	117-81-7	DTXSID5020607
1	Gemfibrozil	25812-30-0	DTXSID0020652
2	Nafenopin	3771-19-5	DTXSID8020911
3	Bezafibrate	41859-67-0	DTXSID3029869
4	Fenofibrate	49562-28-9	DTXSID2029874
5	Pirinixic acid	50892-23-4	DTXSID4020290
6	Ciprofibrate	52214-84-3	DTXSID8020331
7	Clofibrate	637-07-0	DTXSID3020336

```

[7]: compounds = []
for index, row in Chemdata.iterrows():
    compounds.append(row['CAS_ID'])
compounds

```

```

[7]: ['117-81-7', '25812-30-0', '3771-19-5', '41859-67-0',
    ↪'49562-28-9', '50892-23-4', '52214-84-3', '637-07-0']

```

ChemIdConvert

Service description

The ChemIdConverter allows users to submit and translate a variety of chemical descriptors, such as SMILES and InChI, through a REST API.

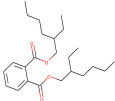
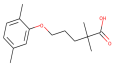
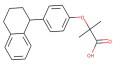
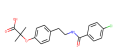
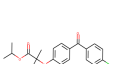
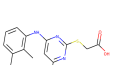
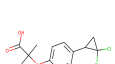
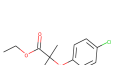
Implementation

Convert selected chemical names and display their chemical structures in a dataframe. It takes CAS IDs as an input, and translates them into Smiles and InChI Keys.

```
[8]: compoundstable = pd.DataFrame(columns=['CAS_ID', 'Image',
→ 'Smiles', 'InChIKey'])
# Fill "compounds" with the "smiles" by the compound name.
for compound in compounds:
    smiles = requests.get(chemidconvert + 'cas/to/smiles',
→ params={'cas': compound}).json()['smiles']
    inchikey = requests.get(chemidconvert + 'smiles/to/
→ inchikey', params={'smiles': smiles}).json()['inchikey']
    compoundstable = compoundstable.append({'CAS_ID':
→ compound, 'Image': smiles, 'Smiles': smiles, 'InChIKey':
→ inchikey}, ignore_index=True)

def smiles_to_image_html(smiles): # "smiles" shadows
→ "smiles" from outer scope, use this function only in
→ "to_html().
    """Gets for each smile the image, in HTML.
    :param smiles: Takes the "smiles" form "compounds".
    :return: The HTML code for the image of the given smiles.
    """
    return ''
# Return a HTML table of "compounds", after "compounds" is
→ fill by "smiles_to_image_html".
HTML(compoundstable.to_html(escape=False,
→ formatters=dict(Image=smiles_to_image_html)))
```

[8] :

CAS_ID	Image	Smiles	InChIKey
0 117-81-7		<chem>CCCCC(CC)COC(=O)c1ccccc1C(=O)OCC(CC)CCCC</chem>	BJQHLLKABXJIVAM-UHFFFAOYSA-N
1 25812-30-0		<chem>Cc1ccc(C)c(OCCCC(C)(C)C(O)=O)c1</chem>	HEMJJKBWTPKOJG-UHFFFAOYSA-N
2 3771-19-5		<chem>CC(C)(Oc1ccc(cc1)C2CCCc3ccccc23)C(O)=O</chem>	XJGBDJOmwKAZJS-UHFFFAOYSA-N
3 41859-67-0		<chem>CC(C)(Oc1ccc(CCNC(=O)c2ccc(Cl)cc2)cc1)C(O)=O</chem>	IIBYAHWJQTYFKB-UHFFFAOYSA-N
4 49562-28-9		<chem>CC(C)OC(=O)C(C)(C)Oc1ccc(cc1)C(=O)c2ccc(Cl)cc2</chem>	YMTINGFKWwxKFG-UHFFFAOYSA-N
5 50892-23-4		<chem>Cc1cccc(Nc2cc(Cl)nc(SCC(O)=O)n2)c1C</chem>	SZRPDCEHvWojX-UHFFFAOYSA-N
6 52214-84-3		<chem>CC(C)(Oc1ccc(cc1)C2CC2(Cl)Cl)C(O)=O</chem>	KPSRODZRAIWAKH-UHFFFAOYSA-N
7 637-07-0		<chem>CCOC(=O)C(C)(C)Oc1ccc(Cl)cc1</chem>	KNHUKKLJHYUCFP-UHFFFAOYSA-N

AOP-DB RDF

Service description

The EPA AOP-DB supports the discovery and development of putative and potential AOPs. Based on public annotations, it integrates AOPs with gene targets, chemicals, diseases, tissues, pathways, species orthology information, ontologies, and gene interactions. The AOP-DB facilitates the translation of AOP biological context, and associates assay, chemical and disease endpoints with AOPs (Pittman et al., 2018; Mortensen et al., 2018). The AOP-DB won the first OpenRiskNet implementation challenge of the associated partner program and is therefore integrated into the OpenRiskNet e-infrastructure. After the conversion of the AOP-DB into an RDF schema, its data will be exposed in a Virtuoso SPARQL endpoint.

Implementation

Extract all genes related to AOP of interest Find all ToxCast assays linked to those genes

```
[9]: Key_Events = str(AOPinfo.iat[4,0]).split(',')
      Genes = []
      #from the KEs, get the AOPs
      for Key_Event in Key_Events:
          sparqlquery = '''
              SELECT DISTINCT ?KE_ID ?Entrez_ID WHERE{
                ?KE_URI edam:data_1027 ?Entrez_URI. ?Entrez_URI edam:
                ↪data_1027 ?Entrez_ID.
                FILTER (?KE_URI = <''' + Key_Event + '''>)}
              '''
          aopdbsparql.setQuery(sparqlquery)
          aopdbsparql.setReturnFormat(JSON)
          results = aopdbsparql.query().convert()
          for result in results["results"]["bindings"]:
              Genes.append(result["Entrez_ID"]["value"])
      print(Genes)
```

```
['5465', '403654', '19013', '25747']
```

BridgeDb to map identifiers

Service description

In order to link databases and services that use particular identifiers for genes, proteins, and chemicals, the BridgeDb platform is integrated into the OpenRiskNet e-infrastructure. It allows for identifier mapping between various biological databases for data integration and interoperability (van Iersel et al., 2010).

Implementation

The genes from AOP-DB are mapped to identifiers from other databases using BridgeDb. Variable values are filled for `inputdatasource` and `outputdatasource` identifiers based on BridgeDb's documentation on system codes. Also, the species is specified as a value in the variable `Species`.

```
[10]: inputdatasource = 'L'
      outputdatasource = ['H', 'En']
      Species = ['Human', 'Dog', 'Mouse', 'Rat']
      Mappings = {}
      HGNC = []

      for source in outputdatasource:
          Mappings[source] = []
          for Entrez in Genes:
              for species in Species:
                  allmappings = re.split('\t|\n', requests.
↳get(bridgedb + species + '/xrefs/' + inputdatasource + '/'
↳' + Entrez + '?dataSource=' + source).text)
                  if allmappings[0] is not '':
                      break
                  Mappings[source].append(allmappings[0])

      ids = {}
      for source in Mappings:
          ids[source] = []
          for identifier in Mappings[source]:
              ids[source].append(identifier)
```

```
GenesTable = pd.  
↳DataFrame(columns=['Entrez', 'HGNC', 'Ensembl'])  
GenesTable['Entrez'] = Genes  
GenesTable['HGNC'] = ids['H']  
GenesTable['Ensembl'] = ids['En']  
  
display(GenesTable)
```

	Entrez	HGNC	Ensembl
0	5465	PPARA	ENSG00000186951
1	403654		ENSCAFG00000000788
2	19013		ENSMUSG00000022383
3	25747		ENSRNOG00000021463

EdelweissData explorer

Curated datasets are made available through the EdelweissData Explorer, the main data provisioning tool in the DataCure case study of OpenRiskNet. It is a web-based data explorer tool that gives users the ability to filter, search and extract data through the use of API calls. The EdelweissData Explorer serves data from ToxCast, ToxRefDB, and TG-GATES.

Prior to using the EdelweissData explorer, the EdelweissData library is initialized and authenticated.

```
[11]: api = API(edelweiss_api_url)  
      api.authenticate()
```

Tox21/ToxCast

Based on the identified genes and chemicals of interest by the AOP-DB RDF and AOP-Wiki RDF, the EdelweissData library is used to find datasets with those particular target genes and chemicals.

First, the assays are searched based on the Entrez gene IDs. These are, along with their metadata, stored in a dataframe.

Next, for those assays, all datasets are retrieved that are generated with the compounds related to the AOP.

```
[12]: columns = [  
    # ("Endpoint", "$.assay.component.endpoint"),  
    ("Endpoint name", "$.assay.component.endpoint.  
    ↪assay_component_endpoint_name.value"),  
    ("Biological target", "$.assay.component.endpoint.target.  
    ↪biological_process_target.value"),  
    ("Entrez gene ID for the molecular target", "$.assay.  
    ↪component.endpoint.target.intended.intended_target_gene.  
    ↪intended_target_entrez_gene_id.value"),  
    ("Symbol", "$.assay.component.endpoint.target.intended.  
    ↪intended_target_gene.intended_target_official_symbol.  
    ↪value"),  
    ("Gene name", "$.assay.component.endpoint.target.  
    ↪intended_target_gene.intended_target_gene_name.  
    ↪value")]  
  
cquery = None  
for gene in Genes:  
    if cquery is None:  
        cquery = Q.search_anywhere("EPA-InVitroDBV3.2") & Q.  
        ↪search_anywhere("summary") & Q.exact_search(Q.  
        ↪column('Entrez gene ID for the molecular target'), gene)  
    else:  
        cquery = cquery | Q.search_anywhere("EPA-InVitroDBV3.  
        ↪2") & Q.search_anywhere("summary") & Q.exact_search(Q.  
        ↪column('Entrez gene ID for the molecular target'), gene)  
  
ToxCast = api.get_published_datasets(limit=200, ↵  
    ↪columns=columns, condition=cquery)  
ToxCast
```

[12]:

id	version	dataset	Endpoint name	Biological target	Entrez gene ID for the molecular target	Symbol	Gene name
d651ef92-c12e-4eba-8979-8dc3f77fc7f3	1	<PublishedDataset 'd65... ...ATG_PPArA_TRANS _dn_summary_tctl>	ATG_PPArA_TRANS_dn	regulation of transcription factor activity	5465	PPARA	peroxisome proliferator-activated receptor alpha
5a9cf864-520d-4ca0-b77f-8333aa8a3d5c	1	<PublishedDataset '5a9... ATG_PPRe_CIS_dn _summary_tctl>	ATG_PPRe_CIS_dn	regulation of transcription factor activity	5465	PPARA	peroxisome proliferator-activated receptor alpha
cd27c421-8273-41cb-9e11-8c96a2b30e10	1	<PublishedDataset 'cd2... NVS_NR_hPPArA _summary_tctl>	NVS_NR_hPPArA	receptor binding	5465	PPARA	peroxisome proliferator-activated receptor alpha
291d8013-662b-4b44-aa96-894be4473d59	1	<PublishedDataset '291... ATG_PPRe_CIS_up _summary_tctl>	ATG_PPRe_CIS_up	regulation of transcription factor activity	5465	PPARA	peroxisome proliferator-activated receptor alpha
f3ebdf70-976f-4b19-92a1-25b189fa13fa	1	<PublishedDataset 'f3e... ATG_PPArA_TRANS_up _summary_tctl>	ATG_PPArA_TRANS_up	regulation of transcription factor activity	5465	PPARA	peroxisome proliferator-activated receptor alpha

```
[13]: cquery = None
for compound in compoundstable['InChIKey'].values:
    if cquery is None:
        cquery = Q.fuzzy_search(Q.column('InChI key'),
        ↳compound)
    else:
        cquery = cquery | Q.fuzzy_search(Q.column('InChI_
        ↳key'), compound)

ToxCastData = pd.DataFrame()
for index, row in ToxCast.iterrows():
    tmpdata = row['dataset'].get_data(limit=None,
    ↳condition=cquery)
    tmpdata['Assay'] = row['Endpoint name']
    tmpdata = tmpdata[['Assay', 'DTXSID', 'Substance name',
    ↳'InChI key', 'CAS', 'IC50', 'Quality check']]
    ToxCastData = pd.concat([ToxCastData, tmpdata])

ToxCastData.sort_values(by=['InChI key', 'Assay'])
```

[13]:

Assay	DTXSID	Substance name	InChI key	CAS	IC50	Quality check
3714 ATG_PPArA_TRANS_dn	DTXSID5020607	Di(2-ethylhexyl) phthalate	B[QHLKABX]IVAM-UHFFFAOYSA-N	117-81-7	NaN	[]
3714 ATG_PPArA_TRANS_up	DTXSID5020607	Di(2-ethylhexyl) phthalate	B[QHLKABX]IVAM-UHFFFAOYSA-N	117-81-7	NaN	[Borderline inactive]
3714 ATG_PPRe_CIS_dn	DTXSID5020607	Di(2-ethylhexyl) phthalate	B[QHLKABX]IVAM-UHFFFAOYSA-N	117-81-7	NaN	[]
3714 ATG_PPRe_CIS_up	DTXSID5020607	Di(2-ethylhexyl) phthalate	B[QHLKABX]IVAM-UHFFFAOYSA-N	117-81-7	1.414849	[]
24 NVS_NR_hPPArA	DTXSID5020607	Di(2-ethylhexyl) phthalate	B[QHLKABX]IVAM-UHFFFAOYSA-N	117-81-7	NaN	[]
919 ATG_PPArA_TRANS_dn	DTXSID0020652	Gemfibrozil	HEMJJKBWTPKOJG-UHFFFAOYSA-N	25812-30-0	NaN	[]
919 ATG_PPArA_TRANS_up	DTXSID0020652	Gemfibrozil	HEMJJKBWTPKOJG-UHFFFAOYSA-N	25812-30-0	1.558449	[]
919 ATG_PPRe_CIS_dn	DTXSID0020652	Gemfibrozil	HEMJJKBWTPKOJG-UHFFFAOYSA-N	25812-30-0	NaN	[]
919 ATG_PPRe_CIS_up	DTXSID0020652	Gemfibrozil	HEMJJKBWTPKOJG-UHFFFAOYSA-N	25812-30-0	1.509815	[]
3245 ATG_PPArA_TRANS_dn	DTXSID3029869	Bezafibrate	IIBYAHWJQTYFKB-UHFFFAOYSA-N	41859-67-0	NaN	[Noisy data]
3245 ATG_PPArA_TRANS_up	DTXSID3029869	Bezafibrate	IIBYAHWJQTYFKB-UHFFFAOYSA-N	41859-67-0	0.918967	[Hit-call potentially confounded by overfitting]
3245 ATG_PPRe_CIS_dn	DTXSID3029869	Bezafibrate	IIBYAHWJQTYFKB-UHFFFAOYSA-N	41859-67-0	NaN	[Noisy data]
3245 ATG_PPRe_CIS_up	DTXSID3029869	Bezafibrate	IIBYAHWJQTYFKB-UHFFFAOYSA-N	41859-67-0	1.208669	[]
4050 ATG_PPArA_TRANS_dn	DTXSID3020336	Clofibrate	KNHUKKLJHYUCFP-UHFFFAOYSA-N	637-07-0	NaN	[]
4050 ATG_PPArA_TRANS_up	DTXSID3020336	Clofibrate	KNHUKKLJHYUCFP-UHFFFAOYSA-N	637-07-0	1.653510	[]
4050 ATG_PPRe_CIS_dn	DTXSID3020336	Clofibrate	KNHUKKLJHYUCFP-UHFFFAOYSA-N	637-07-0	NaN	[]
4050 ATG_PPRe_CIS_up	DTXSID3020336	Clofibrate	KNHUKKLJHYUCFP-UHFFFAOYSA-N	637-07-0	NaN	[Borderline inactive]
353 NVS_NR_hPPArA	DTXSID3020336	Clofibrate	KNHUKKLJHYUCFP-UHFFFAOYSA-N	637-07-0	NaN	[]
2464 ATG_PPArA_TRANS_dn	DTXSID8020331	Ciprofibrate	KPSRODZRAIWAKH-UHFFFAOYSA-N	52214-84-3	NaN	[Noisy data]
2464 ATG_PPArA_TRANS_up	DTXSID8020331	Ciprofibrate	KPSRODZRAIWAKH-UHFFFAOYSA-N	52214-84-3	0.009889	[Hit-call potentially confounded by overfitting]
2464 ATG_PPRe_CIS_dn	DTXSID8020331	Ciprofibrate	KPSRODZRAIWAKH-UHFFFAOYSA-N	52214-84-3	NaN	[]
2464 ATG_PPRe_CIS_up	DTXSID8020331	Ciprofibrate	KPSRODZRAIWAKH-UHFFFAOYSA-N	52214-84-3	1.722385	[]
132 ATG_PPArA_TRANS_dn	DTXSID4020290	Pirinixic acid	SZRPDCCEHVWOJX-UHFFFAOYSA-N	50892-23-4	NaN	[Noisy data]
132 ATG_PPArA_TRANS_up	DTXSID4020290	Pirinixic acid	SZRPDCCEHVWOJX-UHFFFAOYSA-N	50892-23-4	0.745003	[]
132 ATG_PPRe_CIS_dn	DTXSID4020290	Pirinixic acid	SZRPDCCEHVWOJX-UHFFFAOYSA-N	50892-23-4	NaN	[]
132 ATG_PPRe_CIS_up	DTXSID4020290	Pirinixic acid	SZRPDCCEHVWOJX-UHFFFAOYSA-N	50892-23-4	1.307836	[]
409 NVS_NR_hPPArA	DTXSID4020290	Pirinixic acid	SZRPDCCEHVWOJX-UHFFFAOYSA-N	50892-23-4	0.943310	[]
3855 ATG_PPArA_TRANS_dn	DTXSID2029874	Fenofibrate	YMTINGFKWXXKFG-UHFFFAOYSA-N	49562-28-9	NaN	[Noisy data]
3855 ATG_PPArA_TRANS_up	DTXSID2029874	Fenofibrate	YMTINGFKWXXKFG-UHFFFAOYSA-N	49562-28-9	0.727661	[]
3855 ATG_PPRe_CIS_dn	DTXSID2029874	Fenofibrate	YMTINGFKWXXKFG-UHFFFAOYSA-N	49562-28-9	NaN	[]
3855 ATG_PPRe_CIS_up	DTXSID2029874	Fenofibrate	YMTINGFKWXXKFG-UHFFFAOYSA-N	49562-28-9	1.191227	[]
255 NVS_NR_hPPArA	DTXSID2029874	Fenofibrate	YMTINGFKWXXKFG-UHFFFAOYSA-N	49562-28-9	NaN	[]

TG-GATES

Based on the chemicals of interest, all transcriptomics datasets from TG-GATES are queried that are the result of exposure to those chemicals.

```
[14]: columns = [
    ("Compound", "$.Compound.Name"),
    ("InChI Key", "$.Compound.\"InChI Key\""),
    ("CAS", "$.Compound.CAS"),
    ("Organism", "$.Assay.Organism"),
    ("Organ", "$.Assay.Organ"),
    ("Study type", "$.Assay.\"Study type\""),
    ("Dose", "$.Assay.Exposure.Dose"),
    ("Dosing", "$.Assay.Dosing"),
    ("Duration", "$.Assay.Exposure.Duration"),
    ("Duration unit", "$.Assay.Exposure.\"Duration unit\""),
]

cquery = None
for compound in compoundstable['InChIKey'].values:
    if cquery is None:
```

```

cquery = Q.fuzzy_search(Q.column('InChI Key'),
    ↳compound)
else:
    cquery = cquery | Q.fuzzy_search(Q.column('InChI_
    ↳Key'), compound)

condition = Q.search_anywhere("TG-GATES") & Q.
    ↳search_anywhere("FOLD_CHANGES") & (cquery)
TGGATES = api.get_published_datasets(limit=api.
    ↳get_raw_datasets(limit=0, columns=columns,
    ↳condition=condition)['total'], columns=columns,
    ↳condition=condition)
TGGATES['Duration2'] = TGGATES['Duration'].map(str) +
    ↳TGGATES['Duration unit']
TGGATES['Duration'] = TGGATES['Duration2']
TGGATES = TGGATES.drop(['Duration2', 'Duration unit'], axis=1)
TGGATES = TGGATES.sort_values(by=['InChI Key'])
TGGATES

```

[14]:

id	version	dataset	Compound	InChI Key	CAS	Organism	Organ	Study type	Dose	Dosing	Duration
7285e2d4-dd0d		<PublishedDataset '728...									
-458b-83b6-a25	1	gemfibrozil_Rat_Liver_in_vivo_Repeat_15_day_high_FOLD_CHANGES>	gemfibrozil	InChIKey=HEMJJKBWTPKOJG-UHFFFAOYSA-N	25812-30-0	Rat	Liver	in_vivo	high	Repeat	15day
8ed097042		<PublishedDataset '562...									
cb29047e-745f1	1	gemfibrozil_Rat_Liver_in_vivo_Repeat_29_day_low_FOLD_CHANGES>	gemfibrozil	InChIKey=HEMJJKBWTPKOJG-UHFFFAOYSA-N	25812-30-0	Rat	Liver	in_vivo	low	Repeat	29day
-45c2-94ab-510		<PublishedDataset '443...									
406abc89e		gemfibrozil_Rat_Liver_in_vivo_Single_24_hr_high_FOLD_CHANGES>	gemfibrozil	InChIKey=HEMJJKBWTPKOJG-UHFFFAOYSA-N	25812-30-0	Rat	Liver	in_vivo	high	Single	24hr
443072a-ee3d	1	<PublishedDataset '784...									
-4d01-9e12-455		gemfibrozil_Rat_Liver_in_vitro_2_hr_middle_FOLD_CHANGES>	gemfibrozil	InChIKey=HEMJJKBWTPKOJG-UHFFFAOYSA-N	25812-30-0	Rat	Liver	in_vitro	middle	None	2hr
1140eb528		<PublishedDataset '59f...									
7840e0f1-a092	1	gemfibrozil_Human_Liver_in_vitro_2_hr_high_FOLD_CHANGES>	gemfibrozil	InChIKey=HEMJJKBWTPKOJG-UHFFFAOYSA-N	25812-30-0	Human	Liver	in_vitro	high	None	2hr
-45be-9b8d-836		<PublishedDataset '968...									
5b92d383f		fenofibrate_Rat_Liver_in_vivo_Single_9_hr_low_FOLD_CHANGES>	fenofibrate	InChIKey=YMTINGFKWXXKFG-UHFFFAOYSA-N	49562-28-9	Rat	Liver	in_vivo	low	Single	9hr
-49fb-763-dab6	1	<PublishedDataset '361...									
-431c-97a9-6fc		fenofibrate_Rat_Liver_in_vitro_2_hr_high_FOLD_CHANGES>	fenofibrate	InChIKey=YMTINGFKWXXKFG-UHFFFAOYSA-N	49562-28-9	Rat	Liver	in_vitro	high	None	2hr
44307d27a		<PublishedDataset '128...									
...		fenofibrate_Rat_Liver_in_vivo_Single_6_hr_high_FOLD_CHANGES>	fenofibrate	InChIKey=YMTINGFKWXXKFG-UHFFFAOYSA-N	49562-28-9	Rat	Liver	in_vivo	high	Single	6hr
-efeb9b4d-4315		<PublishedDataset '472...									
-4e56-b8a3-151		fenofibrate_Rat_Liver_in_vitro_24_hr_middle_FOLD_CHANGES>	fenofibrate	InChIKey=YMTINGFKWXXKFG-UHFFFAOYSA-N	49562-28-9	Rat	Liver	in_vitro	middle	Single	24hr
bb4a79e77	1	<PublishedDataset '361...									
96834324-4434		fenofibrate_Rat_Liver_in_vitro_2_hr_high_FOLD_CHANGES>	fenofibrate	InChIKey=YMTINGFKWXXKFG-UHFFFAOYSA-N	49562-28-9	Rat	Liver	in_vitro	high	None	2hr
-4314-bab4-a33	1	<PublishedDataset '128...									
0c4fcecd		fenofibrate_Rat_Liver_in_vivo_Single_6_hr_high_FOLD_CHANGES>	fenofibrate	InChIKey=YMTINGFKWXXKFG-UHFFFAOYSA-N	49562-28-9	Rat	Liver	in_vivo	high	Single	6hr
361abe50-9559		<PublishedDataset '472...									
-4033-b614-fc9	1	fenofibrate_Rat_Liver_in_vitro_24_hr_high_FOLD_CHANGES>	fenofibrate	InChIKey=YMTINGFKWXXKFG-UHFFFAOYSA-N	49562-28-9	Rat	Liver	in_vitro	high	None	24hr
13a2d87ae		<PublishedDataset '472...									
128264b2-6bed		fenofibrate_Rat_Liver_in_vitro_24_hr_high_FOLD_CHANGES>	fenofibrate	InChIKey=YMTINGFKWXXKFG-UHFFFAOYSA-N	49562-28-9	Rat	Liver	in_vitro	high	None	24hr
-4cfd-9896-90b	1	<PublishedDataset '472...									
7aec50a3f		fenofibrate_Rat_Liver_in_vitro_24_hr_high_FOLD_CHANGES>	fenofibrate	InChIKey=YMTINGFKWXXKFG-UHFFFAOYSA-N	49562-28-9	Rat	Liver	in_vitro	high	None	24hr
ef2ba35a-7a08	1	<PublishedDataset '472...									
-41f3-878f-424		fenofibrate_Rat_Liver_in_vitro_24_hr_high_FOLD_CHANGES>	fenofibrate	InChIKey=YMTINGFKWXXKFG-UHFFFAOYSA-N	49562-28-9	Rat	Liver	in_vitro	high	None	24hr
65c8c33d		<PublishedDataset '472...									

Filtering the datasets

In most cases, TG-GATES contains multiple datasets for the chemicals of interest, and should therefore be filtered. The next section allows the filtering of the datasets by the different columns from the TGGATES table.

Identify possible values per category

First, all possible values for filters should be identified, summarizing the TGGATES table for the columns: - Compound - Organism - Organ - Study type - Dosing - Dose - Duration

Second, a table will be generated with the possible compounds and their corresponding CAS IDs and InChI Keys.

```
[15]: #Options for categories
compounds = set(TGGATES['Compound'].tolist())
organisms = set(TGGATES['Organism'].tolist())
organs = set(TGGATES['Organ'].tolist())
studytypes = set(TGGATES['Study type'].tolist())
dosings = set(TGGATES['Dosing'].tolist())
doses = set(TGGATES['Dose'].tolist())
durations = set(TGGATES['Duration'].tolist())

print('Data available for compounds: '+str(compounds))
print('Data available for organisms: '+str(organisms))
print('Data available for organs: '+str(organs))
print('Data available for study types: '+str(studytypes))
print('Data available for dosings: '+str(dosings))
print('Data available for doses: '+str(doses))
print('Data available for durations: '+str(durations))

#Table for compounds and corresponding identifiers
df = pd.DataFrame()
chemdict = {}

for index, row in TGGATES.iterrows():
    if not row['Compound'] in chemdict:
        chemdict[row['Compound']] = {}
        chemdict[row['Compound']]['CAS'] = row['CAS']
        chemdict[row['Compound']]['InChI Key'] = row['InChI_
↪Key']

df = pd.DataFrame.from_dict(chemdict, orient='index')
df
```

```
Data available for compounds: {'gemfibrozil', 'fenofibrate', '
↪WY-14643',
```

```
'clofibrate'}
Data available for organisms: {'Rat', 'Human'}
Data available for organs: {'Kidney', 'Liver'}
Data available for study types: {'in_vivo', 'in_vitro'}
Data available for dosings: {'Repeat', None, 'Single'}
Data available for doses: {'low', 'middle', 'high'}
Data available for durations: {'3hr', '24hr', '6hr', '15day', '8hr', '4day', '9hr', '8day', '2hr', '29day'}
```

[15]:

	CAS	InChI Key
gemfibrozil	25812-30-0	InChIKey=HEMJJKBWTPKOJG-UHFFFAOYSA-N
clofibrate	637-07-0	InChIKey=KNHUKKLJHYUCFP-UHFFFAOYSA-N
WY-14643	50892-23-4	InChIKey=SZRPDCCEHVWOJX-UHFFFAOYSA-N
fenofibrate	49562-28-9	InChIKey=YMTINGFKWWXKFG-UHFFFAOYSA-N

Filter the datasets

The values identified for each column can be used to filter the datasets.

In order to do that, two lists should be filled: - A list of all categories involved in the filter, corresponding to the column header. This list is called `list_cat` - A list of the filter values for all categories included in the category list. This list is called `list_input`

Note that the location of the filled values is important. The sequence of the categories, and their values should correspond between the two lists. For each category, one filter value can be filled in the filter list. One can enter as many filters as necessary, from filtering for only the organism variable to filtering for all possible variables captured in the TGGATES dataframe. The `res_df` table contains the dataset(s) that are used later for pathway analysis.

[16]:

```
list_cat = ["Compound", "Organism", "Organ", "Study_
↳type", "Dosing", "Dose", "Duration"]
list_input = ["gemfibrozil", "Rat", "Liver", "in_vivo", "
↳Repeat", "high", "8day"]
```

```

filterlist = pd.DataFrame(
    {'Category': list_cat,
     'Input': list_input
    })
display(filterlist)

def tg_gates(df, list_cat, list_input):
    comb = zip(list_cat, list_input)
    sub_df = df
    for tup in comb:
        sub_df = sub_df[sub_df[tup[0]]==tup[1]]
    return sub_df

res_df = tg_gates(TGGATES, list_cat, list_input)
display(res_df)

```

	Category	Input
0	Compound	gemfibrozil
1	Organism	Rat
2	Organ	Liver
3	Study type	in_vivo
4	Dosing	Repeat
5	Dose	high
6	Duration	8day

id	version	dataset	Compound	InChI Key	CAS	Organism	Organ	Study type	Dose	Dosing	Duration
e45c8706-010e-4597-b490-df1-38f57f06e	1	<PublishedDataset 'e45... gemfibrozil_Rat_Liver_ in vivo_Repeat_8_day_ high_FOLD_CHANGES>	gemfibrozil	InChIKey= HEMJJKBWTPK OJG-UHFFFAOYSA-N	25812-30-0	Rat	Liver	in_vivo	high	Repeat	8day

```

[22]: deg = []
for index, row in res_df.iterrows():
    file = row['dataset'].get_data(limit=100000)
    assay_deg = file[((file['logFC'] > 1) | (file['logFC'] < -1)) & (file['P.Value'] < 0.05)]['ENTREZID'].unique().tolist()
    if len(assay_deg) > 0:
        for gene in assay_deg:
            deg.append(gene)

```

```
print('Number of differentially expressed genes:␣  
↪'+str(len(deg)))  
degforpwanalysis = set(deg)
```

Number of differentially expressed genes: 50

WikiPathways RDF

Service description

WikiPathways is a community-driven molecular pathway database, supporting wide-spread topics and supported by many databases and integrative resources. It contains semantic annotations in its pathways for genes, proteins, metabolites, and interactions using a variety of reference databases, and WikiPathways is used to analyze and integrate experimental omics datasets (Slenter et al., 2017). Furthermore, human pathways from Reactome (Fabregat et al., 2018), another molecular pathway database, are integrated with WikiPathways and are therefore part of the WikiPathways RDF (Waagmeester et al., 2016). On the OpenRiskNet e-infrastructure, the WikiPathways RDF, which includes the Reactome pathways, is exposed via a Virtuoso SPARQL endpoint.

Implementation

The first section is to find all molecular pathways in WikiPathways that contain the genes of interest, by matching the results of a SPARQL query to the list of genes found with the AOP-DB RDF. The SPARQL query extracts all Entrez gene IDs from all pathways in WikiPathways. When overlap between those lists are found, the pathway ID is stored in a dataframe along with its title, and organism.

The next section is for extracting all pathways for the species of interest, along with all genes present. These are used later for pathway analysis with the data extracted from TG-GATES through the EdelweissData explorer.

```
[23]: sparqlquery = '''
    SELECT DISTINCT (str(?wpid) as ?Pathway_ID) (str(?
↪PW_Title) as ?Pathway_title) (fn:substring(?
↪ncbiGeneId,33) as ?Entrez) ?organism
    WHERE {
        ?gene a wp:GeneProduct; dcterms:identifier ?id; dcterms:
↪isPartOf ?pathwayRes; wp:bdbEntrezGene ?ncbiGeneId.
        ?pathwayRes a wp:Pathway; dcterms:identifier ?wpid; dc:
↪title ?PW_Title; wp:organismName ?organism.}
    '''

wikipathwayssparql.setQuery(sparqlquery)
wikipathwayssparql.setReturnFormat(JSON)
results = wikipathwayssparql.query().convert()

def intersection(lst1, lst2):
    lst3 = [value for value in lst1 if value in lst2]
    return lst3

WikiPathwaysGenes = {}
WikiPathwaysNames = {}
WikiPathwaysOrganism = {}
for result in results["results"]["bindings"]:
    WikiPathwaysGenes[result["Pathway_ID"]
    ["value"]] = set([])
for result in results["results"]["bindings"]:
    WikiPathwaysGenes[result["Pathway_ID"]
    ["value"]].add(result["Entrez"]["value"])
for result in results["results"]["bindings"]:
    WikiPathwaysNames[result["Pathway_ID"]
    ["value"]] = result["Pathway_title"]["value"]
for result in results["results"]["bindings"]:
    WikiPathwaysOrganism[result["Pathway_ID"]
    ["value"]] = result["organism"]["value"]

for lst in WikiPathwaysGenes:
    genematch = intersection(WikiPathwaysGenes[lst], Genes)
    WikiPathwaysGenes[lst] = [WikiPathwaysNames[lst],
↪WikiPathwaysOrganism[lst], genematch, len(genematch)]
```

```

WPtable = pd.DataFrame.from_dict(WikiPathwaysGenes,
    ↳orient='index', columns=['Pathway_title', 'Organism',
    ↳'Entrez Gene', 'nGenes'])
WPtable = WPtable[WPtable.nGenes >= 1]
WPtable = WPtable.drop(columns='nGenes')
display(WPtable)

```

	Pathway_title	Organism	Entrez Gene
WP1797	Circadian Clock	Homo sapiens	[5465]
WP3370	RORA activates gene expression	Homo sapiens	[5465]
WP299	Nuclear Receptors in Lipid Metabolism and Toxicity	Homo sapiens	[5465]
WP4721	Eicosanoid metabolism via Lipo Oxygenases (LOX)	Homo sapiens	[5465]
WP4447	SUMOylation of intracellular receptors	Homo sapiens	[5465]
WP4720	Eicosanoid metabolism via Cytochrome P450 Mono-Oxygenases (CYP) pathway	Homo sapiens	[5465]
WP2011	SREBF and miR33 in cholesterol and lipid homeostasis	Homo sapiens	[5465]
WP2882	Nuclear Receptors Meta-Pathway	Homo sapiens	[5465]
WP3594	Circadian rhythm related genes	Homo sapiens	[5465]
WP4396	Nonalcoholic fatty liver disease	Homo sapiens	[5465]
WP431	Nuclear receptors in lipid metabolism and toxicity	Mus musculus	[19013]
WP2084	SREBF and miR33 in cholesterol and lipid homeostasis	Mus musculus	[19013]
WP2316	PPAR signaling pathway	Mus musculus	[19013]
WP447	Adipogenesis genes	Mus musculus	[19013]
WP509	Nuclear Receptors	Mus musculus	[19013]
WP1099	Nuclear receptors in lipid metabolism and toxicity	Canis familiaris	[403654]
WP1105	Adipogenesis	Canis familiaris	[403654]
WP1184	Nuclear Receptors	Canis familiaris	[403654]
WP1541	Energy Metabolism	Homo sapiens	[5465]
WP2706	Activation of gene expression by SREBF (SREBP)	Homo sapiens	[5465]
WP3942	PPAR signaling pathway	Homo sapiens	[5465]
WP2878	PPAR Alpha Pathway	Homo sapiens	[5465]
WP3355	BMAL1:CLOCK,NPAS2 activates circadian gene expression	Homo sapiens	[5465]
WP170	Nuclear Receptors	Homo sapiens	[5465]
WP2881	Estrogen Receptor Pathway	Homo sapiens	[5465]
WP2797	Regulation of lipid metabolism by Peroxisome proliferator-activated receptor alpha (PPARalpha)	Homo sapiens	[5465]
WP236	Adipogenesis	Homo sapiens	[5465]
WP3331	Mitochondrial biogenesis	Homo sapiens	[5465]
WP1822	Generic Transcription Pathway	Homo sapiens	[5465]
WP139	Nuclear receptors in lipid metabolism and toxicity	Rattus norvegicus	[25747]
WP217	Nuclear Receptors	Rattus norvegicus	[25747]
WP155	Adipogenesis	Rattus norvegicus	[25747]
WP2751	Transcriptional regulation of white adipocyte differentiation	Homo sapiens	[5465]

Extracting pathways for pathway analysis

Prior to extracting pathways, the species of interest should be stored in a variable. Note that this should be the latin species name, and

corresponding to the filtered dataset(s) from TG-GATES.

```
[24]: OrganismFilter = 'Rattus norvegicus'
```

```
[25]: sparqlquery = '''
        SELECT DISTINCT (str(?wpid) as ?Pathway_ID) (str(?
        ↪PW_Title) as ?Pathway_title) (fn:substring(?
        ↪ncbiGeneId,33) as ?Entrez)
        WHERE {
            ?gene a wp:GeneProduct; dcterms:identifier ?id; dcterms:
        ↪isPartOf ?pathwayRes; wp:bdbEntrezGene ?ncbiGeneId.
            ?pathwayRes a wp:Pathway; dcterms:identifier ?wpid; dc:
        ↪title ?PW_Title; wp:organismName_
        ↪"'+OrganismFilter+'"^^xsd:string.}
        '''

wikipathwayssparql.setQuery(sparqlquery)
wikipathwayssparql.setReturnFormat(JSON)
results = wikipathwayssparql.query().convert()

WikiPathwaysGenes = {}
WikiPathwaysNames = {}
for result in results["results"]["bindings"]:
    WikiPathwaysGenes[result["Pathway_ID"]["value"]] =
    ↪set([])
for result in results["results"]["bindings"]:
    WikiPathwaysNames[result["Pathway_ID"]["value"]] =
    ↪result["Pathway_title"]["value"]
for result in results["results"]["bindings"]:
    WikiPathwaysGenes[result["Pathway_ID"]["value"]].
    ↪add(result["Entrez"]["value"])

for lst in WikiPathwaysGenes:
    WikiPathwaysGenes[lst] = [WikiPathwaysGenes[lst],
    ↪len(WikiPathwaysGenes[lst])]

WTable = pd.DataFrame.from_dict(WikiPathwaysGenes,
    ↪orient='index', columns=['Genes', 'nGenes'])
display(WTable)
```

	Genes	nGenes
WP1286	{154516, 24861, 292155, 25279, 301264, 29738, 311257, 103690051, 302302, 24424, 494499, 81869, 361631, 316325, 574523, 396551, 293779, 116631, 292915, 25426, 65030, 246767, 65185, 24426, 192242, 286954, 305264, 304322, 25147, 24404, 307092, 113992, 299566, 308190, 114846, 108348061, 246245, 24422, 24192, 24902, 26760, 685402, 301517, 100910526, 310848, 24912, 500257, 361510, 295430, 353498, 81676, 25458, 500892, 363618, 310855, 171341, 154985, 29326, 297029, 286921, 58953, 116632, 246247, 24298, 500359, 116686, 100910462, 303218, 293451, 25315, 312495, 84406, 364476, 25428, 83783, 288108, 24296, 114700, 301595, 81924, 360268, 499302, 64352, 171445, 396527, 290623, 300850, 307838, 25086, 25355, 25427, 25146, 29725, 286989, 303669, 289197, 691394, 362228, 24297, 108348148, ...}	143
WP1290	{24508, 24577, 81525, 84359, 29884, 502004, 64547, 24185, 291100, 103694380, 500592, 309165, 100911660, 100360940, 309295, 117279, 25718, 60434, 292892, 364081, 362675, 492821, 25272, 296953, 78963, 687813, 78971, 25166, 83584, 25402, 24842, 316256, 64625, 290749, 114555, 314856, 25385, 312030, 64044, 140923, 24482, 309361, 25008, 64026, 24516, 116502, 314756, 156767, 25625, 103689977, 64314, 266610, 246775, 103690372, 287398, 246097, 362788, 116667, 64041, 294071, 246756, 24887, 246334, 25513, 84351, 316241, 24483, 25493, 60371, 114214, 64639, 63879, 24224, 293624, 311786, 58918, 81736, 306886}	78
WP1278	{84351, 309361, 313121, 116554, 116590, 309452, 29538, 311245, 103694380, 81736, 309165, 81780, 299331, 24481, 360640, 286908, 246756}	17
WP1279	{84581, 288588, 363481, 498109, 116590, 24185, 288533, 690966, 81649, 84577, 297893, 81504, 50658, 292778, 114486, 314322, 84582, 103695118, 81646, 58919, 100363500, 84578, 363287, 84580, 170922, 313845, 306516, 24400, 81674, 83503, 83828, 310784, 117526, 680149, 361580, 682902, 497672, 171150, 307485, 309361, 171104, 117017, 24790, 24516, 25636, 288651, 294236, 81673, 83805, 289561, 294018, 361365, 309224, 294693, 103690054, 291703, 308415, 84351, 363633, 24890, 24224, 170851, 84389, 293621, 367858, 373541, 314612, 314436, 54244, 81736, 170915, 363067, 266713, 303918}	74
WP1282	{368066, 25331, 690050, 300691, 689330, 171347, 24661, 24267, 81676}	9
...
WP547	{85253, 306761, 24233, 79224, 24946, 25439, 113959, 25619, 24548, 24903, 83580, 24366, 298566, 81750, 116669, 117512, 29251, 25048, 29243, 113936, 295703, 65051, 79126, 288001, 289055, 29333, 29436, 117517, 302470, 25268, 289395, 50692, 24232, 24648, 54249, 24153, 29687, 192262, 313421, 64036, 64459, 24234, 362634, 304917, 155012, 24231, 25584, 287527, 81509, 24441, 25407, 312705, 24617, 24237, 54243, 290757, 84007, 64023, 260320, 25692, 294257}	61
WP505	{25125, 84353, 317376, 50554, 25712, 161452, 59328, 59107, 50658, 25313, 314322, 59086, 85435, 25495, 114208, 24881, 60584, 300054, 311061, 156726, 445442, 29357, 25631, 25296, 24835, 103691556, 311071, 24516, 24373, 81516, 94188, 367264, 25671, 29200, 50689, 25353, 316742, 367218, 25639, 367100, 29591, 83837, 29610, 293621, 367858, 54244, 24617, 81736, 81810, 25124, 170915, 313477, 497010, 84598}	54
WP654	{300711, 25203, 114483, 54237, 25112, 58919, 492821, 78963, 24708, 25402, 24842, 497672, 311562, 116502, 399489, 680110, 114851, 300668, 25729, 24887, 114212, 25309, 362817, 24224, 58918, 298795, 100363502}	27
WP89	{24493, 24471, 64159, 84359, 362456, 117279, 140657, 299625, 313121, 24708, 78963, 116554, 83584, 25402, 29432, 25385, 64044, 24835, 140926, 64026, 24516, 266610, 25591, 103689977, 289014, 287398, 246097, 362491, 315994, 116667, 360748, 114214, 25309, 24224, 58918, 60374, 116685, 100363502, 29431}	39
WP81	{24233, 24232, 362119, 29687, 192262, 312705, 313421, 64036, 362634, 24237, 298566, 64023, 117512, 24231, 117517, 298288}	16

This section is the actual pathway analysis, using the differentially expressed genes from TG-GATES, and the molecular pathways from WikiPathways. With some basic statistics, a Z-score can be calculated for all pathways. A Z-score above 1.96 is considered significant.

```
[26]: ngenepres = []
      for index, row in WTable.iterrows():
          genepres = []
          for gene in row['Genes']:
              for sig in degforpwanalysis:
                  if gene == str(sig):
                      genepres.append(gene)
          ngenepres.append(len(genepres))
      WTable['nSigGenes'] = ngenepres
      WTable['percentSigGenes'] = (WTable['nSigGenes'] /
      ↪ WTable['nGenes']) * 100

      total = []
      for index, row in WTable.iterrows():
          total.append(row['nSigGenes'])

      StandardDeviation = statistics.stdev(total)
      ExpectedValue = (sum(total) / len(WTable))

      WTable['Zscore'] = (WTable['nSigGenes'] - ExpectedValue) /
      ↪ StandardDeviation
      WTable = WTable.sort_values(by=['Zscore'], ascending=False)
      WTable = WTable[WTable.Zscore >= 1.96]
      display(WTable)
```

	Genes	nGenes	nSigGenes	percentSigGenes	Zscore
	{154516, 24861, 292155, 25279, 301264, 29738, 311257, 103690051, 302302, 24424, 494499, 81869, 361631, 316325, 574523, 396551, 293779, 116631, 292915, 25426, 65030, 246767, 65185, 24426, 192242, 286954, 305264, 304322, 25147, 24404, 307092, 113992, 299566, 308190, 114846, 108348061, 246245, 24422, 24192, 24902, 26760, 685402, 301517, 100910526, 310848, 24912, 500257, 361510, 295430, 353498, 81676, 25458, 500892, 363618, 310855, 171341, 154985, 29326, 297029, 286921, 58953, 116632, 246247, 24298, 500359, 116686, 100910462, 303218, 293451, 25315, 312495, 84406, 364476, 25428, 83783, 288108, 24296, 114700, 301595, 81924, 360268, 499302, 64352, 171445, 396527, 290623, 300850, 307838, 25086, 25355, 25427, 25146, 29725, 286989, 303669, 289197, 691394, 362228, 24297, 108348148, ...}	143	4	2.797203	5.401926
WP1286					
	{170670, 113976, 117243, 25330, 25413, 25288, 364975, 25062, 24849, 94340, 361676, 25014, 117035, 171155, 140547, 24539, 311569, 79223, 311849, 24158, 24538, 289481, 113965, 29367, 117543, 25757, 25287, 114024, 29740, 100911615, 25363, 298942, 64304, 25756, 50682}	35	3	8.571429	3.959527
WP1307					
	{170670, 291468, 113976, 25330, 117243, 25413, 25288, 364975, 25062, 94340, 361676, 25014, 117035, 171155, 24539, 100911186, 311569, 311849, 24158, 289481, 24538, 113965, 29367, 117543, 25757, 25287, 114024, 29740, 100911615, 298942, 25363, 64304, 25756, 50682}	34	3	8.823529	3.959527
WP506					
	{29563, 25271, 366791, 310903, 499985, 25703, 24172, 89826, 24705, 29184, 24539, 25086, 361801, 29646, 690953, 83574, 24706, 266603, 100145871, 293049, 25061, 154985, 25056, 155192, 24188, 313689, 116676, 362662, 685072, 246298, 114628, 292915, 432367, 24710, 64047, 25073, 353252, 312495, 100365047, 314264, 83783, 114106}	42	3	7.142857	3.959527
WP1297					
	{170670, 24158, 25363, 113976, 171142, 113965, 117035, 25413, 25288, 64304, 25757, 25287, 114024, 25541, 29740, 113956}	16	3	18.750000	3.959527
WP419					
	{114700, 84356, 24646, 25270, 81924, 25279, 25664, 313210, 24307, 24891, 25303, 24705, 25086, 83569, 24706, 25682, 65035, 85264, 361523, 154985, 140668, 84385, 685072, 24297, 29277, 114628, 60351, 24873, 58852, 170913, 25428, 25747}	32	2	6.250000	2.517128
WP139					
	{25384, 24538, 84497, 25675, 300438, 25292, 310900, 25428, 25728, 296371, 25081, 24539, 25073, 24207, 25080, 299858, 81782, 313210, 108348160, 24530}	20	2	10.000000	2.517128
WP145					
	{116643, 140727, 85420, 100361457, 116590, 684969, 298947, 29224, 83472, 50658, 24451, 116554, 24424, 310392, 291796, 100158233, 81869, 497672, 24919, 29437, 366960, 29739, 24565, 85421, 24534, 287876, 50689, 497931, 304127, 140668, 24314, 29292, 24426, 361632, 295549, 83688, 25445, 25522, 25150, 58960, 24185, 114846, 108348061, 26760, 114495, 25365, 64188, 24842, 100360087, 679217, 25581, 171379, 25458, 24516, 24252, 171341, 114851, 286921, 85430, 361568, 29326, 297029, 117262, 57298, 116667, 25513, 301252, 301555, 117254, 116686, 170538, 24778, 170851, 25283, 293621, 54349, 54244, 25315, 681050, 84027, 25073, 170915, 83619, 117263, 305540, 24552, 24189, 25023, 29741, 25260, 497932, 24908, 64191, 300711, 100912585, 81649, 289623, 25352, 79255, 65052, ...}	167	2	1.197605	2.517128
WP2376					
	{170670, 113976, 117243, 25330, 25413, 25288, 364975, 25062, 24849, 94340, 361676, 25014, 117035, 171155, 140547, 24539, 311569, 79223, 311849, 24158, 24538, 289481, 113965, 29367, 25757, 25287, 114024, 100911615, 25363, 298942, 64304, 25756, 50682}	33	2	6.060606	2.517128
WP372					

```
[33]: SigPathways = list(WPtable.index)
for WP in SigPathways:
    print(WP + '\t' + WikiPathwaysNames[WP])
print('\nBased on dataset(s): ')
display(res_df)
```

WP1286 Metapathway biotransformation
 WP1307 Fatty Acid Beta Oxidation

WP506 Fatty Acid Beta Oxidation
 WP1297 Retinol metabolism
 WP419 Mitochondrial LC-Fatty Acid Beta-Oxidation
 WP139 Nuclear receptors in lipid metabolism
 and toxicity
 WP145 Statin Pathway
 WP2376 Nuclear factor, erythroid-derived 2,
 like 2 signaling pathway
 WP372 Beta Oxidation Meta Pathway

Based on dataset(s):

id	version	dataset	Compound	InChI Key	CAS	Organism	Organ	Study type	Dose	Dosing	Duration
e45c5706-010e-4597-b490-df1-38f57f06e	1	<PublishedDataset 'e45gemfibrozil_Rat_Liver_in_vivo_Repeat_8_day_high_FOLD_CHANGES>	gemfibrozil	HEMJJKBWTPKOJG-UHFFFAOYSA-N	25812-30-0	Rat	Liver	in_vivo	high	Repeat	8day

```
[28]: %load_ext watermark

#python, ipython, packages, and machine characteristics
%watermark -v -m -p sys, pip, SPARQLWrapper, pandas, json, \
↳re, requests, warnings, pyvis, matplotlib, numpy, \
↳IPython, urllib, seaborn, statistics

#dte
print(" ")
%watermark -u -n -t -z
```

CPython 3.6.3

IPython 7.9.0

sys 3.6.3 (default, May 31 2019, 13:05:43)

[GCC 4.8.5 20150623 (Red Hat 4.8.5-36)]

pip 20.0.1

SPARQLWrapper 1.8.5

pandas 0.25.3

json 2.0.9

re 2.2.1

requests 2.22.0

warnings unknown

pyvis 0.1.7.0

matplotlib not installed

numpy 1.18.1

IPython 7.9.0

```
urllib unknown
seaborn not installed
statistics unknown
```

```
compiler   : GCC 4.8.5 20150623 (Red Hat 4.8.5-36)
system     : Linux
release    : 3.10.0-1062.1.2.el7.x86_64
machine    : x86_64
processor  : x86_64
CPU cores  : 60
interpreter: 64bit
```

last updated: Thu Jan 23 2020 14:00:14 UTC

```
[31]: sparqlquery = '''
      SELECT ?originaldata
      WHERE{
        ?dataset a void:Dataset ;
        pav:createdWith ?originaldata .
      }'''
      aopwikisparql.setQuery(sparqlquery)
      aopwikisparql.setReturnFormat (JSON)
      results = aopwikisparql.query().convert()

      Result = results["results"]["bindings"][0]
               ['originaldata']['value']

      print("The underlying dataset for AOP-Wiki RDF:  "+ Result)
```

The underlying dataset for AOP-Wiki RDF: [aop-wiki-xml-2019-07-01](#)

```
[35]: sparqlquery = '''
      select distinct ?dataset ?date where {
        ?dataset a void:Dataset ;
        pav:createdOn ?date .
      }'''
      wikipathwayssparql.setQuery(sparqlquery)
      wikipathwayssparql.setReturnFormat (JSON)
      results = wikipathwayssparql.query().convert()

      Result = results["results"]["bindings"][0]
```

```

['dataset']['value']
Date = results["results"]["bindings"][0]
['date']['value']

print("The WikiPathways RDF used in this notebook: "+Result+
↪ " created on "+ Date)

```

The WikiPathways RDF used in this notebook:
<http://data.wikipathways.org/20191210/rdf/> created on
↪ 2019-12-09T23:28:23.591Z

References

- [1] Paul Jennings et al. *Finalization of case studies and analysis of remaining weaknesses (Deliverable 1.5)*. Mar. 2020. DOI: 10.5281/zenodo.3693636.
- [2] Elisabet Berggren et al. "Ab initio chemical safety assessment: A workflow based on exposure considerations and non-animal methods". *Computational Toxicology* 4 (Nov. 2017), pp. 31–44. DOI: 10.1016/j.comtox.2017.10.001.
- [3] Gerald T. Ankley et al. "Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment". *Environmental Toxicology and Chemistry* 29.3 (Mar. 2010), pp. 730–741. DOI: 10.1002/etc.34.
- [4] T. Burgdorf et al. "Workshop on the validation and regulatory acceptance of innovative 3R approaches in regulatory toxicology – Evolution versus revolution". *Toxicology in Vitro* 59 (Sept. 2019), pp. 1–11. DOI: 10.1016/j.tiv.2019.03.039.
- [5] Marvin Martens, Chris T. Evelo, and Egon L. Willighagen. "Providing Adverse Outcome Pathways from the AOP-Wiki in a Semantic Web Format to Increase Usability and Accessibility of the Content". *Applied in vitro toxicology* 8.1 (Mar. 2022), pp. 2–13. DOI: 10.1089/AIVT.2021.0010.
- [6] Holly M Mortensen et al. "Enhancing the EPA Adverse Outcome Pathway Database (AOP-DB): Recent Updates and Semantic Integration". *The Toxicologist* 174.1 (2020).
- [7] Martijn P van Iersel et al. "The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services". *BMC Bioinformatics* 11.1 (2010), p. 5. DOI: 10.1186/1471-2105-11-5.
- [8] Andra Waagmeester et al. "Using the Semantic Web for Rapid Integration of WikiPathways with Other Biological Online Data Resources". *PLOS Computational Biology* 12.6 (June 2016). Ed. by Christos A. Ouzounis, e1004989. DOI: 10.1371/journal.pcbi.1004989.
- [9] Marvin Martens et al. "WikiPathways: connecting communities". *Nucleic Acids Research* 49.D1 (Jan. 2021), pp. D613–D621. DOI: 10.1093/nar/gkaa1024.

7

Molecular Adverse Outcome Pathways: towards the implementation of transcriptomics data in risk assessments

Adapted from: Marvin Martens et al. "Molecular Adverse Outcome Pathways: towards the implementation of transcriptomics data in risk assessments". *bioRxiv* (Jan. 2023), p. 2023.03.02.530766. DOI: 10.1101/2023.03.02.530766.

Abstract

Adverse Outcome Pathways (AOPs) provide mechanistic insights into toxicological processes and can facilitate a potential replacement for animal studies with in vitro testing systems. However, the majority of Key Events (KEs) in AOP-Wiki describe molecular and cellular processes, which can be difficult to assess using current toxicology methods. Omics technologies, such as transcriptomics, offer a promising approach but have not yet been widely accepted for regulatory risk assessments due to their complexity and lack of consensus on standardization, analysis, and interpretation. In this paper, we propose the use of molecular AOPs to connect KEs of the AOP-Wiki with curated biological pathways in WikiPathways, enabling the analysis and interpretation of transcriptomics data to identify KE activation. We demonstrate the utility of this approach through case studies on liver steatosis and mitochondrial complex I inhibition, where molecular AOPs were developed and transcriptomics datasets were selected. By mapping and analyzing the data, we were able to verify the activation of specific MIEs and KEs and assess progression across the AOP. Our findings show that transcriptomics data can be used to identify the potential activation of KEs, but extensive datasets are required to fully test the capabilities of molecular AOPs. Additionally, linking molecular pathways and KEs can be challenging, and further refinement is necessary to optimize this approach. Despite these challenges, our results suggest that molecular AOPs have the potential to provide valuable insights into toxicological processes and improve the use of transcriptomics data in regulatory risk assessments.

7.1 Introduction

Adverse Outcome Pathways (AOPs) have become useful tools in risk assessments, as they provide a broad overview of events preceding a detrimental outcome. As such, they span multiple biological organization levels, from the molecular level to the effects on a whole organism or even population. These types of pathways consist of three main concepts: Molecular Initiating Events (MIEs), Key Events (KEs), and Adverse Outcomes (AO). KEs are also linked with one another through Key Event Relationships (KERs), from molecular interactions to population dynamics [1]. The KEs in an AOP are not always sufficient on their own to lead to an AO. However, the strength of the KE concept lies in that they are scientifically approved events that are essential to happen in the progression of the AOP. To be a well-defined AOP, the originating events from molecular interactions (i.e., MIE) should have a causal, measurable, and biologically plausible link towards an AO [2, 3]. With the evaluation of newly developed AOPs, one considers tailored Bradford-Hill criteria, in which the causality of observed association in epidemiological studies is determined [4]. This method considers biological plausibility, essentiality, and empirical support for these findings [4].

The practical aspect of AOPs is that they are an integral tool of risk assessment, where they act as scaffolds to collect and structure toxicological information at differing levels of biological organization, used to determine various apical AOs effectively after exposure to a stressor [1, 5]. This is also an essential part of Integrated Approaches to Testing and Assessment (IATA), which can include combinations of methods and integrating results of various types. AOPs can be used as a backbone to develop IATA [6].

However, current research is also focused on the quantification of AOPs in addition to the descriptive mechanisms, as risk assessors typically use numbers and thresholds to study the safety of substances and chemicals [7–10]. These quantitative AOPs can be developed from qualitative AOPs, providing the quantitative descriptors or

annotations for KEs and KERs, and can serve as predictors of adversity [7]. With the rise of genomic technologies and better biological system modelling, it is easier to hypothesize or develop new AOPs [11–14].

The value of transcriptomics data in toxicological research is promising, and the technologies to produce the data are getting cheaper, faster and protocols better established [15–17]. However, transcriptomics technologies are not widely implemented in risk assessments [15]. This is due to the challenges in assessing and interpreting transcriptomics, and hurdles that exist in the general acceptance of transcriptomics data. For example, there is no consensus on the standardisation, reproducibility and experimental setup of the technologies and whether these can be validated against other, well-established approaches for measuring gene expression changes. Furthermore, interpreting the data, distinguishing adaptive and adverse responses, investigating cause and consequence, and the amount of data required to make conclusions are challenges that remain to be solved for working with transcriptomic data [12, 14, 18]. Some of these issues, such as the distinguishing between adaptive, adverse, causative and consequential biological processes, are a key feature of AOPs and therefore help to make these distinctions based on our biological understanding.

For example, integrative approaches of publicly available data have been applied to develop AO networks of biological pathways and diseases which implement transcriptomics data [19], leading to its support of evidence in risk assessments. Also, the integration of transcriptomics data has been studied for adverse pulmonary effects [13, 20]. The Organisation for Economic Co-operation and Development (OECD) has recently published a formal reporting framework to tackle the challenges of transcriptomics and metabolomics technologies in risk assessment applications. These provide guidance on the execution and reporting of data generation, processing, analysis, methodology and metadata [21].

It is expected that the combination of molecular pathways and AOPs allows the use of transcriptomics data for the measurement of KEs in AOPs and AOP Networks [12, 22]. Whereas biomarker genes and proteins exist for the validation of KEs, we believe that more elaborate molecular pathways can provide more meaningful evidence of KE activation using transcriptomics data [23–25]. Also, the connection between AOPs and molecular pathways offers additional insights into the molecular mechanisms underlying the more general KEs [26] and supports the biological plausibility of KERs.

This paper will focus on two AOPs that are well-established and have been studied extensively for application in risk assessments. The first one is an AOP network of liver steatosis. This AOP is particularly well-studied AOP for applications of computational approaches and data integration, and can be initiated through interactions with various nuclear receptors [27–30]. The network is well-defined, and the receptor-specific MIEs allow the investigations of known agonists and antagonists of the receptors. This AOP network has been used to study a range of transcriptomics datasets of multiple chemicals known to activate specific MIEs [19, 31]. The second AOP that this paper will look into is the AOP of mitochondrial complex I inhibition leading to Parkinsonian motor deficits, part of a larger AOP network of mitochondrial inhibition leading to AOs in multiple organs including the brain, liver, and kidneys [32]. A wide array of well-known chemicals are known to interact with the electron transport chain or membrane potential and disturb the production of ATP inside the mitochondria [33, 34]. While multiple downstream KEs are described for this, the most important KEs central in the AOPs are oxidative stress, unfolded protein response, and induction of cell death, described as downstream effects of mitochondrial complex I inhibition. This also counts for the well-established AOP of Parkinsonian motor deficits caused by mitochondrial complex I inhibitors [35, 36].

This paper provides a new method of applying transcriptomics data analyses and interpretations in the framework of AOPs. By doing so, the implementation of such data in risk assessment studies can be fa-

cilitated. To illustrate the value of this method, two case studies are performed, including the liver steatosis AOP Network and the AOP of mitochondrial complex I inhibition AOP leading to Parkinsonian motor deficits.

7.2 Methods

7.2.1 Development of molecular AOP

To model molecular AOPs, PathVisio 3.3.0 [37] was used to draw and upload the molecular AOPs to WikiPathways [38], where they were tagged and stored in the AOP Portal (aop.wikipathways.org, Figure 7.1). For the two case studies presented in this manuscript, the AOP-Wiki and relevant literature were used to construct AOPs. For each KE of the AOP, corresponding molecular pathways were identified in WikiPathways or, when necessary, developed based on the available scientific literature. The molecular AOPs exist as chains of Key Event nodes where each KE was annotated with the corresponding AOP-Wiki KE identifiers if available, linked with directed interactions representing KERs. Attached to the Event nodes are molecular pathway nodes containing identifiers of existing pathways in WikiPathways, using undirected interactions (see Figure 7.2).

7.2.2 Datasets

To illustrate the implementation of gene expression data in molecular AOPs, publicly available datasets were selected in GEO [39] (see Table 7.1). Gene expression data of primary Human hepatocytes exposed to three Pregnane X Receptor (PXR) agonists from GEO dataset GEO:GSE90122 [40] was used in the case study of liver steatosis to explore and compare the effects of agonists of the PXR receptor, one of the MIEs of the liver steatosis AOP network. To explore the effect of time of exposure on gene expression in the liver steatosis case study, we also used data of HepaRG cells exposed to GW3965 (Liver X Receptor (LXR) agonist) from GEO dataset GEO:GSE123053 [41]. For a

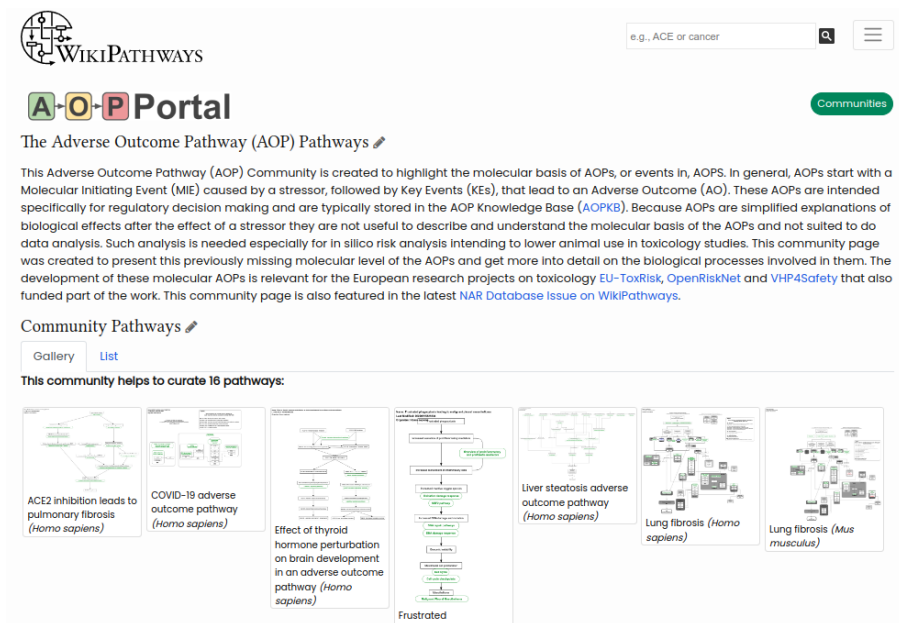


Figure 7.1: The AOP Portal on WikiPathways.org. On the portal, the AOP community can collaborate and comment on molecular AOPs and toxicity pathways.

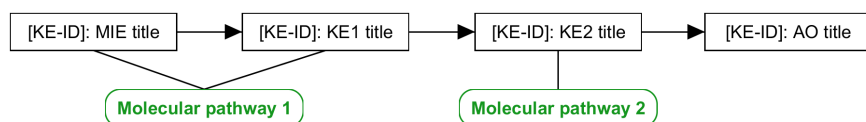


Figure 7.2: Conceptual illustration of a molecular AOP in WikiPathways. Black boxes are Key Events nodes, and green rounded nodes are Pathway nodes.

second case study of mitochondrial complex I inhibition, gene expression data from LUHMES cells (embryonic neuronal precursor cells) exposed to rotenone (mitochondrial complex I inhibitor) from the GEO dataset GEO:GSE116280 [42] was used to explore the effects of both

time and dose on gene expression using molecular AOP. All datasets originated from GEO and were processed with GEO2R [43] to generate Log2 Fold Change (Log2FC) values and perform statistical tests to generate p-values for all reads in the comparison of samples with chemical exposure and without chemical exposure. Next, a custom Jupyter notebook was executed to calculate the average Log2FC values for each individual gene and Fisher's combined probability test was used to calculate p-values for each gene.

Table 7.1: Overview of datasets used and detailed information on relevant samples.

GEO Accession number	Chemical	Cell type	Samples	Reference
GEO:GSE90122	Rifampicin PXR agonist	Primary Human Hepatocytes	1 μ M, 48h, n=3	Lin et al., 2017
GEO:GSE90122	SR12813 PXR agonist	Primary Human Hepatocytes	100nM, 48h, n=3	Lin et al., 2017
GEO:GSE90122	T0901317 LXR/PXR agonist	Primary Human Hepatocytes	30nM, 48h, n=3	Lin et al., 2017
GEO:GSE124053	GW3965 LXR agonist	HepaRG	2 μ M, 4h / 24h, n=3	Wigger et al., 2019
GEO:GSE116280	Rotenone Mitochondrial Complex I inhibitor	LUHMES	50nM / 100nM, 12h / 24h, n=3	Harris et al., 2018

7.2.3 Cytoscape for AOP Network and data visualisation

In Cytoscape (version 3.8.2), the WikiPathways app (version 3.3.7) is used to import the molecular AOPs from WikiPathways as networks. Next, the network was extended with gene identifiers using the CyTargetLinker app (version 4.1.0) [44] with the WikiPathways Linkset (downloaded from cytargetlinker.github.io/pages/linksets/wikipathways, version 20210110). A custom visualisation was applied to distinguish Key Event nodes (orange diamonds) and pathway nodes (yellow squares).

Next, all datasets were imported into Cytoscape, and data were visualised on the gene nodes. These were colored with a blue-white-red gradient to represent the Log2FC values, where blue and red colors indicate the down- and upregulation respectively and white indicates no

change. Node borders are highlighted in green for significantly altered expression levels (p-value < 0.05).

With the data visualised, the remaining genes without available data were removed for clarity and interpretation, although these could also inform which genes were missing in the dataset. Also, in the case of the steatosis AOP network, the AOP was trimmed from irrelevant KEs which were not directly involved with the chemicals that were studied.

7.2.4 Scoring Key Events

To quantify and assess the activation of KEs, the Enrichment Score (ES) is calculated based on the number of significantly affected genes present in molecular pathways linked to each of the KEs in the AOP against all measured genes linked to each KE, compared to the whole data set which was taken as the background. This is done using the formula

$$Enrichment\ Score = \frac{Target\ \%}{Background\ \%} = \frac{\frac{b}{B}}{\frac{n}{N}}$$

where:

n = number of differentially expressed genes in pathways linked to KE

N = total number of genes in pathways linked to KE

b = number of differentially expressed genes in the whole dataset

B = total number of genes in the whole dataset

Furthermore, a hyper-geometric p-value was calculated to indicate the statistical significance (p-value < 0.05) of the enriched pathways (ES > 1). The p-value was calculated using the formula

$$P(x) = \frac{\frac{n}{b} * \frac{N-n}{B-b}}{\frac{N}{B}}$$

where:

n = number of differentially expressed genes in pathways linked to KE

N = total number of genes in pathways linked to KE

b = number of differentially expressed genes in the whole dataset

B = total number of genes in the whole dataset

7.2.5 Pathway data visualisation

For further exploration of the biological roles of differentially expressed genes, PathVisio was used to visualise the data on the molecular pathways linked to activated KEs. The same as for the molecular AOP in Cytoscape, the Log2FC values were visualised using a blue-white-red gradient in the left side of nodes where blue and red colors indicate the down- and upregulation respectively. Similarly, significance was visualised by a bright green color on the right side.

7.3 Results

In order to analyse the transcriptomic data in the case studies, molecular AOPs were developed for the AOP network of hepatic steatosis caused by multiple MIEs ([wikipathways:WP4010](https://www.ehponline.org/doi/10.1371/journal.pone.0141010), [identifiers.org/wikipathways:WP4010](https://www.ehponline.org/doi/10.1371/journal.pone.0141010)) and the AOP of mitochondrial complex I inhibition leading to Parkinsonian motor deficits ([wikipathways:WP4945](https://www.ehponline.org/doi/10.1371/journal.pone.0141010), [identifiers.org/wikipathways:WP4945](https://www.ehponline.org/doi/10.1371/journal.pone.0141010)).

7.3.1 Case study 1: liver steatosis

The molecular AOP network for liver steatosis consists of 30 KE nodes and 20 molecular pathways. In total, 578 unique genes were mapped to 20 of the KEs through the molecular pathways. The most prevalent gene in KEs of the AOP network is *RXRA*, being part of 8 of the molecular pathways. It codes for the Retinoid X Receptor, one of the nuclear receptors involved in MIEs in the AOP network of liver steatosis.

For the datasets of PXR agonists, the PXR section of the liver steatosis AOP network was used for data visualisation and calculation of the KE enrichment scores (Figures 7.3, 7.4, and 7.5). In this subnetwork, eleven KEs were linked to seven molecular pathways, containing 150 unique genes that were measured in the datasets of PXR agonists. The KE with the highest relative number of genes with significantly affected expression levels across the datasets of PXR agonists is the KE245: "PXR activation" and corresponding pathway wikipathways:WP2876, with thirteen out of 28 genes (48%) on average having significantly changed expression levels after exposure to one of the PXR agonists (Figure 7.6). This KE also has the highest enrichment scores of 12.83, 9.29, and 6.31 after exposure to Rifampicin, SR12813 and T0901317, respectively (Table 7.2). The downstream KEs do not show consistency among the three PXR agonist data sets. The dataset of T0901317 exposure notably shows a significant enrichment of most KEs except the on fatty acid lysis. Also of interest is the significant overexpression of most of the prevalent (hub) genes of the network, including *ACSL1*, *ACSL3*, *ACSL4*, *FASN*, and *ACACA*, linked to at least three of the KEs in the network.

Table 7.2: Enrichment Scores for KEs by PXR agonists. Significance is indicated with an asterisk.

Key Event	Pathway	Rifampicin	SR12813	T0901317
KE245: PXR activation	WP2876: Pregnane X Receptor pathway	12.83*	9.29*	6.31*
KE471: FoxA2 inhibition	WP5066: Foxa2 Pathway	4.79*	2.89	3.31*
KE474: HMGCS2 Down Regulation	WP4718: Cholesterol metabolism	0.53	3.85*	2.94*
KE860: Decreased Mitochondrial Fatty Acid Beta Oxidation	WP368: Mitochondrial LC-Fatty Acid Beta-Oxidation	-	3.40	3.25*
KE115: Increased FA Influx	WP5061: Fatty acid transporters	-	-	3.07*
KE89: De Novo FA Synthesis	WP357: Fatty Acid Biosynthesis	-	2.63	5.01*
Fatty Acid Lysis	WP3965: Lipid Metabolism Pathway	0.83	1.00	1.90

Next to the PXR agonists, the second dataset was used to explore the

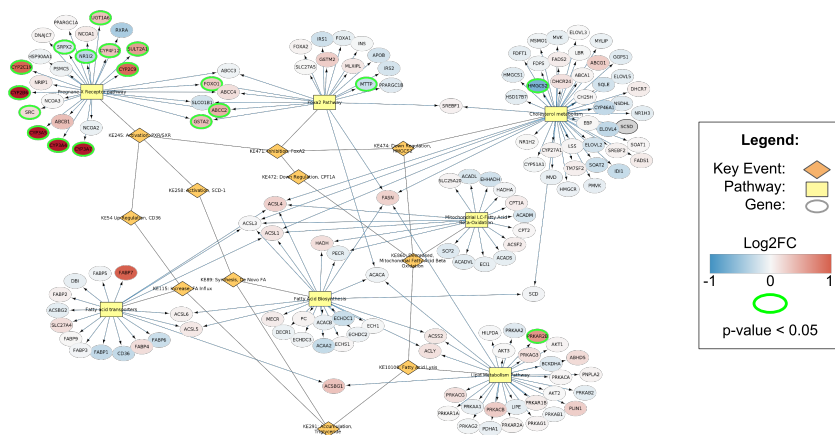


Figure 7.3: Gene expression data after Rifampicin exposure visualised on PXR AOP. Red and blue indicate up- and downregulation of gene expression (Log2FC), respectively. Green borders indicate significance (p -value < 0.05).

effects of the LXR agonist GW3965 at different time points on the liver steatosis AOP. The AOP network starts with the MIE of LXR activation and includes all downstream KEs, leading to a new AOP network of six KEs, of which four were linked to five molecular pathways. In total, 226 unique genes were part of the network and were measured in the dataset. The gene that is involved in most KEs was *FASN*, which is present in all five pathways, followed by *SCD* and *SREBF1* which were part of four of the molecular pathways (Figures 7.7 and 7.8). A major difference between the time points is the number of significantly altered genes across the AOP, where nine genes are differentially expressed at four hours of exposure, and 27 genes at 24 hours of exposure. At four hours, the majority of the gene expression changes happen to the highly connected genes.

From the data on the LXR AOP, it can be noted that the MIE of LXR activation and the KE of Fatty Acid Biosynthesis were significantly

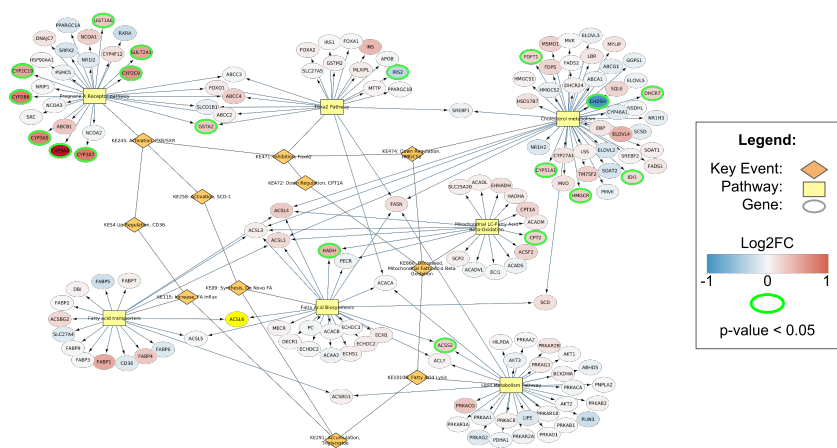


Figure 7.4: Gene expression data after SR12813 exposure visualised on PXR AOP. Red and blue indicate up- and downregulation of gene expression (Log2FC), respectively. Green borders indicate significance (p-value < 0.05).

enriched at both time points (Table 7.3). However, the largest KE of SREBP-1c activation, linking to 113 genes, is only significantly enriched at 24h of exposure to GW3965. This difference is most visible in the SREBP pathway, whereas the AMPK pathway does not differ between the time points.

Table 7.3: Enrichment Scores of KEs after exposure to GW3965 (LXR agonist). Significance is indicated with an asterisk

Key Event	Pathway(s)	GW3965 4h	GW3965 24h
KE167: LXR activation	WP2874: Liver X Receptor Pathway	12.31*	11.46*
KE66: ChREBP activation	WP3915: Angiopoietin Like Protein 8 Regulatory Pathway	1.72	1.60
KE264: SREBP-1c activation	WP1982: SREBP signaling WP1403: AMPK signaling	1.63	3.88*
KE89: De Novo FA synthesis	WP357: Fatty Acid Biosynthesis	6.71*	11.72*

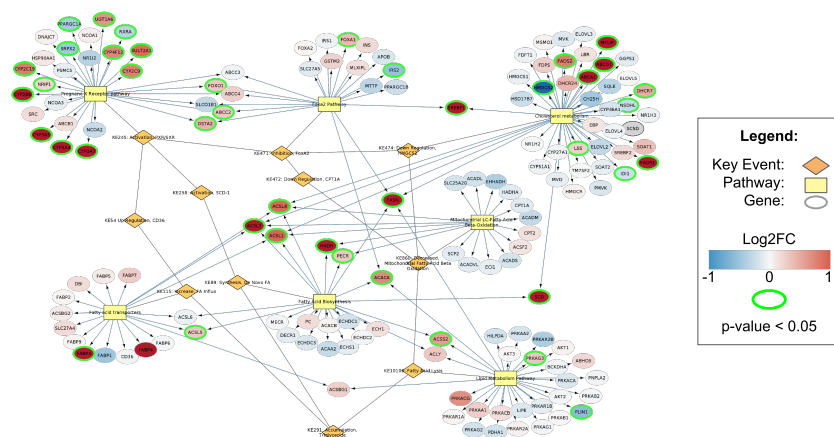


Figure 7.5: Gene expression data after T0901317 exposure visualised on PXR AOP. Red and blue indicate up- and downregulation of gene expression (Log2FC), respectively. Green borders indicate significance (p -value < 0.05).

7.3.2 Case study 2: mitochondrial inhibition

The second case study was focused on the AOP of mitochondrial complex I inhibition leading to Parkinsonian motor deficits. The developed molecular AOP contains seven KE nodes and seven pathway nodes. When extended using CyTargetLinker, a total of 199 gene nodes are added to the molecular AOP network and are measured in the datasets of Rotenone exposure to LUHMES cells. Of all the genes in this network, approximately 25% show significantly altered gene expression levels upon exposure to Rotenone.

The datasets for rotenone on LUHMES cells were visualised and they show the significantly altered gene expression for all KEs, where 50nM dose exhibits a stronger effect in the early KEs of mitochondrial complex I and oxidative phosphorylation, and the 100nM dose causes a higher number of gene expression changes in the AO of Parkinsonian motor deficits (Figures 7.10 and 7.11). The only KEs that are signifi-

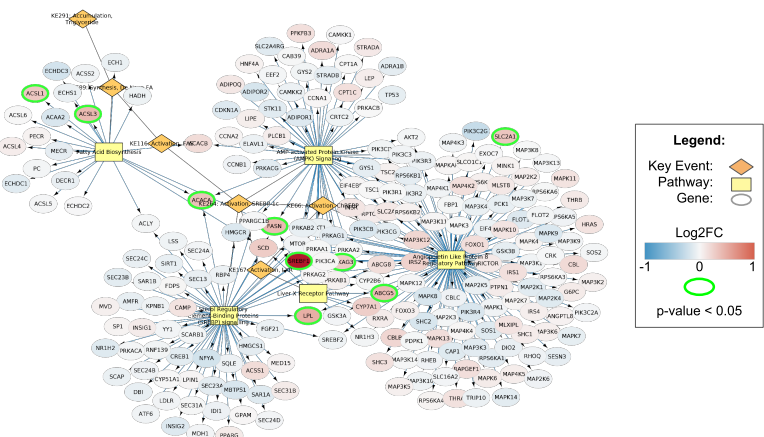


Figure 7.7: Gene expression data after 4 hour exposure to GW3965 (LXR agonist) visualised on LXR AOP. Red and blue indicate up- and down-regulation of gene expression (Log2FC), respectively. Green borders indicate significance (p-value < 0.05).

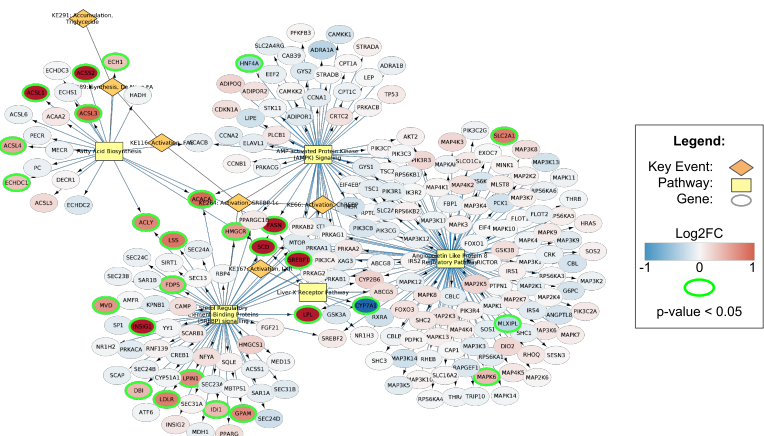


Figure 7.8: Gene expression data after 24 hour exposure to GW3965 (LXR agonist) visualised on LXR AOP. Red and blue indicate up- and down-regulation of gene expression (Log2FC), respectively. Green borders indicate significance (p-value < 0.05).

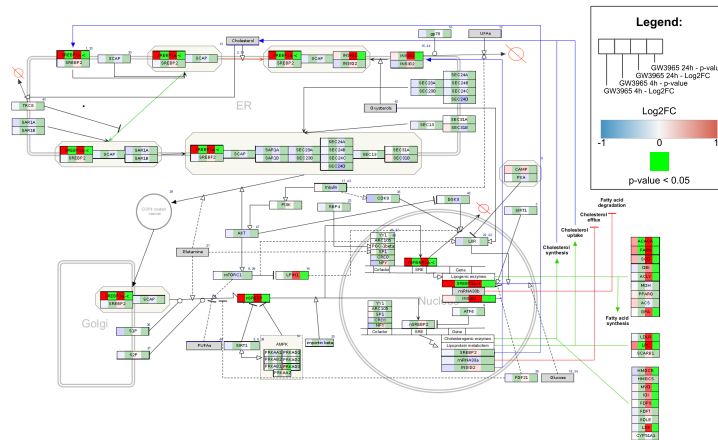


Figure 7.9: Gene expression data after exposure to GW3965 (LXR agonist) visualised on the SREBP signaling pathway (wikipathways:WP1982). The left side of each data node represents data from 4h exposure, and the right side represents the data from 24h exposure. Red and blue indicate the up- and down-regulation of gene expression (Log2FC). Bright green marks indicate significance (p-value < 0.05).

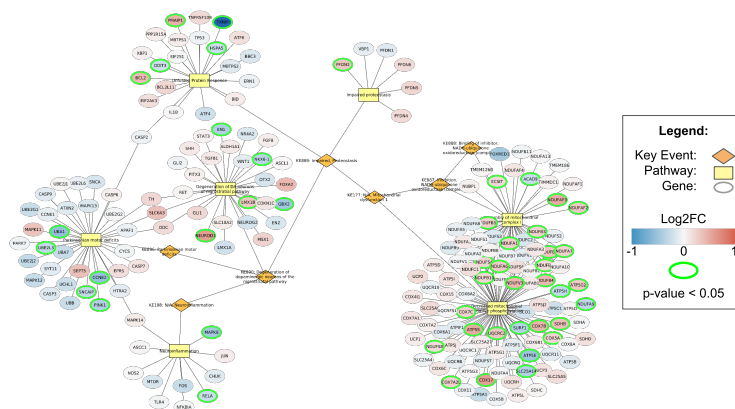


Figure 7.10: Rotenone (50nM) data visualised on mitochondrial complex I inhibition AOP. Red and blue indicate the up- and downregulation of gene expression (Log2FC). Bright green marks indicate significance (p-value < 0.05)

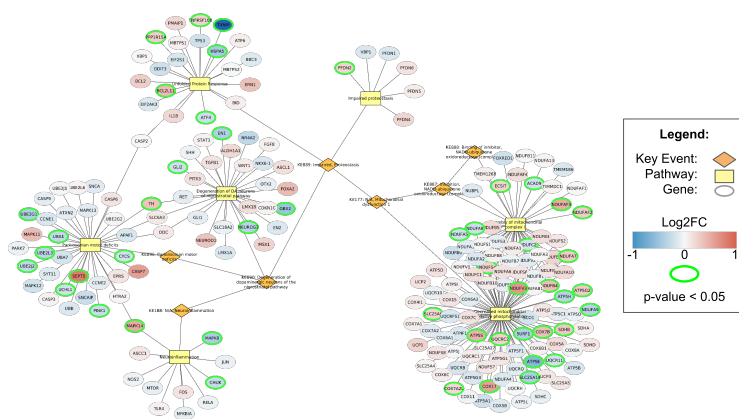


Figure 7.11: Rotenone (100nM) data visualised on mitochondrial complex I inhibition AOP. Red and blue indicate the up- and downregulation of gene expression (Log2FC). Bright green marks indicate significance (p-value < 0.05)

7.4 Discussion

With this work, we have created the molecular counterparts of AOPs, tightly linked to matching AOPs in the AOP-Wiki through the linking of molecular pathways explicitly to KEs. The addition of molecular pathways to KEs to analyse transcriptomic data allows a broader sense of the effects of toxicants when compared to biomarkers for process activation, which most often provide limited information and are used to measure a single KE. Therefore, these molecular AOPs can be used to get a more thorough understanding of traditional AOPs by exploring molecular pathway models, but they also provide a method of visualising and analysing omics datasets, which was explored in this manuscript.

When looking at gene expression changes upon exposure to a toxicant, the predictive value of each individual gene in the network could differ. This is where pathway models can provide additional insights into the overall connectivity of genes and their roles in pathways [45, 46].

Table 7.4: Enrichment Scores for KEs by exposure to Rotenone at 50nM and 100nM. Significance is indicated with an asterisk

Key Event	Pathway(s)	Rotenone 50nM	Rotenone 100nM
KE888: Binding of inhibitor, Complex I	WP4324: Assembly of mitochondrial complex I	1.38*	0.93
KE887: Inhibition, Complex I	WP4324: Assembly of mitochondrial complex I	1.38*	0.93
KE177: Mitochondrial dysfunction	WP111: Electron transport chain	1.21	0.90
KE889: Impaired, Proteostasis	WP4918: Cellular proteostasis WP4925: Unfolded protein response	0.94	1.01
KE890: Degeneration of dopaminergic neurons	WP2855: Dopaminergic neurogenesis	0.73	0.67
KE188: Neuroinflammation	WP4919: Neuroinflammation	0.71	0.97
KE896: Parkinsonian s motor deficit	WP2371: Parkinson's disease pathway	0.59	1.08

The first case study was focused on the liver steatosis AOP, a widely studied and well-established AOP [27, 28, 47, 48], starting with a range

of nuclear receptors known to be involved in maintaining lipid balance in the liver, including LXR, PXR, PPAR, and CAR, among others [48]. With the case study, we showed that molecular AOPs can be used to visualise transcriptomic datasets and perform enrichment analyses of KEs. Based on the results, we can clearly identify the activated MIEs and make comparisons between chemicals or exposure scenarios. For example, the dataset of GW3965 could be used to investigate the differences in gene expressions at two time points. By performing an enrichment analysis, many processes can be assessed simultaneously and generate hypotheses of KE activation.

Although the molecular AOP can highlight which KEs are affected based on pathway-level gene expression changes, the mapping of KEs to molecular pathways does not always fill its purpose. Some of the KEs in the liver steatosis AOP network are described in AOP-Wiki as single gene expression changes, rather than processes being affected. For example, the expression level of *HMGCS2*, which plays an essential role in cholesterol metabolism and ketogenesis [49], is significantly decreased after exposure to T0901317 and Rifampicin. However, since the focus of our analysis lies on the molecular pathways to expand our biological understanding, the KE is only regarded as significantly enriched within the dataset of T0901317 exposure. This can be a limitation for KEs that have a clear transcriptional marker gene or gene set, or KEs that are confined to single gene expression changes rather than processes, when compared to our KE enrichment calculation without taking the size of changes into account. This distinction of single markers and pathways would also be important to address for the KE of SCD-1 (SCD) activation, which plays a role in the regulation of energy metabolism and lipid synthesis, and has significantly increased expression levels after exposure to T0901317. In the molecular AOP network, however, it is linked to the Cholesterol metabolism and Fatty Acid Biosynthesis pathways, and not the specific KE of SCD-1 activation. The data shows that only T0901317 exposure led to the enrichment of the majority of KEs whereas the other PXR agonists only led to the enrichment of a handful of KEs. This could be due to the described dual

agonistic role of T0901317, as it is also an LXR agonist [50], affecting the downstream KEs through multiple pathways.

On the other hand, the KEs of *CD36* upregulation, being a fatty acid transporter [51], does not show up as significant in the gene expression data but it is part of the Fatty Acid Transporters pathway which is significantly enriched after T0901317 exposure. This is also the case for the KE of *CPT1A* downregulation which does not show up in the data of the PXR agonists. However, it is part of the pathway linked to the downstream KE of Decreased Mitochondrial Fatty Acid Beta Oxidation, which is significantly enriched after T0901317 exposure.

As discussed, there can be value in regarding single transcriptional marker genes to investigate KE activation. Since individual or groups of (computed) transcriptional biomarkers can provide great insights into the activation of individual processes [52], further developments of this approach should focus on combining the molecular AOPs with stress response marker genes relevant to individual KEs. The value of using biomarkers or defined gene sets to explore the effects of toxicants and xenobiotics has been shown in various studies related to AOPs and KE activation. For example, well-described transcription factor modulations can be explored with predictive gene sets [23] or well-established transcriptional biomarkers based on responsive transcription factors [53]. An approach to combine the well-studied and carefully selected expression biomarkers for KE activation and molecular pathways can be most informative to validate the activation of KEs and understand the biological responses at the cellular level in more detail. Alternatively, a ranking gene set enrichment method can be used to take advantage of not always knowing the biomarkers of particular processes.

With the case studies presented in this work, we limited the investigation to AOPs that involve only a single cell type. While that can be the case for sequential KEs that are mostly on the molecular, cellular or tissue level, KEs can also involve cellular communication, such as the secretion of signalling molecules or recruitment of inflammatory cells, as

inflammation plays an important role in toxic adverse outcomes [54]. For such AOPs to be fit for use as molecular AOP, multiple datasets or single-cell transcriptomic data would be required to assess the KEs across cell types, making the overall assessment more complex. However, the flexible nature of WikiPathways, the identifier handling by BridgeDb and the integration of these tools in Cytoscape facilitate the integration of data and the model.

Based on our analyses and calculations of enrichment scores, the case studies provide a great insight into the usability and value of transcriptomics data within the molecular AOPs, showing the potential activations of KEs. With the visualisation and enrichment calculations, this method provides a quick overview of the overall activation of pathways that are linked to the KEs of interest. Also, it shows the interplay between processes within the AOP, highlighting central, highly connected genes, whose role can be further explored in the molecular pathways in which they exist. This is for example clearly visible in the PXR AOP, with various members of the ACSL family being involved in multiple KEs, as well as *FASN* being the most connected gene in the network. A significant alteration of their expression levels is expected to have a stronger impact on the overall assessment within the AOP when compared to genes that are part of only a single pathway or KE.

With the case study on liver steatosis, we focused on highly specific MIEs with well-studied stressors, and these show the activation of MIE-linked pathways based on gene expression data, which was consistent across the chemicals that we investigated for the different MIEs. Furthermore, with the multiple time points in the dataset for stressor GW3965, it is clear that the exposed cells progress through the KEs of the AOP towards the AO, which is promising for the application of time series exposure data on molecular AOPs. However, more extensive datasets with additional time points and doses would be required to assess the progression through the AOP based on gene expression data.

The case study of mitochondrial complex I inhibition by rotenone is

showing the challenges of interpreting transcriptomics data and variations in dose-response data. While in low-dose exposure we see an abundant upregulation of genes involved at the early KEs to counter the initial inhibition of mitochondrial complex I, this is not as clear in the higher-dose exposure. This suggests a switch in response from recovery towards adverse, but this is not clearly visible in the late KEs which represent stress response processes, none of which have been affected significantly. With the current calculation of KE enrichment, the whole dataset is taken as the background data, of which approximately 25% had significantly altered gene expression levels. Since the datasets of the liver steatosis case study contain between 3% and 9% of significantly affected genes, it could be that the disturbance of cellular energy production causes many more processes to be affected.

This relates to our approach to developing the molecular AOPs because we limit ourselves to the known KEs and do not include additional processes or responsive (feedback) pathways in the molecular AOP model. This constitutes a challenge in our approach, in which the pathway's level of detail and focus originate from their initial creation by the many contributors [55]. Since the method for creating the molecular AOP involves the inclusion of existing pathways developed by the community, curation might be necessary to ensure the expected quality of pathways to support KEs and to explore the outputs of the analyses. However, an enrichment evaluation of the whole WikiPathways database and exploring gene and pathway interactions with the AOP can potentially be used to identify missing KEs of the AOP.

Another challenge lies in the distinction between causative and responsive processes to the toxicity of stressors and understanding their sequence based on gene expression changes. As an example where this was possible, the MIEs of the steatosis AOP are nuclear receptor activation pathways that cause downstream effects upon their activation or inhibition, and clearly show transcriptional changes based on the exposure to stressors. However, the effects on pathways linked to downstream KEs are much more subtle and respond to the changes that occur because of the activated MIE. On the other hand, the case

study of mitochondrial inhibition does not provide a clear-cut activation of a pathway related to the MIE but instead consists mostly of responsive pathways to adapt to the new situation caused by the stressor. To develop useful molecular AOPs, one needs to be aware of the expected effects on the transcriptional level, which does not always match the KE description in the AOP-Wiki, where feedback loops and context are less present.

Whereas some KEs describe the activation of cellular responses and processes and are therefore easily linked to their molecular pathways in WikiPathways, other KEs can be more simple or more complex. For example, KEs can merely describe individual molecular interactions such as receptor activation, or describe the larger, more general interplay of processes, such as cell death where the measurement is focused on cellular viability. This varying level of complexity should automatically also be represented in molecular AOPs.

Based on these results, taking into account the limitations, we find that we can explain the biological plausibility of KE activation by visualizing experimental transcriptomics data to the molecular pathways underlying the KEs. The extension of AOPs with molecular markers, gene sets, or pathways would be an essential step towards the integration of high-throughput transcriptomics data into risk assessment studies. With these technologies getting cheaper, faster, and more reliable, they are becoming more frequently used in toxicological research. Hence, a framework to connect the established AOPs on AOP-Wiki with molecular entities through biological pathways is the logical intermediate.

Work is already ongoing to expand the molecular AOP contents and on validating their utility through additional case studies and comparing the outcomes to other methods to measure KE activation using gene expression data, such as the TXGMAPr tool [56] or AOP fingerprints [57]. When new biological mechanisms are resolved, these can be converted into molecular pathways and associated with KEs for which that was not known so far. Second, by introducing more spe-

cific ontological annotations of AOP-Wiki content, new literature can be discovered, allowing a dynamic process of describing the biology behind the AOP. Furthermore, by improving annotations of AOP content and making those annotations available through RDF [58], the automatic creation of molecular AOPs based on AOP-Wiki contents will be possible. Validation of the method is required by comparing it with other transcriptomic-based analysis methods or other *in vitro* methods that measure KE activation. This could be done through more extensive case studies on well-established AOPs in the AOP-Wiki.

References

- [1] Gerald T. Ankley et al. "Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment". *Environmental Toxicology and Chemistry* 29.3 (Mar. 2010), pp. 730–741. DOI: 10.1002/etc.34.
- [2] Daniel L. Villeneuve et al. "Adverse outcome pathway (AOP) development I: strategies and principles." *Toxicological Sciences* 142.2 (Dec. 2014), pp. 312–320. DOI: 10.1093/toxsci/kfu199.
- [3] Mathieu Vinken. "The adverse outcome pathway concept: A pragmatic tool in toxicology". *Toxicology* 312.1 (Oct. 2013), pp. 158–165. DOI: 10.1016/j.tox.2013.08.011.
- [4] Richard A. Becker et al. "Increasing Scientific Confidence in Adverse Outcome Pathways: Application of Tailored Bradford-Hill Considerations for Evaluating Weight of Evidence". *Regulatory Toxicology and Pharmacology* 72.3 (Aug. 2015), pp. 514–537. DOI: 10.1016/j.yrtph.2015.04.004.
- [5] Cameron MacKay et al. "From pathways to people: applying the adverse outcome pathway (AOP) for skin sensitization to risk assessment". *ALTEX* 30.4 (Nov. 2013), pp. 473–486. DOI: 10.14573/altex.2013.4.473.
- [6] Knut Erik Tollefsen et al. "Applying Adverse Outcome Pathways (AOPs) to support Integrated Approaches to Testing and Assessment (IATA)". *Regulatory Toxicology and Pharmacology* 70.3 (Dec. 2014), pp. 629–640. DOI: 10.1016/j.yrtph.2014.09.009.
- [7] S. Jannicke Moe et al. "Quantification of an Adverse Outcome Pathway Network by Bayesian Regression and Bayesian Network Modeling". *Integrated Environmental Assessment and Management* 17.1 (Jan. 2021), pp. 147–164. DOI: 10.1002/ieam.4348.
- [8] Edward J. Perkins et al. "Building and Applying Quantitative Adverse Outcome Pathway Models for Chemical Hazard and Risk Assessment". *Environmental Toxicology and Chemistry* 38.9 (Sept. 2019), pp. 1850–1865. DOI: 10.1002/etc.4505.
- [9] Nicoleta Spinu et al. "Quantitative adverse outcome pathway (qAOP) models for toxicity prediction". *Archives of Toxicology* 94.5 (May 2020), pp. 1497–1510. DOI: 10.1007/s00204-020-02774-7.

- [10] Gavin Maxwell et al. "Applying the skin sensitisation adverse outcome pathway (AOP) to quantitative risk assessment". *Toxicology in Vitro* 28.1 (Feb. 2014), pp. 8–12. DOI: 10.1016/j.tiv.2013.10.013.
- [11] Mathieu Vinken. "Omics-based input and output in the development and use of adverse outcome pathways". *Current Opinion in Toxicology* 18 (Dec. 2019), pp. 8–12. DOI: 10.1016/J.COTOX.2019.02.006.
- [12] Erica K. Brockmeier et al. "The Role of Omics in the Application of Adverse Outcome Pathways for Chemical Risk Assessment". *Toxicological Sciences* 158.2 (Aug. 2017), pp. 252–262. DOI: 10.1093/toxsci/kfx097.
- [13] Penny Nymark et al. "A Data Fusion Pipeline for Generating and Enriching Adverse Outcome Pathway Descriptions". *Toxicological Sciences* 162.1 (Mar. 2018), pp. 264–275. DOI: 10.1093/toxsci/kfx252.
- [14] Roland Buesen et al. "Applying 'omics technologies in chemicals risk assessment: Report of an ECETOC workshop". *Regulatory Toxicology and Pharmacology*. Vol. 91. Academic Press, Dec. 2017, S3–S13. DOI: 10.1016/j.yrtph.2017.09.002.
- [15] Ursula G. Sauer et al. "The challenge of the application of 'omics technologies in chemicals risk assessment: Background and outlook". *Regulatory Toxicology and Pharmacology* 91 (Dec. 2017), S14–S26. DOI: 10.1016/j.yrtph.2017.09.020.
- [16] Heidrun Ellinger-Ziegelbauer and Hans-Juergen Ahr. "Omics in Toxicology". *Regulatory Toxicology*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2020, pp. 1–13. DOI: 10.1007/978-3-642-36206-4_40-2.
- [17] Gemma L. D'Adamo, James T. Widdop, and Edward M. Giles. "The future is now? Clinical and translational aspects of "Omics" technologies". *Immunology and Cell Biology* 99.2 (Feb. 2021), pp. 168–176. DOI: 10.1111/IMCB.12404.
- [18] Hans Martin Kauffmann et al. "Framework for the quality assurance of 'omics technologies considering GLP requirements". *Regulatory Toxicology and Pharmacology* 91 (Dec. 2017), S27–S35. DOI: 10.1016/J.YRTPH.2017.10.007.
- [19] Noffisat O. Oki et al. "Integrated analysis of in vitro data and the adverse outcome pathway framework for prioritization and regulatory applications: An exploratory case study using publicly available data on piperonyl butoxide and liver models". *Toxicology in Vitro* 54 (Feb. 2019), pp. 23–32. DOI: 10.1016/j.tiv.2018.09.002.
- [20] Karolina Jagiello et al. "Transcriptomics-Based and AOP-Informed Structure–Activity Relationships to Predict Pulmonary Pathology Induced by Multiwalled Carbon Nanotubes". *Small* 17.15 (Apr. 2021), p. 2003465. DOI: 10.1002/smll.202003465.
- [21] Joshua A. Harrill et al. "Progress towards an OECD reporting framework for transcriptomics and metabolomics in regulatory toxicology". *Regulatory Toxicology and Pharmacology* 125 (Oct. 2021), p. 105020. DOI: 10.1016/j.yrtph.2021.105020.
- [22] Marvin Martens et al. "Introducing WikiPathways as a Data-Source to Support Adverse Outcome Pathways for Regulatory Risk Assessment of Chemicals and Nanomaterials". *Frontiers in Genetics* 9 (Dec. 2018), p. 661. DOI: 10.3389/fgene.2018.00661.
- [23] J. Christopher Corton. "Integrating gene expression biomarker predictions into networks of adverse outcome pathways". *Current Opinion in Toxicology* 18 (Dec. 2019), pp. 54–61. DOI: 10.1016/J.COTOX.2019.05.006.

-
- [24] Kirsten A. Baken et al. "A strategy to validate a selection of human effect biomarkers using adverse outcome pathways: Proof of concept for phthalates and reproductive effects". *Environmental Research* 175 (Aug. 2019), pp. 235–256. DOI: 10.1016/j.envres.2019.05.013.
- [25] Jin Wuk Lee et al. "Significance of adverse outcome pathways in biomarker-based environmental risk assessment in aquatic organisms". *Journal of Environmental Sciences* 35 (Sept. 2015), pp. 115–127. DOI: 10.1016/J.JES.2015.05.002.
- [26] Vinita Chauhan et al. "Bringing together scientific disciplines for collaborative undertakings: a vision for advancing the adverse outcome pathway framework". *International Journal of Radiation Biology* 97.4 (2021), pp. 431–441. DOI: 10.1080/09553002.2021.1884314.
- [27] Claudia Luckert et al. "Adverse Outcome Pathway-Driven Analysis of Liver Steatosis in Vitro: A Case Study with Cyproconazole". *Chemical Research in Toxicology* 31.8 (Aug. 2018), pp. 784–798. DOI: 10.1021/acs.chemrestox.8b00112.
- [28] Dajana Lichtenstein et al. "An adverse outcome pathway-based approach to assess steatotic mixture effects of hepatotoxic pesticides in vitro". *Food and Chemical Toxicology* 139 (May 2020), p. 111283. DOI: 10.1016/j.fct.2020.111283.
- [29] Lyle D. Burgoon et al. "Predicting the Probability that a Chemical Causes Steatosis Using Adverse Outcome Pathway Bayesian Networks (AOPBNs)". *Risk Analysis* 40.3 (Mar. 2020), pp. 512–523. DOI: 10.1111/RISA.13423.
- [30] Michelle M. Angrish et al. "Mechanistic toxicity tests based on an adverse outcome pathway network for hepatic steatosis". *Toxicological Sciences* 159.1 (2017), pp. 159–169. DOI: 10.1093/TOXSCI/KFX121.
- [31] Alejandro Aguayo-Orozco et al. "Analysis of Time-Series Gene Expression Data to Explore Mechanisms of Chemical-Induced Hepatic Steatosis Toxicity". *Frontiers in Genetics* 9.SEP (Sept. 2018), p. 396. DOI: 10.3389/FGENE.2018.00396.
- [32] Florentina Troger et al. "Identification of mitochondrial toxicants by combined in silico and in vitro studies – A structure-based view on the adverse outcome pathway". *Computational Toxicology* 14 (May 2020), p. 100123. DOI: 10.1016/J.COMTOX.2020.100123.
- [33] Joel N. Meyer et al. "Mitochondria as a Target of Environmental Toxicants". *Toxicological Sciences* 134.1 (July 2013), pp. 1–17. DOI: 10.1093/TOXSCI/KFT102.
- [34] Julie Eakins et al. "A combined in vitro approach to improve the prediction of mitochondrial toxicants". *Toxicology in Vitro* 34 (Aug. 2016), pp. 161–170. DOI: 10.1016/J.TIV.2016.03.016.
- [35] Andrea Terron et al. "An adverse outcome pathway for parkinsonian motor deficits associated with mitochondrial complex I inhibition". *Archives of Toxicology* 92.1 (Jan. 2018), pp. 41–82. DOI: 10.1007/s00204-017-2133-4.
- [36] Anna Bal-Price et al. "Adverse Outcome Pathway on Inhibition of the mitochondrial complex I of nigro-striatal neurons leading to parkinsonian motor deficits". 7 (2018). DOI: 10.1787/b46c3c00-en.
- [37] Martina Kutmon et al. "PathVisio 3: An Extendable Pathway Analysis Toolbox". *PLOS Computational Biology* 11.2 (Feb. 2015). Ed. by Robert F. Murphy, e1004085. DOI: 10.1371/journal.pcbi.1004085.
- [38] Marvin Martens et al. "WikiPathways: connecting communities". *Nucleic Acids Research* 49.D1 (Jan. 2021), pp. D613–D621. DOI: 10.1093/nar/gkaa1024.

- [39] Emily Clough and Tanya Barrett. "The Gene Expression Omnibus Database". *Methods in Molecular Biology*. Vol. 1418. Humana Press Inc., 2016, pp. 93–110. DOI: 10.1007/978-1-4939-3578-9_5.
- [40] Wenwei Lin et al. "SPA70 is a potent antagonist of human pregnane X receptor". *Nature Communications* 2017 8:1 8.1 (Sept. 2017), pp. 1–14. DOI: 10.1038/s41467-017-00780-5.
- [41] Leonore Wigger et al. "System analysis of cross-talk between nuclear receptors reveals an opposite regulation of the cell cycle by LXR and FXR in human HepaRG liver cells". *PloS one* 14.8 (Aug. 2019). DOI: 10.1371/JOURNAL.PONE.0220894.
- [42] Georgina Harris et al. "Toxicity, recovery, and resilience in a 3D dopaminergic neuronal in vitro model exposed to rotenone". *Archives of Toxicology* 2018 92:8 92.8 (June 2018), pp. 2587–2606. DOI: 10.1007/s00204-018-2250-8.
- [43] Tanya Barrett et al. "NCBI GEO: archive for functional genomics data sets—update". *Nucleic acids research* 41.Database issue (Jan. 2013). DOI: 10.1093/NAR/GKS1193.
- [44] Martina Kutmon et al. "CyTargetLinker app update: A flexible solution for network extension in Cytoscape". *F1000Research* 7 (Aug. 2019), p. 743. DOI: 10.12688/f1000research.14613.2.
- [45] Martina Kutmon et al. "WikiPathways: Capturing the full diversity of pathway knowledge". *Nucleic Acids Research* 44.D1 (2016), pp. D488–D494. DOI: 10.1093/nar/gkv1024.
- [46] Donny Soh et al. "Consistency, comprehensiveness, and compatibility of pathway databases". *BMC Bioinformatics* 11.1 (Dec. 2010), p. 449. DOI: 10.1186/1471-2105-11-449.
- [47] Mathieu Vinken. "Adverse Outcome Pathways and Drug-Induced Liver Injury Testing". *Chemical Research in Toxicology* 28.7 (July 2015), pp. 1391–1397. DOI: 10.1021/acs.chemrestox.5b00208.
- [48] Claire L. Mellor, Fabian P. Steinmetz, and Mark T. D. Cronin. "The identification of nuclear receptors associated with hepatic steatosis to develop and extend adverse outcome pathways". *Critical Reviews in Toxicology* 46.2 (Feb. 2016), pp. 138–152. DOI: 10.3109/10408444.2015.1089471.
- [49] Shaza Asif et al. "Hmgcs2-mediated ketogenesis modulates high-fat diet-induced hepatosteatosis". *Molecular Metabolism* 61 (July 2022), p. 101494. DOI: 10.1016/J.MOLMET.2022.101494.
- [50] Nico Mitro et al. "T0901317 is a potent PXR ligand: implications for the biology ascribed to LXR". *FEBS letters* 581.9 (May 2007), pp. 1721–1726. DOI: 10.1016/J.FEBSLET.2007.03.047.
- [51] Han Zeng et al. "CD36 promotes de novo lipogenesis in hepatocytes through INSIG2-dependent SREBP1 processing". *Molecular Metabolism* 57 (Mar. 2022), p. 101428. DOI: 10.1016/J.MOLMET.2021.101428.
- [52] Irmgard Riedmaier and Michael W. Pfaffl. "Transcriptional biomarkers – High throughput screening, quantitative verification, and bioinformatical validation methods". *Methods* 59.1 (Jan. 2013), pp. 3–9. DOI: 10.1016/J.YMETH.2012.08.012.
- [53] Paul Jennings. "Stress response pathways, toxicity pathways and adverse outcome pathways". *Archives of Toxicology* 87.1 (Jan. 2013), pp. 13–14. DOI: 10.1007/s00204-012-0974-4.

-
- [54] Linlin Chen et al. "Inflammatory responses and inflammation-associated diseases in organs". *Oncotarget* 9.6 (Jan. 2018), pp. 7204–7218. DOI: 10.18632/oncotarget.23208.
- [55] Saikat Chowdhury and Ram Rup Sarkar. "Comparison of human cell signaling pathway databases—evolution, drawbacks and challenges". *Database* 2015 (Jan. 2015), p. 126. DOI: 10.1093/DATABASE/BAU126.
- [56] Giulia Callegaro et al. "The human hepatocyte TXG-MAPr: gene co-expression network modules to support mechanism-based risk assessment". *Archives of Toxicology* 95.12 (Dec. 2021), pp. 3745–3775. DOI: 10.1007/S00204-021-03141-W.
- [57] Laura Aliisa Saarimäki et al. "Toxicogenomics Data for Chemical Safety Assessment and Development of New Approach Methodologies: An Adverse Outcome Pathway-Based Approach". *Advanced Science* 10.2 (Jan. 2023), p. 2203984. DOI: 10.1002/ADVS.202203984.
- [58] Marvin Martens, Chris T. Evelo, and Egon L. Willighagen. "Providing Adverse Outcome Pathways from the AOP-Wiki in a Semantic Web Format to Increase Usability and Accessibility of the Content". *Applied in vitro toxicology* 8.1 (Mar. 2022), pp. 2–13. DOI: 10.1089/AIVT.2021.0010.

8

Discussion

In current practice, risk assessments cannot keep up with the number of chemicals that require testing [1]. The risk assessment community aims to drive the shift from costly, time-consuming, ethically challenged *in vivo* experimentation on animals towards *in vitro* and computational methods to inform risk assessments about chemical risks and hazards and promote human safety [2]. However, the transition from traditional chemical risk assessment towards mechanism-based toxicity testing is slow because it relies on initiatives to develop and evaluate novel techniques to replace animal testing [3]. Additionally, the transition toward the implementation of those novel techniques in risk assessments and acceptance by regulators face barriers of validation, reproducibility, reliability, and overall confidence by the risk assessment community [4, 5].

In order to guide the paradigm shift, Adverse Outcome Pathways (AOPs) have been introduced to capture mechanistic knowledge about toxicological processes, which summarizes existing knowledge and organizes it in a set of measurable endpoints called Key Events (KEs) [6–9]. While there is an increased momentum for

AOPs towards application in risk assessments, the use of *in vitro* assays and in particular large-scale omics datasets face challenges of validation of their toxicological relevance and that hampers their inclusion in testing strategies [10]. While getting a lot of interest and having proven value in the exploration of toxicity pathways, the use of transcriptomic data in risk assessment is debated because of the complexity of analysing and interpretation of the data [10–12]. In order to resolve these challenges, we expect that the integration of biological databases containing biological knowledge with databases containing experimental data can improve the usability of existing data and scientific knowledge to generate hypotheses and guide risk assessment approaches.

It was hypothesized that we can use large-scale gene expression profile studies to explore the molecular modes of action of potential stressors. This can be followed by linking such observations to risk assessment endpoints using the established AOPs as templates to connect KEs with molecular pathways [13–16]. With the linking of molecular pathways with AOPs and the resulting potential to perform transcriptomic data analyses, we can use transcriptomic data to measure the activity of biological processes and therefore also KEs. This leads to the concept that if we can measure and visualise KE activation utilising transcriptomic data we could better support risk assessments. Consequently, we want to evaluate whether using the molecular structures of KEs strengthens the potential of transcriptomics data to become a viable resource for risk assessments. This thesis demonstrates the strength and utility of data integration with other databases and thereby facilitates transcriptomics data into AOP-based risk assessments. At the same time, by linking the AOPs to underlying biological processes, this integration also outlines how other types of biological data could be linked to KEs, which can facilitate the replacement of animal testing with transcriptomics-based AOP assessment.

8.1 WikiPathways

8.1.1 WikiPathways as an integration resource

In order to perform large-scale analysis of transcriptomic data, there is a need for a pathway database for the biological interpretation of data, to explain how genes are involved in biological processes. WikiPathways is one such database. It was a very relevant resource in this thesis. Especially the open science nature and flexibility of WikiPathways allowed us to experiment with the integration of molecular pathways with AOPs and develop molecular AOPs as meta-pathways [17]. The ability to update biological pathways with the latest research insights and add pathways that were not yet sufficiently described allows the use in processes yet not well understood [18], as is often the case in toxicology. It is important to not only perform analyses and interpretation of data but also build on the common knowledge of molecular processes of toxicity. Together with its accompanying pathway editor tool PathVisio [19] and Cytoscape [20] which has a WikiPathways plugin [21], the omics data analysis and visualisation capabilities of related tools are other essential aspects that make WikiPathways such a useful resource.

We previously described in Chapter 2 the power of WikiPathways to engage research communities to work together. We observed that WikiPathways has had a steady growth in terms of content and contributors, focused on community-driven pathway development and curation [22, 23], as described in Chapter 2. Also, the resource allows for many ways to interact with the content which are also machine-readable, through user interfaces, through coding environments and APIs, and through third-party tools that have integrated WikiPathways or can easily load the content, as well as through a SPARQL Protocol and Resource Description Framework (RDF) Query Language (SPARQL) endpoint loaded with the data [17, 24]. That makes the WikiPathways database and its corresponding tools perfect to fulfil the goal of linking the AOPs in AOP-Wiki with molecular pathways. Furthermore, the resource has all the requirements to allow communities to collaborate

and develop new pathway diagrams [25]. We think it has the potential to become a main hub for molecular AOPs and bring experts from the field together. Also, the flexible nature of identifier handling of WikiPathways through BridgeDb [26] allows the inclusion of KE nodes in pathway models, enabling interoperability with external resources such as the AOP-Wiki [27]. Besides the possibility of analysing transcriptomics data being the main focus of this thesis, WikiPathways can be used to analyse all types of omics data, including proteomics [28–30], metabolomics [31, 32], epigenomics [33], genomics [34, 35], and combinations thereof [36], highlighting the data analysis potential of WikiPathways with regards to molecular AOPs.

8.1.2 Linking WikiPathways to the AOP-Wiki

Upon establishing that WikiPathways is a relevant resource to link molecular pathways with AOPs, we wanted to explore how the AOP-Wiki, the central repository for AOPs, could be integrated with WikiPathways. We chose here to link KEs to pathways instead of genes, as one would do when using gene activity instead of pathway-level changes as a biomarker for KEs, adding to our biological understanding of how these biomarkers are involved in biological systems. Chapter 3 shows that the majority of early KEs (molecular, cellular and tissue-level) can be linked to relevant molecular pathways in WikiPathways and that 30% of chemical stressors of the AOP-Wiki were found in existing pathways [27], showing that molecular pathways can potentially cover the majority of the known AOPs. Since WikiPathways contains mostly endogenous metabolites and chemicals within cellular pathways, it does not contain the majority of stressors in the AOP-Wiki which are mostly exogenous. However, since AOPs are meant to be chemical agnostic, meaning that the focus is on processes activated by exposure to any stressor, there is no necessity to include all potential toxicants within molecular pathways. On the other hand, over 70% of all mapped genes on textual descriptions in AOP-Wiki were found in

WikiPathways. This shows that the majority of described genes are involved in biological processes already captured in the database, which include stress response pathways or signaling pathways. The remaining 30% could be genes involved in regulatory processes such as transcription factors and micro RNA, or structural genes not part of active processes and are therefore not represented in molecular pathways.

As discussed in the last paragraph, the process of gene-based matching of KEs to molecular pathways has its limitations which cause gaps and uncertainties in the linking of KEs with molecular pathways. Therefore, in Chapter 3, we also performed a manual analysis of all early KEs in AOP-Wiki, showing that approximately 67% of KEs can potentially be linked to molecular pathways in WikiPathways. As already indicated, there are challenges in making the connection, where KEs and pathways do not always match one-to-one, the biology of KEs is not yet fully understood, or could simply not be represented as molecular pathways because the KEs do not describe pathway processes or the pathway is not represented in WikiPathways. Taken together with the usability of WikiPathways and the identified connectivity between it and the AOP-Wiki, the integration of the resources was expected to facilitate the use of omics approaches in risk assessments by directly connecting molecular pathways and KEs [27]. For example, although the two resources apply ontologies to annotate pathways and biological processes in KEs, these resources do not align and could therefore not be utilized to make connections. Furthermore, the depiction of molecular processes in WikiPathways that underlie KEs are not always one-to-one mirrors of those KEs because the biological complexity, which includes feedback loops and are typically not included in AOPs [9]. This challenge, in particular, was discussed in more depth in Chapter 7 with the introduction of molecular AOPs as a manually curated model to match KEs with molecular pathways. On the one hand, this has shown us that multiple KEs can be part of the same molecular pathway, especially at the molecular and cellular level of organization. On the other hand, single KEs can potentially involve more than

one molecular pathway, which is the case mostly for late KEs, such as the KE of cell death which is not always specifically described and could involve multiple cell death pathways. The development of the molecular AOP models and the curation or new creation of molecular pathways linked to KEs were performed in PathVisio and uploaded to WikiPathways.

With the challenges of automated matching of KEs and molecular pathways discussed before based on genes and ontology terms, the proposed method that emerged was the molecular AOP model, which is a manual process, introduced in Chapter 7. The idea was to develop a method that is simple and flexible, utilizes as much of the existing pathways as possible without duplication, and allows for data analysis and visualisation. These aspects would be essential to expand the pool of potential users and applications and have an approach that is in line with the existing AOP framework. The resulting framework of the molecular AOP that we implemented is a so-called meta-pathway that only contains links to molecular pathways and KEs, and the connections between those, resembling the original AOPs. This model follows the current description of AOPs, consisting of separated KEs which as modules can exist in multiple AOPs, which is also the case for the molecular AOP model.

To illustrate the capabilities of the molecular AOP as a meta pathway, case studies were performed on an AOP network of liver steatosis [37] and an AOP of neurodegeneration [38] using public, *in vitro* transcriptomics datasets. The molecular AOPs for these case studies were developed using contents of the AOP-Wiki, and scientific literature that supports the AOPs to ensure that we selected the correct molecular pathways for each KE. Based on the molecular AOPs enriched with gene expression data, we calculated KE enrichment scores, assessing the potency of exposure scenarios on affecting KE processes. While showing dose-response patterns in liver steatosis AOPs, the patterns in KE activation with increasing dose, time, or stressor type were not consistent across the case studies, possibly due to the differences in the case studies, misalignment of KEs and underlying processes,

or regulatory and compensatory processes that are involved in the molecular pathways. Additionally, the datasets had only limited exposure scenarios, making it difficult to identify patterns. The complexity of biology represented not only in molecular pathways but also in omics datasets does not align with the simplification of biology in AOPs. Therefore, molecular AOPs, as a literal bridge between traditional AOPs and molecular pathways, should be regarded as a link from simplified biology to biological complexity. To create and curate molecular AOPs, serious efforts are required to ensure that all relevant biology can be captured and linked to the AOP, including the compensatory, modulatory or feedback processes that are purposely kept out of AOPs in the AOP-Wiki. These aspects pose challenges to the overall acceptance of this type of analysis for application in risk assessments, and additional studies with more exposure scenarios would be required to evaluate the method for its utility.

Taken together, we argue that, based on these results and taking into account the current limitations, the integration of AOPs and molecular pathways can become a powerful tool to support risk assessments. We can explain the biological plausibility of KE activation by exploring and interpreting molecular pathways, and we can use transcriptomics data to assess KE activation based on the effect on underlying molecular pathways, although this depends on the particular case study and dataset as described before. The molecular AOPs can serve as a flexible scaffold to combine all KEs with molecular pathways and perform transcriptomics data analyses, which can be applied for the majority of AOPs [27]. With increasing insights into the molecular understanding of KEs, focused curation and pathway development approaches can be initiated to increase the overall coverage of KEs with corresponding molecular pathways. For example, curation efforts can be focused on the remaining 33% of KEs that were not yet able to link to a molecular pathway, or exploring the 30% of genes in KEs that were not yet present in WikiPathways.

8.1.3 Expanding the utility of molecular AOPs

With WikiPathways being community-driven and freely accessible, it provides ample opportunity for AOP developers to get involved and develop molecular AOPs. With the setup of the AOP community portal in WikiPathways, a community of pathway developers with shared expertise is still needed. Since there is no centralized team of pathway developers and topic experts for curation, contrary to other molecular pathway databases such as Reactome [39–41] and Kyoto Encyclopedia of Genes and Genomes (KEGG) [42], pathways in WikiPathways can be inconsistent in the level of detail in their pathways [43, 44]. It has also been known that the same biological pathway can be represented differently in alternative databases, which influences statistical enrichment analyses and the overall interpretation of gene expression data [45]. It would therefore be more optimal if we could utilize multiple pathway databases and develop integrative pathway models, including all relevant biological knowledge that underlie the KEs [43, 45, 46]. However, where WikiPathways is flexible and allows communities to co-create, other resources do generally not use a crowdsourcing approach like WikiPathways. For the AOP community to get involved with WikiPathways and molecular AOPs, and use them for analysing their transcriptomics data, there is a need for training, and a need to prove the value and application of molecular AOPs to encourage the AOP community. Such training materials exist for the basics of WikiPathways and PathVisio, such as the WikiPathways Academy [47], but these are not specifically tailored towards the development of molecular AOPs and using these for analyses.

The molecular AOP model can combine transcriptomics data and AOPs, it can be used to perform data analyses to find activated KEs, to explore the molecular processes impacted by exposure to stressors, and to generate hypotheses based on these analyses on the molecular pathways that underlie KEs. Although the molecular AOPs are meant to capture the complexity of cellular biology in pathways and allow for detailed interpretation of the data, other methods to use transcriptomics data in the risk assessment domain

have been studied. For example, large public transcriptomics datasets could be used for weighted gene co-expression network analysis to develop visualisations with functional modules based on clustered gene sets [48, 49], being a data-driven approach rather than the literature-based molecular AOPs.

These data-driven frameworks could then be used to analyse transcriptomic datasets to explore biological perturbations caused by toxicants, using the functional annotations of the modules. While this type of analysis can be of great value to rapidly explore the potential mechanisms of a toxicant, the applicability is limited to a handful of well-studied endpoints as it requires large datasets to develop. Furthermore, whereas both the modules in such analysis and the KE enrichment in molecular AOPs generally comprise gene lists, the molecular AOPs have the additional benefit of specifically linking to molecular pathways for detailed interpretation and understanding of the biological processes. This also supports our understanding of the biological systems and how these are connected, which relates to the KERs of AOPs where the biological plausibility of the link between KEs is described based on our understanding of biology. Another difference between the methods is the basis on which the data models were built. Whereas the TXG-MAPr models were data-driven, the molecular pathways and AOPs are generally based on literature and are developed by researchers and topic experts. This makes the development and improvement of the molecular AOPs more flexible and undergo continuous improvements based on new scientific insights, without the dependence on large amounts of data to build new data models [17].

The human-driven strategy was also used to annotate the AOPs in AOP-Wiki to generate AOP-derived *in vitro* transcriptional biomarkers for KEs, which could be used to derive AOP fingerprints and serve as robust biomarkers for pulmonary fibrosis when studying Multi-Walled Carbon Nanotube toxicity [50]. However, the process of validation of these transcriptional biomarkers is case-dependent and requires extensive experimental confirmation. Also, similar to the TXG-MAPr tool, these biomarker sets can be used to assess the activation of partic-

ular processes, but are limited in the overall biological interpretation on the pathway level and molecular interactions. Furthermore, the data-driven and case-specific approaches are generally aimed at providing insights into specific endpoints of toxicants and require additional experimentation to expand to other endpoints or AOPs. Also, in general, biomarker sets such as those used in the described approaches encounter problems involving the number of false positives in the original dataset, and most robust biomarkers correspond to late, general effects. The combination of such approaches with the AOPs could help filter false positives that do not occur in pathways and aggregate responses on the pathway level. Since molecular AOPs are based on literature and are linked to an extensive library of existing molecular pathways, it is simple to expand molecular AOPs to cover other endpoints, AOPs, or AOP networks.

The concept of linking gene sets to particular processes as transcriptional biomarkers was shown to have high predictive value on whether a particular pathway is activated. For example, genotoxic compounds and skin sensitizers can accurately be identified with curated gene sets [51, 52]. However, an additional goal of our method was to have an additional level of biological understanding and data interpretation which is possible through molecular pathway models. To move forward with the development of molecular AOPs, known transcriptional biomarkers could be used to verify or strengthen the matching of KEs and molecular pathways in WikiPathways, potentially expanding our mechanistic understanding of the toxicity. Whereas Chapter 7 introduces the method and shows the application of the molecular AOPs using two example cases, the actual validation of this approach for application and acceptance in risk assessments was not yet investigated. We expect that the integrated approach of molecular AOPs would make transcriptomic data analysis more robust, consistent, and acceptable for omics-derived regulatory risk assessments. By the time molecular AOPs cover most of the AOPs in AOP-Wiki, probably over the upcoming few years, the network-based analyses would allow rapid exploration of process activation based on single exten-

sive transcriptomic datasets instead of batteries of individual assays that measure single KEs. Although the outcomes of molecular AOPs might not be as robust as these specific assays, the transcriptomic data can provide many insights, increase our understanding of the KEs, and generate hypotheses by showing which molecular processes are affected based on changes in gene expressions.

All molecular AOPs developed so far are stored in the AOP portal of WikiPathways, where the toxicology community can contribute and discuss the pathway models. Thus far, the development of molecular AOPs was driven by project case studies, which provide a practical approach for testing new methodologies, comparing with other knowledge or previous findings, and refining strategies. However, these case studies were performed with limited curation by experts. This is why the current approach is limited, and there is a need for more experts to evaluate the results in a crowdsourced approach. While our approaches allow such a community approach for knowledge, with WikiPathways for example, a general equivalent for data analysis results does not yet exist. For further application and implementation of molecular AOPs, expert curation of the underlying pathways can help improve confidence in the use of molecular AOPs. This can be achieved through tailored curation workshops with domain experts, as was shown with the community-driven molecular pathway on mesothelioma, which is known to be an adverse outcome of inhalation of asbestos particulates [53].

8.2 Toxicological data

8.2.1 Omics approaches in the life sciences

Whereas the previous section focused on the linking of molecular pathways and AOPs, the second goal of this thesis was to make transcriptomics data more accepted in risk assessment approaches. It is generally understood that omics technologies provide a tremendous amount of data to describe the complexity of molecular biology, and have many applications in the life sciences. This is also the case in toxicology,

where various types of biomarkers are commonly used to assess the activation of processes or adverse effects. For example, gene expression levels based on transcriptomic experiments can show the transcriptional activation of molecular pathways, and metabolites from metabolomics approaches can in some cases highlight functional disturbances of processes. The project that is presented in this thesis focuses on transcriptomic data, which is widely applied in toxicology projects to uncover the mode of action of toxicants. Although transcriptomics approaches can highlight altered gene expression levels [54], the data type does have drawbacks. For example, transcriptomic data only show changes in the expression of genes, after which a myriad of processing and regulatory steps needs to happen to produce proteins and achieve functional changes within the cell. It is therefore complicated to correlate the changes in transcription to the potential for adverse effects.

However, to use transcriptomics data in risk assessment, a thorough mechanistic understanding of the pathways is needed. That is where the simplified, linear nature of AOPs is a limiting factor, which generally does not describe molecular pathways but rather focuses on a larger scale of biology. AOPs have been described as the bridge between toxicological scientists and the risk assessment community, capturing current understanding of toxicological processes, tailored for use and informing risk assessors. However, the current concept limits the potential integration with experimental data and in particular omics data that provides molecular insights and addresses complexity. This is why the suggested solutions for implementing omics through AOPs in risk assessments include molecular annotations through pathways or gene lists serving as biomarkers [12, 55]. This is in line with our approach to manually develop and curate molecular AOPs to connect molecular pathways to KEs and thus serve as the templates required for transcriptomics analyses. The potential for such pathway-level generalization of particular KEs has been shown for various stress response pathways which respond similarly, independent from the MIEs or earlier KEs that

might cause their activation [56, 57]. Such observations are promising for the expansion of molecular AOPs and networks as described in Chapter 7. Whereas our connection of KEs to molecular pathways comprises all genes and proteins involved in particular processes, a particular focus on well-established transcriptional response gene sets might serve for more direct read-outs on KE activation.

8.2.2 Applying Findable, Accessible, Interoperable, and Reusable (FAIR) principles to improve the integration of resources

As part of studying the linking of transcriptomics data and AOPs, we explored how databases involved in that effort could be made more accessible and reusable to facilitate the integration of other types of data and knowledge. This would also be of more general interest since the field of toxicology produces vast amounts of data of many different types, which can be stored in various data repositories. For example, public repositories such as Gene Expression Omnibus (GEO) [58, 59] and the European Nucleotide Archive which contains ArrayExpress [60] are used to store transcriptomics datasets, and some specific to toxicological data, such as ToxBank and Comparative Toxicogenomics Database (CTD).

Furthermore, the transition in risk assessment to become more data-driven and based on existing mechanistic knowledge of toxicological processes relies on databases and scientific knowledge. This strengthens the need for efficient use and reuse of data. Since the amount of experimental data in publicly accessible repositories is expanding rapidly, there is a need for agreement and guidance in data handling and standardization of life science data to improve their usability and integrative capabilities. This is where the FAIR principles play a role in improving the reuse of data by making the data more findable, accessible, interoperable and reusable [61].

In the life sciences, the FAIR principles are increasingly applied to promote data reuse and have been tailored for research software as

well [62]. For example, the application of FAIR principles is central to the goals of the ELIXIR Toxicology Community [63], and the field of nanomaterial toxicology also aims to apply the FAIR principles on nanosafety data [64]. Also, a variety of tools have become available to assess the FAIRness of data and resources, for example by using FAIR maturity indicators on data repositories [65].

Increasing the FAIRness of a data resource can be achieved by applying semantic web technologies, for example through the transformation of data to RDF [66]. This makes the data accessible in more ways and since Linked Open Data (LOD) principles [67] are applied by using ontologies and unique, persistent and resolvable identifiers, the interoperable capabilities of the resource increase as well. This approach was used, for example, to create linked open data for ChEMBL [68], WikiPathways [24], various databases of the European Bioinformatics Institute (EBI) [69], and DisGeNET [70], and Wikidata [71], among others.

8.2.3 Making AOP knowledge FAIR

Within the realm of AOPs, the main database to develop and distribute AOPs is the AOP-Wiki, comprising hundreds of qualitative descriptions of AOPs including over a thousand KEs. The resource, existing mostly of free-text descriptions and ontological annotations, has few options to interact with the data, leaving the users to explore the extensive database by manually searching or downloading the data in Extensible Markup Language (XML). Therefore, we envisioned a semantic version of the AOP-Wiki to improve the overall usability of the data and make the resource more FAIR.

Therefore, Chapter 4 explores methods of integration of the AOP-Wiki with other resources using Linked Open Data standards. The RDF model was selected to create an extended, semantically annotated version of the contents in the AOP-Wiki. With the development of the RDF, various ontologies and resource identifiers were added to form the knowledge graph, making the data more interoperable and acces-

sible [72]. When loaded into a public SPARQL endpoint, the data has been made findable and accessible using SPARQL queries, all of which are in line with the FAIR principles for improved data usability. Besides the possibility to explore the data from computational environments and workflow systems, the SPARQL endpoint allows data flow to and from remote resources with federated queries. Being part of the LOD world, there is potential integration with a vast amount of external resources.

Additional to the AOP-Wiki, the AOP-DB, an effort by the United States Environmental Protection Agency (US EPA), contained additional data relevant to AOPs that is not part of the AOP-Wiki [73]. The resource expands and integrates AOP knowledge by combining various resources linked to AOPs and the AOP-Wiki, including genes, chemicals, ToxCast assays, Single Nucleotide Polymorphisms (SNPs), pathways, diseases, and more [74]. Chapter 5 describes the creation of a semantic version of the AOP-DB, where the contents of seven of its core data tables were converted into RDF [75]. This was done in line with the AOP-Wiki RDF for optimal interoperability between the two resources and has led to improved access to AOP data and associated data of toxicological interest. Whereas the majority of the core framework of the RDF could be directly linked to the AOP-Wiki RDF, some of the predicates and object types required additional attention to select the most fitting ontological annotations.

Overall, the development of semantic versions of the AOP-Wiki and AOP-DB makes it possible to explore the activated biological processes in AOPs in more ways and integrate the resources with other sources to compare or validate found results with independent knowledge bases. With the implementation of RDF and allowing exploration with SPARQL queries, the types of questions that one can pose to investigate the AOP contents exceed far beyond what is possible with the original user interface. For example, Chapter 4 describes connections between the AOP-Wiki RDF and chemical databases or molecular pathway databases. However, there are many more potential connections with the resource, such as Wikidata [76].

Wikidata aims to assemble a knowledge graph containing all information of relevance in the life sciences, from chemical data, to genomic data, pathways, and disease data. With the addition of semantic data availability of the AOP-Wiki and AOP-DB, there are ample opportunities to expand AOP knowledge and integrate with resources such as Wikidata. Making the contents of the AOP-Wiki more FAIR promotes the reuse of AOP knowledge. If more toxicological resources would apply FAIR principles and expose their data for exploration as RDF, the potential to integrate such resources would change how we would use these resources and answer questions relevant to risk assessments as will be explained later.

8.2.4 Showing the utility of FAIR data resources in a workflow

We wanted to know how easy it was to use these integrated services. Therefore, to illustrate the possibilities and strengths of making resources FAIR, we created a Jupyter notebook that utilizes both AOP resources. By itself, the AOP-Wiki has limited content on quantitative data nor does it refer to data to a large extent. Therefore, our goal was to find and analyse experimental data that supports an AOP of interest, by integrating the resource with other tools and databases within a workflow. Chapter 6 illustrates how combining services and data can perform this task relevant to risk assessment of identifying experimental data to support an AOP. The developed workflow supports the re-use of data by finding experimental data that fits the context, and could therefore limit the number of experiments that would be needed to measure all KEs that occur after exposure to a stressor. This is also in support of minimizing the need for experimental studies on animals, by exploring experimental data in Toxicogenomics Project-Genomics Assisted Toxicity Evaluation System (TG-GATES) [77], which contains data not only on human cell lines but also on rats. By making this Jupyter Notebook automated and only requiring the AOP ID as an input, the notebook is reusable for any AOP, and the outcomes are reproducible. This is in line with the larger movement in science to increase the reproducibility of workflow results and repeatability of workflow

execution [78, 79]. While the amount of new data in biomedical research grows at an increasing rate, it has become clear that much of experimental results are not meaningful and hard to reproduce. This is referred to as the reproducibility crisis and has become a major issue that needs to be addressed in current research practice [78]. Generally, Jupyter notebooks are great tools for developing understandable, reproducible workflows, where code, narrative text, functions and visualisations are combined in a single document, while also taking care of all dependencies to run the workflow. However, since the current workflow sends requests to live, external services and databases, it is dependent on these services being available and consistent without changing API or data to have fully reproducible results.

Ideally, all resources that are used for such workflow comply with the FAIR principles for data and software, in order to ensure the longevity of the functional workflow [80, 81]. To optimize the reproducibility of computational workflows that rely on online resources and services, they should be more stable and sustainability should be ensured. Large projects such as the European Open Science Cloud (EOSC) have a clear vision to ensure the implementation of FAIR principles and on the need for sustainable research tools [82].

8.3 Adverse Outcome Pathways as a tool in risk assessments

The central theme in this thesis is the AOP concept. Since its introduction following an era of the Mode of Action (MoA) as the focus of risk assessment studies, the generalised AOPs have taken over toxicology research. Generally supported as a driving force for the paradigm shift from traditional animal testing in risk assessments toward *in vitro* assays and *in silico* predictions, AOPs have become a central theme in recent toxicology projects (see Table 8.1). Furthermore, the Organisation for Economic Co-operation and Development (OECD) launched the AOP Development Programme in 2016 to support AOP development, thus far leading to nineteen reviewed and endorsed AOPs in the AOP-

Wiki [83]. These AOPs are for various adverse effects such as liver fibrosis, learning and memory impairment, and various adversities for fish. Also, multiple human-relevant AOPs were approved and have been implemented successfully in Integrated Approaches to Testing and Assessment (IATA) case studies for skin sensitization [84–86], liver steatosis [87], Parkinsonian motor deficits [88], and developmental neurotoxicity [89, 90].

However, the AOP framework is not without flaws or drawbacks. While informative and efficient in displaying current knowledge of the sequential biological disturbances after stressor exposure, the actual construction of AOPs is labour-intensive. Besides that, there is little incentive to develop AOPs to the extent of fully usable AOPs fit for application and push these into the public AOP-Wiki, also because current research is focused on publications in scientific journals which do not generally follow a format of AOPs. These are some of the reasons for the relatively limited number of reviewed and endorsed AOPs in the AOP Knowledge Base (AOP-KB). Furthermore, AOPs as single, linear chains of KEs, do not always serve their purpose by themselves for applications in risk assessments. AOP networks can be more realistic in that respect as they describe all pathways that can lead to a particular apical endpoint. This is one of the reasons for the development of the AOP-Wiki RDF. Not only can we utilize the created knowledge graph to expand contents by linking external resources such as molecular pathway databases, but we can also take the modular nature of AOPs to generate new paths of KEs to form AOPs. The possibilities of data and knowledge integration to both expand AOP knowledge and hypothesize new AOPs have been described earlier, utilizing the existing Key Event Components (KECs) or toxicological data to generate computationally predicted AOPs (cpAOPs) [91].

A solution is needed to make AOPs easier to explore and use existing AOP knowledge. This is where semantic web approaches can facilitate improving the accessibility and interoperability of AOP resources. Chapters 4 and 5 contribute to the overall interoperability of

AOP knowledge through the introduction of global, persistent identifiers that allow the linking to other databases. For example, adding chemical identifiers for all stressors can provide all relevant information about them, from their structure to their role in biology, or disturbance thereof, by linking the stressors to ChEBI [92], ChEMBL [68, 93], Wikidata [76], and ToxBank [94], among other resources. This also counts for the proteins that are mapped from the biological object annotations of KEs, providing additional links to protein and gene databases such as UniProt [95] and Ensembl [96], among others.

Table 8.1: Recent projects in which AOPs are a central theme.

Project name	EU H2020 or NWO NWA Grant number	Website
EU-ToxRisk	681002	https://www.eu-toxrisk.eu/
OpenRiskNet	731075	https://openrisknet.org/
NanoSolveIT	814572	https://nanosolveit.eu/
RiskGONE	814425	https://riskgone.eu/
ONTOX	963845	https://ontox-project.eu/
VHP4Safety	NWA 1292.19.272	https://vhp4safety.nl/
CIAO		https://www.ciao-covid.net/
PATROLS	760813	https://www.patrols-h2020.eu/
EDCMET	825762	https://sites.uef.fi/edcmet/
OpenTox	200787	https://opentox.net/
ERGO	825753	https://ergo-project.eu/
SmartNanoTox	686098	http://www.smartnanotox.eu/
EuroMix	633172	https://www.euromixproject.eu/
HBM4EU	733032	https://www.hbm4eu.eu/

8.4 Conclusion

In conclusion, this thesis project was aimed at making AOPs and transcriptomic data more usable for risk assessment by more accurately

describing the knowledge and data, specifically focusing on the analyses and interpretation of the data. The main reason for this is the current refrain from implementing transcriptomic data in risk assessments, despite the proven value of the widely-applied technology. Because AOPs are a more commonly used knowledge framework to support risk assessments, our goal was to create a connection between transcriptomic data and AOPs. Therefore, the first chapters of this thesis explored how the molecular pathway database of WikiPathways could be integrated with the AOP-Wiki, which was thus far a stand-alone resource with limited integrative ability. Various approaches were evaluated and one was chosen that links KEs to full biological processes represented in molecular pathways, making activity of biological processes rather than individual gene expression biomarkers of KEs. It was clear that the AOP-Wiki should focus on making the data more interoperable and FAIR, allowing more ways to explore and analyse its contents. Therefore, the following chapters focused on the creation of LOD of the AOP-Wiki and AOP-DB, making their data accessible through a SPARQL Application Programming Interface (API) that allows scripted querying from coding environments. Their application was shown in an automated, flexible workflow in a Jupyter notebook that automatically finds and analyses experimental data to support an AOP of interest, which allows us to perform many more actions and analyses when compared to manual approaches to data exploration and analysis. Finally, this thesis has shown that the combined approaches enable the application of molecular AOPs with WikiPathways to allow the visualisation and reproducible analyses of transcriptomic data for identifying KE activation. This provides a new approach to visualize the data, validate the biological plausibility, and generate new hypotheses for high-throughput studies and KE activation. This approach can potentially bridge the gap between the commonly used big data approaches and the risk assessment community, where such methods have not yet played a large role. In order to prove the usefulness of this work, future research should focus on testing the data analysis and interpretation approaches and comparing them to existing risk assessment strategies. As this thesis provided only a

handful of example applications and limitations of molecular AOPs, there is a need for validation and comparison to more traditional assays to assess AOP activation. By performing additional case studies with molecular AOPs, as is currently ongoing in VHP4Safety, we will be able to show the utility of transcriptomics in risk assessment approaches of chemicals and nanomaterials.

References

- [1] Zhanyun Wang et al. "Toward a Global Understanding of Chemical Pollution: A First Comprehensive Analysis of National and Regional Chemical Inventories". *Environmental Science and Technology* 54.5 (Mar. 2020), pp. 2575–2584. DOI: 10.1021/ACS.EST.9B06379.
- [2] William Moy Stratton Russell and Rex Leonard Burch. *The principles of humane experimental technique*. Methuen, 1959.
- [3] Gary L. Ginsberg et al. "New toxicology tools and the emerging paradigm shift in environmental health decision-making". *Environmental Health Perspectives* 127.12 (Dec. 2019), p. 125002. DOI: 10.1289/EHP4745.
- [4] Robert J. Kavlock et al. "Accelerating the Pace of Chemical Risk Assessment". *Chem. Res. Toxicol.* 31.5 (May 2018), pp. 287–290. DOI: 10.1021/acs.chemrestox.7b00339.
- [5] Stanley T. Parish et al. "An evaluation framework for new approach methodologies (NAMs) for human health safety assessment". *Regulatory Toxicology and Pharmacology* 112 (Apr. 2020), p. 104592. DOI: 10.1016/j.yrtph.2020.104592.
- [6] Gerald T. Ankley et al. "Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment". *Environmental Toxicology and Chemistry* 29.3 (Mar. 2010), pp. 730–741. DOI: 10.1002/etc.34.
- [7] Mathieu Vinken. "The adverse outcome pathway concept: A pragmatic tool in toxicology". *Toxicology* 312.1 (Oct. 2013), pp. 158–165. DOI: 10.1016/j.tox.2013.08.011.
- [8] Gerald T. Ankley and Stephen W. Edwards. "The adverse outcome pathway: A multifaceted framework supporting 21st century toxicology". *Current Opinion in Toxicology* 9 (June 2018), pp. 1–7. DOI: 10.1016/j.cotox.2018.03.004.
- [9] Marcel Leist et al. "Adverse outcome pathways: opportunities, limitations and open questions". *Archives of Toxicology* 91.11 (Nov. 2017), pp. 3477–3505. DOI: 10.1007/s00204-017-2045-3.
- [10] Roland Buesen et al. "Applying 'omics technologies in chemicals risk assessment: Report of an ECETOC workshop". *Regulatory Toxicology and Pharmacology*. Vol. 91. Academic Press, Dec. 2017, S3–S13. DOI: 10.1016/j.yrtph.2017.09.002.
- [11] Ursula G. Sauer et al. "The challenge of the application of 'omics technologies in chemicals risk assessment: Background and outlook". *Regulatory Toxicology and Pharmacology* 91 (Dec. 2017), S14–S26. DOI: 10.1016/j.yrtph.2017.09.020.

- [12] Erica K. Brockmeier et al. "The Role of Omics in the Application of Adverse Outcome Pathways for Chemical Risk Assessment". *Toxicological Sciences* 158.2 (Aug. 2017), pp. 252–262. DOI: 10.1093/toxsci/kfx097.
- [13] Jessica J.Y. Y Lee et al. "Knowledge base and mini-expert platform for the diagnosis of inborn errors of metabolism". *Genetics in Medicine* 20.1 (Jan. 2018), pp. 151–158. DOI: 10.1038/gim.2017.108.
- [14] J. Christopher Corton. "Integrating gene expression biomarker predictions into networks of adverse outcome pathways". *Current Opinion in Toxicology* 18 (Dec. 2019), pp. 54–61. DOI: 10.1016/j.cotox.2019.05.006.
- [15] Kirsten A. Baken et al. "A strategy to validate a selection of human effect biomarkers using adverse outcome pathways: Proof of concept for phthalates and reproductive effects". *Environmental Research* 175 (Aug. 2019), pp. 235–256. DOI: 10.1016/j.envres.2019.05.013.
- [16] Laura Alisa Saarimäki et al. "Toxicogenomics Data for Chemical Safety Assessment and Development of New Approach Methodologies: An Adverse Outcome Pathway-Based Approach". *Advanced Science* 10.2 (Jan. 2023), p. 2203984. DOI: 10.1002/ADVS.202203984.
- [17] Marvin Martens et al. "WikiPathways: connecting communities". *Nucleic Acids Research* 49.D1 (Jan. 2021), pp. D613–D621. DOI: 10.1093/nar/gkaa1024.
- [18] Marek Ostaszewski et al. "COVID19 Disease Map, a computational knowledge repository of virus–host interaction mechanisms". *Molecular Systems Biology* 17.10 (Oct. 2021), e10387. DOI: 10.15252/msb.202110387.
- [19] Martina Kutmon et al. "PathVisio 3: An Extendable Pathway Analysis Toolbox". *PLOS Computational Biology* 11.2 (Feb. 2015). Ed. by Robert F. Murphy, e1004085. DOI: 10.1371/journal.pcbi.1004085.
- [20] Paul Shannon et al. "Cytoscape: A software Environment for integrated models of biomolecular interaction networks". *Genome Research* 13.11 (2003), pp. 2498–2504. DOI: 10.1101/gr.1239303.
- [21] Martina Kutmon et al. "WikiPathways App for Cytoscape: Making biological pathways amenable to network analysis and visualization". *F1000Research* 3 (Sept. 2014), p. 152. DOI: 10.12688/f1000research.4254.2.
- [22] Alexander R. Pico et al. "WikiPathways: Pathway Editing for the People". *PLoS Biology* 6.7 (July 2008), e184. DOI: 10.1371/journal.pbio.0060184.
- [23] Thomas Kelder et al. "WikiPathways: building research communities on biological pathways". *Nucleic Acids Research* 40.D1 (Jan. 2012), pp. D1301–D1307. DOI: 10.1093/nar/gkr1074.
- [24] Andra Waagmeester et al. "Using the Semantic Web for Rapid Integration of WikiPathways with Other Biological Online Data Resources". *PLOS Computational Biology* 12.6 (June 2016). Ed. by Christos A. Ouzounis, e1004989. DOI: 10.1371/journal.pcbi.1004989.
- [25] Kristina Hanspers et al. "Ten simple rules for creating reusable pathway models for computational analysis and visualization". *PLOS Computational Biology* 17.8 (Aug. 2021). Ed. by Scott Markel, e1009226. DOI: 10.1371/journal.pcbi.1009226.
- [26] Martijn P van Iersel et al. "The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services". *BMC Bioinformatics* 11.1 (2010), p. 5. DOI: 10.1186/1471-2105-11-5.

-
- [27] Marvin Martens et al. "Introducing WikiPathways as a Data-Source to Support Adverse Outcome Pathways for Regulatory Risk Assessment of Chemicals and Nanomaterials". *Frontiers in Genetics* 9 (Dec. 2018), p. 661. DOI: 10.3389/fgene.2018.00661.
- [28] Freek G. Bouwman et al. "2D-electrophoresis and multiplex immunoassay proteomic analysis of different body fluids and cellular components reveal known and novel markers for extended fasting". *BMC Med. Genomics* 4.1 (Mar. 2011), pp. 1–12. DOI: 10.1186/1755-8794-4-24.
- [29] Aoife M. Curran et al. "Sexual Dimorphism, Age, and Fat Mass Are Key Phenotypic Drivers of Proteomic Signatures". *J. Proteome Res.* 16.11 (Nov. 2017), pp. 4122–4133. DOI: 10.1021/acs.jproteome.7b00501.
- [30] Qi Qiao et al. "Adipocyte abundances of CES1, CRYAB, ENO1 and GANAB are modified in-vitro by glucose restriction and are associated with cellular remodelling during weight regain". *Adipocyte* 8.1 (Jan. 2019), pp. 190–200. DOI: 10.1080/21623945.2019.1608757.
- [31] Monica Chagoyen and Florencio Pazos. "Tools for the functional interpretation of metabolomic experiments". *Brief. Bioinform.* 14.6 (Nov. 2013), pp. 737–744. DOI: 10.1093/bib/bbs055.
- [32] Stefan Jenkins et al. *Global LC/MS Metabolomics Profiling of Calcium Stressed and Immunosuppressant Drug Treated Saccharomyces cerevisiae*. 2013. DOI: 10.3390/metabo3041102.
- [33] Martina Kutmon et al. "Integrative network-based analysis of mRNA and microRNA expression in 1,25-dihydroxyvitamin D3-treated cancer cells". *Genes Nutr.* 10.5 (2015), p. 35. DOI: 10.1007/s12263-015-0484-0.
- [34] Sarah Mount et al. "Network Analysis of Genome-Wide Association Studies for Chronic Obstructive Pulmonary Disease in the Context of Biological Pathways". *Am. J. Respir. Crit. Care Med.* 200.11 (July 2019), pp. 1439–1441. DOI: 10.1164/rccm.201904-0902LE.
- [35] Elisa Cirillo et al. "From SNPs to pathways: Biological interpretation of type 2 diabetes (T2DM) genome wide association study (GWAS) results". *PLoS One* 13.4 (Apr. 2018). Ed. by Qingyang Huang, e0193515. DOI: 10.1371/journal.pone.0193515.
- [36] Isabel Rubio-Aliaga et al. "Alterations in hepatic one-carbon metabolism and related pathways following a high-fat dietary intervention". *Physiol. Genomics* 43.8 (Feb. 2011), pp. 408–416. DOI: 10.1152/physiolgenomics.00179.2010.
- [37] Mathieu Vinken. "Adverse Outcome Pathways and Drug-Induced Liver Injury Testing". *Chemical Research in Toxicology* 28.7 (July 2015), pp. 1391–1397. DOI: 10.1021/acs.chemrestox.5b00208.
- [38] Andrea Terron et al. "An adverse outcome pathway for parkinsonian motor deficits associated with mitochondrial complex I inhibition". *Archives of Toxicology* 92.1 (Jan. 2018), pp. 41–82. DOI: 10.1007/s00204-017-2133-4.
- [39] Antonio Fabregat et al. "The Reactome Pathway Knowledgebase". *Nucleic Acids Research* 46.D1 (Jan. 2018), pp. D649–D655. DOI: 10.1093/nar/gkx1132.
- [40] Anwesha Bohler et al. "Reactome from a WikiPathways Perspective". *PLoS Computational Biology* 12.5 (May 2016). DOI: 10.1371/journal.pcbi.1004941.

- [41] Marc Gillespie et al. "The reactome pathway knowledgebase 2022". *Nucleic Acids Research* 50.D1 (Jan. 2022), pp. D687–D692. DOI: 10.1093/nar/gkab1028.
- [42] M. Kanehisa. "KEGG: Kyoto Encyclopedia of Genes and Genomes". *Nucleic Acids Research* 28.1 (Jan. 2000), pp. 27–30. DOI: 10.1093/nar/28.1.27.
- [43] David S. Wishart et al. "PathBank: a comprehensive pathway database for model organisms". *Nucleic Acids Research* 48.D1 (Jan. 2020), pp. D470–D478. DOI: 10.1093/NAR/GKZ861.
- [44] Saikat Chowdhury and Ram Rup Sarkar. "Comparison of human cell signaling pathway databases—evolution, drawbacks and challenges". *Database* 2015 (Jan. 2015), p. 126. DOI: 10.1093/DATABASE/BAU126.
- [45] Sarah Mubeen et al. "The Impact of Pathway Database Choice on Statistical Enrichment Analysis and Predictive Modeling". *Frontiers in Genetics* 10 (Nov. 2019). DOI: 10.3389/fgene.2019.01203.
- [46] Donny Soh et al. "Consistency, comprehensiveness, and compatibility of pathway databases". *BMC Bioinformatics* 11.1 (Dec. 2010), p. 449. DOI: 10.1186/1471-2105-11-449.
- [47] Denise N. Slenter et al. "WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research". *Nucleic Acids Research* 46.D1 (Nov. 2018), pp. D661–D667. DOI: 10.1093/nar/gkx1064.
- [48] Jeffrey J. Sutherland et al. "Assessing Concordance of Drug-Induced Transcriptional Response in Rodent Liver and Cultured Hepatocytes". *PLOS Computational Biology* 12.3 (Mar. 2016), e1004847. DOI: 10.1371/JOURNAL.PCBI.1004847.
- [49] Giulia Callegaro et al. "The human hepatocyte TXG-MAPr: gene co-expression network modules to support mechanism-based risk assessment". *Archives of Toxicology* 95.12 (Dec. 2021), pp. 3745–3775. DOI: 10.1007/S00204-021-03141-W.
- [50] Laura A. Saarimäki et al. "Prospects and challenges for FAIR toxicogenomics data". *Nature Nanotechnology* 2021 17:1 17.1 (Dec. 2021), pp. 17–18. DOI: 10.1038/s41565-021-01049-1.
- [51] Heng Hong Li et al. "TGx-DDI, a Transcriptomic Biomarker for Genotoxicity Hazard Assessment of Pharmaceuticals and Environmental Chemicals". *Frontiers in Big Data* 2 (Oct. 2019), p. 36. DOI: 10.3389/FDATA.2019.00036.
- [52] Henrik Johansson et al. "A genomic biomarker signature can predict skin sensitizers using a cell-based in vitro alternative to animal tests". *BMC Genomics* 12.1 (Aug. 2011), pp. 1–19. DOI: 10.1186/1471-2164-12-399.
- [53] Marvin Martens et al. "A Community-Driven, Openly Accessible Molecular Pathway Integrating Knowledge on Malignant Pleural Mesothelioma". *Frontiers in Oncology* 12 (Apr. 2022), p. 1. DOI: 10.3389/fonc.2022.849640.
- [54] Pekka Kohonen et al. "A transcriptomics data-driven gene space accurately predicts liver cytopathology and drug-induced liver injury". *Nature Communications* 8.1 (July 2017), pp. 1–15. DOI: 10.1038/ncomms15932.
- [55] Yuan Jin et al. "High throughput data-based, toxicity pathway-oriented development of a quantitative adverse outcome pathway network linking AHR activation to lung damages". *Journal of Hazardous Materials* 425 (Mar. 2022), p. 128041. DOI: 10.1016/J.JHAZMAT.2021.128041.

-
- [56] Paul Jennings. "Stress response pathways, toxicity pathways and adverse outcome pathways". *Archives of Toxicology* 87.1 (Jan. 2013), pp. 13–14. DOI: 10.1007/s00204-012-0974-4.
- [57] Bas ter Braak et al. "Systematic transcriptome-based comparison of cellular adaptive stress response activation networks in hepatic stem cell-derived progeny and primary human hepatocytes". *Toxicology in Vitro* 73 (June 2021), p. 105107. DOI: 10.1016/J.TIV.2021.105107.
- [58] Tanya Barrett et al. "NCBI GEO: archive for functional genomics data sets—update". *Nucleic acids research* 41.Database issue (Jan. 2013). DOI: 10.1093/NAR/GKS1193.
- [59] Ron Edgar, Michael Domrachev, and Alex E. Lash. "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository". *Nucleic Acids Research* 30.1 (Jan. 2002), pp. 207–210. DOI: 10.1093/NAR/30.1.207.
- [60] H. Parkinson et al. "ArrayExpress—a public database of microarray experiments and gene expression profiles". *Nucleic Acids Research* 35.Database issue (Jan. 2007), p. D747. DOI: 10.1093/NAR/GKL995.
- [61] Mark D. Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific Data* 3.1 (Mar. 2016), p. 160018. DOI: 10.1038/sdata.2016.18.
- [62] Anna-Lena Lamprecht et al. "Towards FAIR principles for research software". *Data Science* 3.1 (June 2020). Ed. by Paul Groth, Paul Groth, and Michel Dumontier, pp. 37–59. DOI: 10.3233/DS-190026.
- [63] Marvin Martens et al. "ELIXIR and Toxicology: a community in development [version 2; peer review: 2 approved]". *F1000Research* 10 (Oct. 2023), p. 1129. DOI: 10.12688/f1000research.74502.1.
- [64] Nina Jeliaskova et al. "Towards FAIR nanosafety data". *Nature Nanotechnology* 16.6 (June 2021), pp. 644–654. DOI: 10.1038/s41565-021-00911-6.
- [65] N. A. Krans et al. "FAIR assessment tools: evaluating use and performance". *NanoImpact* 27 (July 2022), p. 100402. DOI: 10.1016/J.IMPACT.2022.100402.
- [66] Richard Cyganiak, David Wood, and Markus Lanthaler. *RDF 1.1 Concepts and Abstract Syntax*. 2014.
- [67] Florian Bauer and Martin Kaltenböck. *Linked Open Data: The Essentials A Quick Start Guide for Decision Makers*. 2012.
- [68] Egon L Willighagen et al. "The ChEMBL database as linked open data". *Journal of Cheminformatics* 5.5 (Dec. 2013), p. 23. DOI: 10.1186/1758-2946-5-23.
- [69] Simon Jupp et al. "The EBI RDF platform: linked open data for the life sciences." *Bioinformatics (Oxford, England)* 30.9 (May 2014), pp. 1338–1339. DOI: 10.1093/bioinformatics/btt765.
- [70] Núria Queralt-Rosinach et al. "DisGeNET-RDF: Harnessing the innovative power of the Semantic Web to explore the genetic basis of diseases". *Bioinformatics* 32.14 (2016), pp. 2236–2238. DOI: 10.1093/bioinformatics/btw214.
- [71] Fredo Erxleben et al. "Introducing Wikidata to the Linked Data Web". *Mika P. et al. (eds) The Semantic Web – ISWC 2014. ISWC 2014. Lecture Notes in Computer Science*. Vol. 8796. Springer, Cham, Oct. 2014, pp. 50–65. DOI: 10.1007/978-3-319-11964-9_4.
- [72] Marvin Martens, Chris T. Evelo, and Egon L. Willighagen. "Providing Adverse Outcome Pathways from the AOP-Wiki in a Semantic Web Format to Increase

- Usability and Accessibility of the Content". *Applied in vitro toxicology* 8.1 (Mar. 2022), pp. 2–13. DOI: 10.1089/AIVT.2021.0010.
- [73] P. Langley et al. *Aop-Db Frontend: A User Interface for the Adverse Outcome Pathways Database* (2017). Tech. rep. RTP, NC: Genetics and Environmental Mutagenesis Society Fall Meeting, 2017.
- [74] Maureen E. Pittman et al. "AOP-DB: A database resource for the exploration of Adverse Outcome Pathways through integrated association networks". *Toxicology and Applied Pharmacology* 343 (Mar. 2018), pp. 71–83. DOI: 10.1016/j.taap.2018.02.006.
- [75] Holly M. Mortensen et al. "The AOP-DB RDF: Applying FAIR Principles to the Semantic Integration of AOP Data Using the Research Description Framework". *Frontiers in Toxicology* 4 (Feb. 2022), pp. 1–6. DOI: 10.3389/ftox.2022.803983.
- [76] Andra Waagmeester et al. "Wikidata as a knowledge graph for the life sciences". *eLife* 9 (Mar. 2020). DOI: 10.7554/eLife.52614.
- [77] Yoshinobu Igarashi et al. "Open TG-GATES: a large-scale toxicogenomics database". *Nucleic Acids Research* 43.Database issue (Jan. 2015), p. D921. DOI: 10.1093/NAR/GKU955.
- [78] C. Glenn Begley and John P.A. Ioannidis. "Reproducibility in science: Improving the standard for basic and preclinical research". *Circulation Research* 116.1 (2015), pp. 116–126. DOI: 10.1161/CIRCRESAHA.114.303819.
- [79] Remzi Celebi et al. "Towards FAIR protocols and workflows: the OpenPRE-DICT use case". *PeerJ. Computer science* 6 (2020), pp. 1–29. DOI: 10.7717/PEERJ-CS.281.
- [80] Malcolm Atkinson et al. "Scientific workflows: Past, present and future". *Future Generation Computer Systems* 75 (Oct. 2017), pp. 216–227. DOI: 10.1016/J.FUTURE.2017.05.041.
- [81] Carole Goble et al. "Fair computational workflows". *Data Intelligence* 2.1-2 (Jan. 2020), pp. 108–121. DOI: 10.1162/DINT_A_00033.
- [82] Paolo Budroni, Jean Claude-Burgelman, and Michel Schouppe. "Architectures of Knowledge: The European Open Science Cloud". *ABI Technik* 39.2 (2019), pp. 130–141. DOI: 10.1515/abitech-2019-2006.
- [83] *OECD Series on Adverse Outcome Pathways | OECD iLibrary.*
- [84] Sebastian Hoffmann et al. "Non-animal methods to predict skin sensitization (I): the Cosmetics Europe database". *Critical reviews in toxicology* 48.5 (May 2018), pp. 344–358. DOI: 10.1080/10408444.2018.1429385.
- [85] Nicole C. Kleinstreuer et al. "Non-animal methods to predict skin sensitization (II): an assessment of defined approaches *". *Critical reviews in toxicology* 48.5 (May 2018), pp. 359–374. DOI: 10.1080/10408444.2018.1429386.
- [86] J. M. Fitzpatrick and G. Patlewicz. "Application of IATA–A case study in evaluating the global and local performance of a Bayesian network model for skin sensitization". *SAR and QSAR in Environmental Research* 28.4 (Apr. 2017), pp. 297–310. DOI: 10.1080/1062936X.2017.1311941.
- [87] Sylvia E. Escher et al. "Integrate mechanistic evidence from new approach methodologies (NAMs) into a read-across assessment to characterise trends in shared mode of action". *Toxicology in Vitro* 79 (Mar. 2022), p. 105269. DOI: 10.1016/J.TIV.2021.105269.

-
- [88] Wanda van der Stel et al. "New approach methods (NAMs) supporting read-across: Two neurotoxicity AOP-based IATA case studies". *ALTEX - Alternatives to animal experimentation* 38.4 (Oct. 2021), pp. 615–635. DOI: 10.14573/ALTEX.2103051.
- [89] Antonio Hernandez-Jerez et al. "Development of Integrated Approaches to Testing and Assessment (IATA) case studies on developmental neurotoxicity (DNT) risk assessment". *EFSA Journal* 19.6 (June 2021), e06599. DOI: 10.2903/J.EFSA.2021.6599.
- [90] Lola Bajard et al. "Application of AOPs to assist regulatory assessment of chemical risks – Case studies, needs and recommendations". *Environmental Research* 217 (Jan. 2023), p. 114650. DOI: 10.1016/J.ENVRES.2022.114650.
- [91] Clemens Wittwehr et al. "How adverse outcome pathways can aid the development and use of computational prediction models for regulatory toxicology". *Toxicological Sciences* 155.2 (Feb. 2017), pp. 326–336. DOI: 10.1093/toxsci/kfw207.
- [92] Kirill Degtyarenko et al. "ChEBI: a database and ontology for chemical entities of biological interest." *Nucleic Acids Research* 36 (Jan. 2008), pp. D344–D350. DOI: 10.1093/nar/gkm791.
- [93] Anna Gaulton et al. "ChEMBL: a large-scale bioactivity database for drug discovery." *Nucleic Acids Research* 40.Database issue (Jan. 2012), pp. D1100–D1107. DOI: 10.1093/nar/gkr777.
- [94] Pekka Kohonen et al. "The ToxBank data warehouse: Supporting the replacement of in vivo repeated dose systemic toxicity testing". *Molecular Informatics* 32.1 (Jan. 2013), pp. 47–63. DOI: 10.1002/minf.201200114.
- [95] Alex Bateman and UniProt Consortium. "UniProt: a worldwide hub of protein knowledge". *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D506–D515. DOI: 10.1093/nar/gky1049.
- [96] Andrew D. Yates et al. "Ensembl 2020". *Nucleic Acids Research* 48.D1 (Jan. 2020), pp. D682–D688. DOI: 10.1093/NAR/GKZ966.

Impact

This chapter outlines the impact this thesis has on society. It outlines how this research is being picked up outside academia, and explains how which societal problems this work addresses.

Molecular plausibility of Adverse Outcome Pathways

The main problem that this thesis focused on was the currently limited implementation of transcriptomic data in risk assessments of chemicals and other toxicants, while it has proven potential in studying and understanding toxicological mechanisms. This is mainly due to the complexity of the data, its analysis and interpretation, and the lack of consensus and validation of producing and handling such data. Whereas the production and handling of the data have been addressed by the Organisation for Economic Co-operation and Development (OECD) in the past years resulting in an OECD reporting framework [1], the analysis and interpretation of such data remain unspecified and unformulated. This is why we introduced the molecular Adverse Outcome Pathways (AOPs) to provide a clear, simple method to perform transcriptomic data analysis that is directly aligned with AOPs, which have become an accepted framework of toxicological knowledge to support Integrated Approaches to Testing and Assessment (IATA) development. Additionally, the molecular AOPs can bridge the AOPs and molecular biological and toxicological studies, forming a graphical representation of complex biological systems. By integrating transcriptomic data into AOPs, researchers can better understand the potential adverse effects of exposure and identify early biomarkers of toxicity. This knowledge can inform regulatory decisions and ultimately lead to the development and use of safer chemicals and nanomaterials. Additionally, the integration of transcriptomic data into AOPs can facilitate collaborations among

researchers from diverse fields, leading to more comprehensive and innovative approaches to toxicity assessment. Overall, extending AOPs with molecular entities through pathways and transcriptomic data can accelerate scientific progress and promote the protection of public health and the environment.

With the introduction of molecular AOPs in the WikiPathways database and showing their value in analysing transcriptomic datasets, the results of this thesis provide new and informative ways to analyse and interpret transcriptomic data. Case studies were performed on various AOPs involving mitochondrial dysfunction, liver steatosis, Pleural Mesothelioma (PM), liver cancer, and pulmonary fibrosis by multi-walled carbon nanotubes and by SARS-CoV-2 exposure, thyroid-related neurodevelopmental toxicity, and pharmacovigilance in kidneys. All of these case studies resulted in molecular AOP models, or drafts thereof, that are stored in the AOP Portal on WikiPathways (aop.wikipathways.org). This has shown us that not all toxicological pathways are fully understood yet and that in some cases, the approach to linking pathways to Key Events (KEs) requires additional refinement regarding directionality and differentiating between causal and consequential gene expression changes. While the approach of molecular AOP development is clarified in Chapter 7, we should aim to develop general guidelines for other researchers to engage in molecular AOP development, using them for data analysis of omics data, and compare with other methods of measuring KE activation. This would give us better insights into the validity and generalizability of using transcriptomic data to assess KE activation and generate hypotheses or inform IATA strategies.

While this thesis presents a method of connecting molecular pathways to AOPs to perform analyses of omics data, the process of generating these has its limitations and assumptions, as discussed in Chapter 7. To further explore methods of linking molecular entities to AOPs and their KEs, we have submitted a project proposal to the European Food Safety Authority (EFSA) to test various methods of (semi-)automatically annotating KEs with genes, proteins and

molecular pathways. This proposal, which has been accepted, also involves comparing the approach of molecular AOPs with the more data-driven approach of the TXG-MAPr tool [2].

Reusable AOPs

The thesis also aimed to enhance the usability of AOP content through a Findable, Accessible, Interoperable, and Reusable (FAIR) approach in the AOP-Wiki. This was achieved by implementing the Resource Description Framework (RDF), improving accessibility and interoperability of the AOP-Wiki for seamless integration with other datasets and tools. We maintain a publicly accessible and regularly updated SPARQL endpoint to reflect the latest AOP-Wiki data release. Additionally, in collaboration with the United States Environmental Protection Agency (US EPA), we developed RDF for the AOP-DB as part of the OpenRiskNet implementation challenge. This expanded the available AOP-related content for exploration and integration. The AOP-Wiki RDF and SPARQL Protocol and RDF Query Language (SPARQL) endpoint are utilized in ongoing projects such as VHP4Safety and NanoSolveIT, where virtual infrastructures host the SPARQL endpoints for data integration and AOP-Wiki exploration. These resources are also utilized in the Partnership for the Assessment of Risks from Chemicals (PARC) (eu-parc.eu), a significant European partnership focused on chemical risk assessment for human and environmental protection.

While Chapters 3, 4, 5, and 6 describe, apply and utilize methods to make the contents of the AOP-Wiki more accessible and interoperable, there are other aspects of the AOP-Wiki that can be improved for increased FAIRness. With the release of the AOP-Wiki version 2.5 on July 16, 2022, the resource was made more structured and introduced direct links to third-party tools. This is also the case for the AOP-Wiki SNORQL User Interface (UI) and AOP-DB SPARQL endpoint described in Chapter 4 and 5, respectively. The majority of current efforts in the AOP-Wiki are aimed at complying with the FAIR

principles and increasing data usability, and extensive analysis on the FAIRness of the AOP-Wiki is currently ongoing to explore how the underlying data model can be improved to comply with the FAIR principles. Additionally, the AOP-Wiki has started linking to Wiki Kaptis (wikikaptis.lhasacloud.org), a tool by the UK-based company Lhasa, from KE pages, which is also planned for the molecular pathways of WikiPathways based on the KE-pathway mapping performed for molecular AOPs.

An initiative that preceded this thesis project and focused on the future integration of tools, data and information on AOPs was the development of the AOP Ontology (AOPO) [3]. As its original implementation, the AOPXplorer utilizes the AOPO to visualize AOP networks, and it has been used for studying various AOPs including neurotoxicity [4] and hepatotoxicity [5], with a focus on gene expression data just like the molecular AOPs. However, the AOPO has not yet been implemented to annotate AOP-related content in the AOP-Wiki itself. Therefore, developing the AOP-Wiki RDF model (Chapter 4) also aimed to include the AOPO to annotate relationships within the data model. The chapter also highlighted the potential additions to the AOPO in order to cover the full domain of AOPs and related information, and work is ongoing to expand the AOPO for complete coverage of the AOP-Wiki data model. The alignment of resources with the AOP-Wiki and the direction of its development was also a result of our involvement with the AOP Knowledge Base (AOP-KB) development group, which manages the AOP-Wiki and future developments are discussed. This involvement also offered the opportunity to attend the AOP-KB face-to-face meeting in 2019 at the US EPA in Research Triangle Park, North Carolina, USA, to discuss approaches and share ideas for improving the AOP-Wiki contents and structure.

As is clear from this thesis and the previous paragraphs, the AOP-Wiki plays a central role in this thesis, which is a resource where researchers can collaborate and develop AOPs. Although there are centralized AOP development efforts pushed by the OECD AOP Development Programme work plan, most of the AOPs in the

resource are developed by researchers across the globe, based on individual projects, or large consortia focusing on particular toxicities. We were involved in coordinating and pushing AOP development in EU-ToxRisk (eu-toxrisk.eu), leading to a public deliverable (eu-toxrisk.eu/media/articles/files/EU-ToxRisk_D5.1_FINAL_R1.0.pdf) and a total of twelve AOPs, of which the majority was entered into the AOP-Wiki, including AOPs on adverse effects on the brain, liver, kidneys, lungs and tissue development. These are currently available for AOP users to explore or refine further. These AOPs have the potential to establish AOP-informed IATAs for the risk assessment of a variety of chemicals. Besides our role in EU-ToxRisk, we were involved in the development of COVID-19 AOPs in the community-driven project called Modelling the Pathogenesis of COVID-19 Using the Adverse Outcome Pathway Framework (CIAO), which is described in a later section.

Overall, this thesis had a clear focus on making AOP-related content accessible, interoperable and therefore reusable, in line with the FAIR principles. By making AOP contents more FAIR, researchers, policymakers, and the public can better access and understand the potential risks associated with exposure to certain chemicals, substances, or nanomaterials. This increased accessibility and understanding can lead to more informed decision-making regarding the regulation and use of these substances, ultimately improving public health and safety. Additionally, making AOP content more FAIR can help promote transparency and accountability in the scientific community, leading to more trustworthy and reliable scientific research.

WikiPathways: community collaboration

As described in the section on molecular AOPs, the WikiPathways database served as the second main resource of this thesis, which is widely used in various biological research fields, as an information resource, an integration resource, and to perform analyses of omics

datasets. As presented in Chapter 2, the focus of WikiPathways' current and future developments is on the involvement of user communities. This community-driven aspect of WikiPathways can have a significant societal impact by promoting open and collaborative knowledge-sharing in the field of biological pathways. By allowing researchers, educators, and the public to contribute to and access high-quality pathway information, WikiPathways helps to disseminate scientific knowledge and promote openness and responsibility in scientific research and knowledge sharing. This can lead to faster and more accurate scientific discoveries, improved education efforts, and ultimately, better health outcomes for individuals and communities. Additionally, the community-driven aspect of WikiPathways can help foster a sense of belonging and shared purpose among individuals interested in advancing the field of biology, leading to more robust and impactful collaborations. WikiPathways has also been utilized in the development of a literature-based molecular pathway of PM, a rare type of lung cancer (wikipathways.org/instance/WP5087) [6]. The development of this molecular pathway of PM allows researchers to analyse and interpret their data, and the pathway figure can serve as an educational resource for understanding the molecular aspects of the disease, and this will be extended with the development of a molecular AOP of asbestos leading to PM. This ultimately aids in the increased awareness of the biological complexity and causes of the disease and therefore, supports the research toward better diagnosis, prognosis and treatment.

One project that had a specific interest in making the AOP-Wiki more interoperable with other databases such as WikiPathways is the CIAO project (ciao-covid.net), consisting of a network of partners from industry, policymakers, clinicians, and academic institutes. This project focused on the development of AOP networks for COVID-19, and evaluating the AOP-Wiki data model through workshops and case studies regarding the data structure, annotation of its components, and overall FAIRness. During this project, a molecular AOP model was developed for ACE2 inhibition leading to

pulmonary fibrosis, which was used to analyse transcriptomic data. This work was presented during the final workshop in February 2023.

In summary, WikiPathways played a key role in this thesis and related parallel projects, acting as a central hub that fosters collaboration between academia, research institutes, and industry, and facilitates the modeling of intricate molecular pathways. The pathways developed through WikiPathways offer valuable resources for analysing diverse datasets and interpreting the complexities of biology. Moreover, the utility of WikiPathways extends to the development of molecular AOPs, introducing an innovative approach to analyse transcriptomic data and evaluate KE activation through gene expression data. The integration of these molecular AOPs into risk assessments has the potential to enhance the practicality of transcriptomic data and, consequently, contribute to the promotion of human and environmental safety.

References

- [1] Joshua A. Harrill et al. "Progress towards an OECD reporting framework for transcriptomics and metabolomics in regulatory toxicology". *Regulatory Toxicology and Pharmacology* 125 (Oct. 2021), p. 105020. DOI: 10.1016/j.yrtph.2021.105020.
- [2] Giulia Callegaro et al. "The human hepatocyte TXG-MAPr: gene co-expression network modules to support mechanism-based risk assessment". *Archives of Toxicology* 95.12 (Dec. 2021), pp. 3745–3775. DOI: 10.1007/S00204-021-03141-W.
- [3] Lyle D. Burgoon. "The AOPOntology: A semantic artificial intelligence tool for predictive toxicology". *Applied In Vitro Toxicology* 3.3 (Sept. 2017), pp. 278–281. DOI: 10.1089/aivt.2017.0012.
- [4] Nicoleta Spinu et al. "Development and analysis of an adverse outcome pathway network for human neurotoxicity". *Archives of Toxicology* 93.10 (Oct. 2019), pp. 2759–2772. DOI: 10.1007/S00204-019-02551-1.
- [5] Emma Arnesdotter et al. "Derivation, characterisation and analysis of an adverse outcome pathway network for human hepatotoxicity". *Toxicology* 459 (July 2021), p. 152856. DOI: 10.1016/J.TOX.2021.152856.
- [6] Marvin Martens et al. "A Community-Driven, Openly Accessible Molecular Pathway Integrating Knowledge on Malignant Pleural Mesothelioma". *Frontiers in Oncology* 12 (Apr. 2022), p. 1. DOI: 10.3389/fonc.2022.849640.

Summary

Risk assessors struggle to keep up with the growing number of chemicals requiring testing. The aim is to shift from costly and ethically challenging animal experimentation to more efficient and humane *in vitro* methods, *in silico* models, and human data for risk assessments and human safety promotion. However, the transition involves the challenges that come with the development of novel techniques as alternatives to animal testing.

In order to support risk assessments, the Adverse Outcome Pathway (AOP) approach has been introduced to capture and organize literature-derived mechanistic knowledge of toxicological processes to guide the paradigm shift in risk assessments toward alternative models. It does this by separating the cascade of biological perturbations upon stressor interaction into smaller, measurable effects called Key Events (KEs). Despite the increasing number of AOPs and the growing momentum of the use of AOPs in risk assessments, there are challenges in validating and incorporating *in vitro* targeted assays and large-scale omics datasets into the testing strategies. Although a promising tool in many fields of biomedical research to study molecular processes such as the understanding of toxicological responses, the production and use of transcriptomic data in risk assessment face barriers related to reproducibility, reliability, and acceptance within the risk assessment community. This is why this thesis had the two aims of improving AOP usability and establishing a method to analyse and interpret transcriptomic data that utilises and extends AOPs to facilitate better insights into KE activation, with the ultimate goal to increase the overall acceptance of transcriptomic data in risk assessments.

Improving AOP usability

Before being able to make a link between transcriptomic data and AOPs, the work presented in this thesis involved the exploration of the overall usability of AOPs which are generally stored in the AOP-Wiki and seek interoperability with the established molecular pathway database called WikiPathways to expand KEs with molecular entities and processes. This has led to an introductory description of WikiPathways and highlighting the strengths of community-driven developments and methods of accessing data in Chapter 2. This was followed by investigating the level of coverage of KEs in AOP-Wiki as molecular pathways in WikiPathways in Chapter 3. This has shown that the majority of early KEs in AOP-Wiki have corresponding molecular pathways in WikiPathways, and that opportunities exist to make the AOP-Wiki more linked to other biological databases by using the ontological annotations of KEs and molecular entities captured in their description. To further increase the usability of AOPs in the AOP-Wiki, this thesis has resulted in a more Findable, Accessible, Interoperable, and Reusable (FAIR) version of the AOP-Wiki by employing semantic web technologies and producing an Resource Description Framework (RDF) version of the data, as described in Chapter 4. As an extension to the AOP-Wiki RDF, various tables of the AOP-DB have been modelled into RDF format using the same principles, which was presented in Chapter 5. The producing of the AOP-Wiki RDF and AOP-DB RDF and loading these into SPARQL Protocol and RDF Query Language (SPARQL) endpoints allow the exploration and integration of the data with external resources and allow computational querying of the contents, which has been illustrated in Chapter 6 as a Jupyter notebook. The flexible, reproducible workflow that is presented accesses and uses a range of public services to find and analyse data to support an AOP of interest, showing the utility of the semantic web versions of AOP-Wiki, AOP-DB, and WikiPathways.

Extending AOPs with molecular pathways

With the establishment of the AOP-Wiki and WikiPathways, Chapter 7 presents the establishment of an analysis method for transcriptomic data, utilising WikiPathways as an integrative platform between AOPs and molecular pathways with an intermediate model called the molecular AOP. It was expected that integrating these biological databases with experimental data holds promise for improving transcriptomic data usability by enabling data interpretation and links to KEs to measure biological processes and KE activation. This integration potentially enables the use of transcriptomics data to support AOP-based risk assessment strategies by providing measurements and visualizations of KE activity. As illustrations and proof of principles, case studies were performed on a liver steatosis AOP network and on an AOP that initiates with mitochondrial complex I inhibition in neuronal cells. This has shown us that there is value in the model to analyse and interpret transcriptomic data and generate hypotheses on KE activation. However, the case studies have shown the challenges of modelling molecular AOPs and their use with more extensive datasets, and require more comprehensive testing to define their domain of applicability and technology readiness level.

Impact of this research

The overall goal of this thesis was to make better use of existing mechanistic knowledge in AOPs and utilise large-scale omics approaches based on *in vitro*, to drive the transition away from animal testing for the risk assessment of chemicals and nanomaterials. The increased accessibility and interoperability of the AOP-Wiki and AOP-DB could lead to more effective use of AOP knowledge, leading to a more efficient establishment of knowledge-driven Integrated Approaches to Testing and Assessment (IATA). Ultimately, this can facilitate better and faster risk assessment approaches to ensure human and environmental safety. Regarding the proposed method of analysing transcriptomic data by utilising molecular AOPs, this novel method can aid the

integration and utility of transcriptomic data in risk assessment, providing a clear analysis and interpretation model to assess KE activation.

Conclusion

This thesis project aimed to enhance the usability of AOPs and transcriptomic data for risk assessment. First, this thesis explored integrating the molecular pathway database of WikiPathways with AOP-Wiki to allow the establishment of a connection between transcriptomic data and AOPs by linking KEs to biological processes represented in molecular pathways. This thesis also emphasized the importance of making the data in AOP-Wiki more interoperable and FAIR. This facilitated an automated workflow in a Jupyter Notebook that could find and analyze experimental data to support a specific AOP of interest. Finally, the introduced molecular AOP approach allows for the visualization and reproducible analysis of transcriptomic data to identify KE activation, which can bridge the gap between big data approaches and the risk assessment community.

Samenvatting

Risicobeoordelaars hebben moeite om het groeiende aantal chemicaliën dat getest moet worden bij te houden. Het doel is om over te stappen van kostbare en ethisch uitdagende dierexperimenten naar efficiëntere en diervriendelijkere *in vitro* methoden, *in silico* modellen en menselijke gegevens voor risicobeoordelingen en bevordering van menselijke veiligheid. Deze overgang brengt echter uitdagingen met zich mee die gepaard gaan met de ontwikkeling van nieuwe technieken als alternatieven voor dierproeven.

Om risicobeoordelingen te ondersteunen, is de benadering van de Adverse Outcome Pathway (AOP) geïntroduceerd om op literatuur gebaseerde mechanismen van toxicologische processen vast te leggen en te organiseren. Hiermee wordt de paradigmaverschuiving in risicobeoordelingen naar alternatieve modellen begeleid. Dit wordt bereikt door de cascade van biologische verstoringen als gevolg van interactie met een stressor op te splitsen in kleinere, meetbare effecten die Key Events (KE's) worden genoemd. Ondanks het groeiende aantal AOP's en het toenemende momentum van het gebruik van AOP's in risicobeoordelingen, zijn er uitdagingen bij het valideren en opnemen van *in vitro* gerichte assays en grootschalige omics-datasets in teststrategieën. Hoewel transcriptomische gegevens een veelbelovend instrument zijn in veel gebieden van biomedisch onderzoek om moleculaire processen zoals toxicologische reacties te bestuderen, worden ze geconfronteerd met belemmeringen op het gebied van reproduceerbaarheid, betrouwbaarheid en acceptatie binnen de risicobeoordelingsgemeenschap. Daarom had dit proefschrift twee doelen: het verbeteren van de bruikbaarheid van AOP's en het vaststellen van een methode voor het analyseren en interpreteren van transcriptomische gegevens en welke gebruikmaakt van AOP's en deze uitbreidt om beter inzicht te bieden in KE-activatie, met als ultiem doel het vergroten van de algehele acceptatie van

transcriptomische gegevens in risicobeoordelingen.

Verbetering van de bruikbaarheid van AOP's

Voordat er een link kon worden gelegd tussen transcriptomische gegevens en AOP's, diende de algehele bruikbaarheid van AOP's verkend te worden. Deze worden over het algemeen opgeslagen in de AOP-Wiki. Er wordt gestreefd naar interoperabiliteit met de gevestigde moleculaire pathway-database genaamd WikiPathways om KE's uit te breiden met moleculaire entiteiten en processen. Dit heeft geleid tot een introductie van WikiPathways en het benadrukken van de sterke punten van door de gemeenschap gedreven ontwikkelingen en methoden om toegang te krijgen tot gegevens, beschreven in Hoofdstuk 2. Vervolgens is in Hoofdstuk 3 de mate van overlap tussen KE's in AOP-Wiki en moleculaire pathways in WikiPathways onderzocht. Dit heeft aangetoond dat de meerderheid van de vroege KE's in AOP-Wiki overeenkomstige moleculaire pathways heeft in WikiPathways. Bovendien zijn er mogelijkheden om de AOP-Wiki meer te verbinden met andere biologische databases door gebruik te maken van de ontologische annotaties van KE's en moleculaire entiteiten die in hun beschrijving zijn vastgelegd. Om de bruikbaarheid van AOP's in de AOP-Wiki verder te vergroten, heeft dit proefschrift geleid tot een meer FAIR (Findable, Accessible, Interoperable, Reusable) versie van de AOP-Wiki door gebruik te maken van semantische webtechnologieën en het produceren van een RDF (Resource Description Framework)-versie van de gegevens, zoals beschreven in Hoofdstuk 4. Na de productie van de AOP-Wiki RDF zijn verschillende tabellen van de AOP Data Base (AOP-DB) gemodelleerd in RDF-formaat met dezelfde principes, wat gepresenteerd is in Hoofdstuk 5. Het genereren van de AOP-Wiki RDF en AOP-DB RDF en het laden van deze data in SPARQL-eindpunten maakt de verkenning en integratie van de gegevens met externe bronnen en het computationeel opvragen van de inhoud mogelijk, zoals geïllustreerd in Hoofdstuk 6 door

middel van een Jupyter-notebook. De gepresenteerde flexibele en reproduceerbare workflow maakt gebruik van een reeks openbare diensten om gegevens te vinden en te analyseren ter ondersteuning van een AOP naar keuze, en toont de bruikbaarheid van de semantische webversies van AOP-Wiki, AOP-DB en WikiPathways.

Uitbreiding van AOP's met moleculaire pathways

Met de oprichting van de AOP-Wiki en WikiPathways presenteert Hoofdstuk 7 de oprichting van een analysemethodiek voor transcriptomische gegevens, waarbij WikiPathways wordt gebruikt als een integratief platform tussen AOP's en moleculaire pathways met een tussenmodel genaamd de moleculaire AOP. Er werd verwacht dat de integratie van deze biologische databases met experimentele gegevens veelbelovend zou zijn om de bruikbaarheid van transcriptomische gegevens te bevorderen door middel van verbeterde gegevensinterpretatie en door koppelingen tussen KE's en biologische processen om de KE-activatie te meten. Door meting en visualisaties van KE-activiteit te bieden, maakt deze integratie het mogelijk om, maakt deze integratie maakt het mogelijk om transcriptomische gegevens te gebruiken ter ondersteuning van op AOP gebaseerde risicobeoordelingsstrategieën. Ter illustratie en principieel bewijs zijn casestudy's uitgevoerd naar een AOP-netwerk van leververvetting en een AOP die begint met remming van mitochondriale complex I in neuronale cellen. Dit heeft aangetoond dat het model veelbelovend is om transcriptomische gegevens te analyseren en interpreteren en om hypothesen te genereren over KE-activatie. De casestudy's hebben echter ook de uitdagingen aangetoond van het modelleren van moleculaire AOP's en hun gebruik. Uitgebreidere testen zijn vereist om het toepassingsgebied en technologische gereedheidsniveau van moleculaire AOP's te definiëren.

Impact van dit onderzoek

Het overkoepelende doel van dit proefschrift was om het bestaande mechanistische kennis in AOP's beter te benutten en grootschalige omics-benaderingen te gebruiken om de overgang van dierproeven naar efficiëntere en diervriendelijkere risicobeoordeling van chemicaliën en nanomaterialen te bevorderen. De verhoogde toegankelijkheid en interoperabiliteit van de AOP-Wiki en AOP-DB kunnen leiden tot effectiever gebruik van AOP-kennis, wat vervolgens kan leiden tot een efficiëntere oprichting van op kennis gebaseerde Integrated Approaches to Testing and Assessment (IATA). Dit kan uiteindelijk betere en snellere risicobeoordelingsbenaderingen faciliteren om de veiligheid en gezondheid van mens en milieu te waarborgen. De voorgestelde methode die gebruikmaakt van moleculaire AOPs om transcriptomische gegevens te analyseren, kan deze nieuwe methode de integratie en bruikbaarheid van transcriptomische gegevens in risicobeoordeling ondersteunen. Hiermee wordt een duidelijk analyse- en interpretatiemodel geboden om KE-activatie te beoordelen.

Conclusie

Dit proefschrift had als doel de bruikbaarheid van AOP's en transcriptomische gegevens voor risicobeoordeling te verbeteren. Ten eerste heeft dit proefschrift onderzocht hoe de moleculaire pathway-database van WikiPathways kan worden geïntegreerd met de AOP-Wiki. De connectie hiertussen zorgt ervoor dat er een verbinding kan worden gelegd tussen transcriptomische gegevens en AOP's door door KE's te koppelen aan moleculaire pathways. Bovendien benadrukte dit proefschrift het belang van het interoperabel en FAIR maken van de gegevens in de AOP-Wiki. Dit maakte een geautomatiseerde workflow mogelijk in een Jupyter Notebook dat experimentele gegevens kon opzoeken en analyseren ter ondersteuning van een specifieke AOP. Ten slotte maakt de geïntroduceerde moleculaire AOP-benadering de visualisatie en

reproduceerbare analyse van transcriptomische gegevens mogelijk, wat de identificatie van KE-activatie kan faciliteren. Dit kan zorgen voor een overbrugging van de kloof tussen big data-benaderingen en de risicobeoordelingsgemeenschap en uiteindelijk een verbetering van de veiligheid en gezondheid van mens en milieu.

List of Abbreviations

AO	Adverse Outcome
AOP	Adverse Outcome Pathway
AOP-KB	AOP Knowledge Base
AOP-O	AOP Ontology
API	Application Programming Interface
CIAO	Modelling the Pathogenesis of COVID-19 Using the Adverse Outcome Pathway Framework
cpAOP	computationally predicted AOP
CTD	Comparative Toxicogenomics Database
DEG	differentially expressed gene
EAGMST	Extended Advisory Group on Molecular Screening and Toxicogenomics
EBI	European Bioinformatics Institute
EFSA	European Food Safety Authority
EOSC	European Open Science Cloud
FAIR	Findable, Accessible, Interoperable, and Reusable
GEO	Gene Expression Omnibus
IATA	Integrated Approaches to Testing and Assessment
IRI	Internationalized Resource Identifier
KE	Key Event
KEC	Key Event Component

KEGG Kyoto Encyclopedia of Genes and Genomes

KER Key Event Relationship

LOD Linked Open Data

MIE Molecular Initiating Event

MoA Mode of Action

NAMs New Approach Methodologies

OECD Organisation for Economic Co-operation and Development

OWL Web Ontology Language

PARC Partnership for the Assessment of Risks from Chemicals

PM Pleural Mesothelioma

qAOP quantitative AOP

RDF Resource Description Framework

SNP Single Nucleotide Polymorphism

SPARQL SPARQL Protocol and RDF Query Language

TG-GATES Toxicogenomics Project-Genomics Assisted Toxicity Evaluation System

UI User Interface

US EPA United States Environmental Protection Agency

XML Extensible Markup Language

Acknowledgments

While a PhD project is typically considered an individual endeavour, my journey as a PhD candidate has been enriched by the incredible support of numerous individuals. From advisors providing guidance to stimulating sparring partners for discussions, invaluable supporters, collaborators who shared their expertise, and those who welcomed me into their academic communities, my heartfelt appreciation extends to a multitude of individuals. The first chapter of my academic career has been significantly shaped by the contributions of these remarkable people, and I am sincerely grateful for their involvement.

This chapter encapsulates the gratitude I feel towards those who have played a significant role in my journey, reflecting not only on the past years but also on the ongoing commitment to collaboration, growth, and shared success. Each person mentioned here has contributed in unique ways, leaving a mark on my academic career. As I transition into the role of a postdoctoral researcher, I look forward to building on the lessons learned and achieving new milestones with these remarkable individuals who have been an integral part of my journey.

I would like to start by expressing my gratitude to my esteemed supervisors, Prof. Dr. Chris Evelo and Dr. Egon Willighagen, for their invaluable contribution to shaping the person I am today. Under your guidance, I found not only mentorship but also constant support that significantly impacted my academic journey. Although the PhD position at BiGCaT I initially applied for was not within my reach, your recognition of my potential led to my inclusion in a project within the group that aligned with my skills and interests. I felt genuinely welcomed, and your continuous support has been the cornerstone of my professional development.

Chris, I am grateful for your pivotal role as my promoter. Your guid-

ance throughout my PhD journey has been instrumental in my growth as a researcher. Your wealth of knowledge and expertise, coupled with your ability to provide constructive feedback and make connections, has been crucial. Approving my extended time as a PhD candidate at BiGCaT and granting me the opportunity to continue as a postdoc in the department is a confirmation of your belief in my capabilities.

Egon, your commitment as a co-promotor to bringing out the best in me by allowing me the autonomy to learn from my mistakes and explore diverse opportunities has been extremely valuable. Your guidance not only taught me the intricacies of the research world but also the values of doing open science. Your steadfast support, even in the face of unexpected challenges and delays, has been invaluable and greatly appreciated. Your mentorship extended beyond the academic realm, allowing me to navigate the complexities of research with confidence. I am truly grateful for the vast network of researchers I became a part of under your influence. Egon, you were always approachable, providing guidance whenever needed. I cannot emphasize enough how crucial your support was during my time as a PhD candidate. Your mentorship has left a permanent mark on my academic journey, and I am sincerely thankful for the profound impact you have had on my professional growth.

I am profoundly grateful to the numerous BiGCaT colleagues who have been integral to my journey at the department, offering support, advice, and engaging discussions that have enriched my work. Foremost, I extend my deepest appreciation to Denise, my close colleague who has shared the office with me since the initiation of my PhD. Her consistent insights and thoughtful discussions have been invaluable. Special thanks to Lars, Freddie, and Susan, who provided expertise and feedback from the start, and for the opportunities to contribute to teaching various courses.

Acknowledging the distinctive contributions of Elisa, Mirella, Ryan, Amadeo, Nuno, Jonathan, Woosub, Martina, Myrtle, Iris, Helena, Finterly, Nhung, Lauren, Duygu, and Laurent—each leaving their

mark during different phases of my PhD before they left BiGCaT. My gratitude extends to Ozan and Elena, postdocs collaborating on the VHP4Safety project with whom I am happy to continue working as a postdoctoral researcher, as well as Jeaphianne and Tooba for their roles in nanosafety projects and the new TXG-MAP project, respectively. Continuing with BiGCaTs that have made an impact on my experience and learnings, a special mention to Aishwarya, Javier, Ammar, and Jente, the current PhD candidates whose enthusiasm and commitment significantly contribute to the office atmosphere. Working with each one of you on various topics and projects has been a pleasure. Lastly, heartfelt thanks to Luc and Petra for their indispensable contributions to the smooth functioning of our research environment and office. Together, this immensely diverse and dedicated group has played a crucial role in shaping my PhD experience.

I extend my sincere gratitude to numerous collaborators who have played pivotal roles in my academic journey outside the department. Clemens, Holly, Stephen, and the entire AOP-Wiki community, thank you for directly involving me with the AOP-KB and facilitating my travel to the USA twice during my PhD. Your collaborative spirit has greatly enriched my experience. Thomas, your support through OpenRiskNet and later through other consortia, has been invaluable. The engaging discussions and collaborative work have been both enriching and rewarding. Iseult, Antreas, Penny, and many others involved in nanosafety projects within the NanoSafety Cluster, I appreciate your contributions and your collaborative efforts have significantly shaped my research in this domain, seeing my work applied in the nanosafety domain. Bob, Marcel, Rabea, and the AOP development team in EU-ToxRisk, have been integral to the start of my journey. Rabea's guidance on managing the main AOP deliverable that I was given responsibility for was pivotal to its success, and Marcel's initiative of a sitting ovation after the presentation of this major effort was a memorable highlight. It was an overall enjoyable collaboration and I look forward to continue working with Bob, Marcel and others on the TXG-

MAP project. Ferdinando and Lydie, thank you for bringing me into mesothelioma research and the fruitful collaborations. Kristina, Anders, Alex, and the entire WikiPathways team, it has been a pleasure working together and being part of the WikiPathways curation team. Our collaborative efforts and joint database publications have been truly rewarding.

I express heartfelt gratitude to the students I had the privilege of working with – Joost, Jelle, Tim, Ado, Ulas, Franziska, Anna Baya, Jeroen, Ches, Martijn, Stefan, Aria, Celine, Julia, Alexandra, Zuzanna, and Pierre. I am grateful for the invaluable lessons in supervision that your contributions have provided, shaping my role as a mentor. Your insights into testing methodologies like developing Adverse Outcome Pathways, transcriptomic data analysis, and toxicological pathway modeling have been instrumental. Some of you have become co-authors on papers, directly impacting the success of my publications. Your dedication and hard work have not only left a lasting impact on our projects but have also significantly influenced my professional growth. Thank you for your involvement in fostering collaboration and contributing to the success of our endeavours.

Ten slotte wil ik mijn familie bedanken. Ik wil mijn oprechte waardering uitdrukken aan mijn ouders, John en Janny, en mijn zus, Nikita, voor hun steun gedurende en voorafgaand aan mijn academische reis. Jullie constante aanmoediging, begrip en voortdurende geloof in mijn capaciteiten waren essentieel voor het succesvol voltooien van deze promotie. In momenten van blijdschap, verdriet en uitdaging in mijn professionele en persoonlijke leven is jullie steun een bron van kracht geweest die me heeft voortgedreven. Ik ben dankbaar voor de liefde, begeleiding en aanmoediging die jullie mij hebben gegeven en hebben bijgedragen aan de persoon die ik vandaag ben. Ook wil ik mijn oprechte dank betuigen aan tante Annelize en Eugène, die gedurende de eerste 2,5 jaar van mijn studie in Maastricht hun huis voor me openden en van onschatbare waarde zijn geweest in hun steun, met name aan het begin van mijn academische reis.

Professionele groei en succes vereisen een solide fundering in het persoonlijke leven. Daarom eindig ik dit hoofdstuk met mijn diepste dank uit te spreken aan mijn vrouw, Carolina Prado. Je kwam in mijn leven tijdens de laatste maand van mijn master en hebt aan mijn zijde gestaan sinds het begin van mijn promotieonderzoek. Jouw steun, inzichtelijke perspectieven en constante gezelschap hebben als bouwstenen gediend die me gedurende deze jaren hebben gedragen. Ik ben diep dankbaar voor jouw voortdurende aanwezigheid en de liefde die me omringt door jouw voortdurende steun. De uitdagingen en triomfen van mijn promotiereis kregen betekenis en structuur dankzij jouw constante aanmoediging en geloof in mijn capaciteiten.

Onze verloving, aan het begin van de COVID-19-pandemie was een bevestiging van onze veerkracht en toewijding, en de daaropvolgende viering van ons huwelijk in augustus 2021 markeerde het begin van een nieuw hoofdstuk in ons leven. Carolina, jouw liefde en steun hebben een cruciale rol gespeeld in het vormgeven van mijn academische en persoonlijke leven, en ik kijk uit naar de voortdurende groei en gedeelde successen die voor ons liggen. Niet te vergeten onze hond Diva, die naast mij lag en fungeerde als mentale ondersteuning terwijl ik de afgelopen jaren vanuit huis werkte.

As I reflect on the multitude of individuals and experiences mentioned, I want to emphasize that while these acknowledgements capture moments from the past years, my commitment to collaboration, growth, and shared success persists into the future. As I transition into the role of a postdoctoral researcher at BiGCaT, I look forward to continuing our work together, building on the lessons learned, and achieving new milestones. The invaluable contributions, unwavering support, and shared endeavours have not only shaped my academic journey but have also laid a foundation for ongoing collaboration. Thank you for being an integral part of my journey, and I eagerly anticipate the continued opportunities for collaboration and success that lie ahead.

About the author

Marvin Martens was born on the 23rd of September, 2023, in Oosterhout, the Netherlands, and he grew up in the small town of Raamsdonk. After completing his high school education in 2011 at Dongemonnd College in Raamsdonksveer, where he focused on the nature and health study profile, he relocated to Banholt to commence his Bachelor studies in Biomedical Sciences at Maastricht University. During the final year of his Bachelor's, he moved to the city of Maastricht, and he conducted a thesis internship with Dr. Jos Adam, exploring the development of inhibitory control in the human brain through the application of cognitive assessment tools.

In 2014, he initiated his Master's in Biomedical Sciences at Maastricht University. His first Master's internship, supervised by Prof. Dr. Frederik-Jan van Schooten and Dr. Rianne Fijten at the Department of Toxicology, focused on studying cisplatin resistance in non-small cell lung cancer through a combination of *in vitro* experiments and *in silico* approaches. For the final year of his Master's, he embarked on a year-long Erasmus exchange to Université Pierre et Marie Curie in Paris, France, where he participated in the Master's program in Developmental Biology. During this period, his thesis research, under the guidance of Prof. Dr. Delphine Duprez and Dr. Mickael Orgeur, investigated limb development in chickens, combining laboratory experiments with gene annotation using bioinformatics approaches. He successfully graduated in 2016.

He started his PhD studies at the Department of Bioinformatics (BiG-CaT) at Maastricht University in 2017 under the supervision of Prof. Dr. Chris Evelo and Dr. Egon Willigagen. His work as a PhD Candidate involved Adverse Outcome Pathway development, use, FAIRification, and their implementation in workflows and transcriptomic data analysis. He continues his work at the Department of Bioinformatics as a postdoctoral researcher.

Published work

1. Leist M, Ghallab A, Graepel R, Marchan R, Hassan R, Bennekou SH, Limonciel A, Vinken M, Schildknecht S, Waldmann T, Danen E, van Ravenzwaay B, Kamp H, Gardner I, Godoy P, Bois FY, Braeuning A, Reif R, Oesch F, Drasdo D, Höhme S, Schwarz M, Hartung T, Braunbeck T, Beltman J, Vrieling H, Sanz F, Forsby A, Gadaleta D, Fisher C, Kelm J, Fluri D, Ecker G, Zdrazil B, Terron A, Jennings P, van der Burg B, Dooley S, Meijer AH, Willighagen E, **Martens M**, Evelo C, Mombelli E, Taboureau O, Mantovani A, Hardy B, Koch B, Escher S, van Thriel C, Cadenas C, Kroese D, van de Water B, Hengstler JG. Adverse outcome pathways: opportunities, limitations and open questions. *Archives of Toxicology* 91.11 (Nov. 2017), pp. 3477-3505. DOI: 10.1007/s00204-017-2045-3
2. Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, Mélius J, Cirillo E, Coort SL, Digles D, Ehrhart F, Giesbertz P, Kalafati M, **Martens M**, Miller R, Nishida K, Rieswijk L, Waagmeester A, Eijssen LMT, Evelo CT, Pico AR, Willighagen EL. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research* 46.D1 (Jan. 2018), pp. D661-D667. DOI: 10.1093/nar/gkx1064
3. **Martens M**, Verbruggen T, Nymark P, Grafström R, Burgoon LD, Aladjov H, Torres Andón F, Evelo CT, Willighagen EL. Introducing WikiPathways as a Data-Source to Support Adverse Outcome Pathways for Regulatory Risk Assessment of Chemicals and Nanomaterials. *Frontiers in Genetics* 9 (Dec. 2018), p. 661. DOI: 10.3389/fgene.2018.00661
4. **Martens M**, Ammar A, Riutta A, Waagmeester A, Slenter DN, Hanspers K, Miller R, Digles D, Lopes EN, Ehrhart F, Dupuis LJ, Winckers LA, Coort SL, Willighagen EL, Evelo CT,

- Pico AR, Kutmon M. WikiPathways: connecting communities. *Nucleic Acids Research* 49.D1 (Jan. 2021). pp. D613-D621. DOI: 10.1093/nar/gkaa1024
5. Kyzer JL, **Martens M**. Metabolism and Toxicity of Fluorine Compounds. *Chemical Research in Toxicology* 34.3 (Mar. 2021). pp. 678-680. DOI: 10.1021/acs.chemrestox.0c00439
 6. Hanspers K, Kutmon M, Coort SL, Digles D, Dupuis LJ, Ehrhart F, Hu F, Lopes EN, **Martens M**, Pham N, Shin W, Slenter DN, Waagmeester A, Willighagen EL, Winckers LA, Evelo CT, Pico AR. Ten simple rules for creating reusable pathway models for computational analysis and visualization. *PLoS Computational Biology* 17.8 (Aug. 2021). e1009226. DOI: 10.1371/journal.pcbi.1009226
 7. Anfray C, Mainini F, Digifico E, Maeda A, Sironi M, Erreni M, Anselmo A, Ummarino A, Gandoy S, Expósito F, Redrado M, Serrano D, Calvo A, **Martens M**, Bravo S, Mantovani A, Allavena P, Andón FT. Intratumoral combination therapy with poly(I:C) and resiquimod synergistically triggers tumor-associated macrophages for effective systemic antitumoral immunity. *Journal of Immunotherapy of Cancer* 9.9 (Sep. 2021). e002408. DOI: 10.1136/jitc-2021-002408
 8. Murugadoss S, Vinković Vrček I, Pem B, Jagiello K, Judzinska B, Sosnowska A, **Martens M**, Willighagen EL, Puzyn T, Dusinska M, Cimpan MR, Fessard V, Hoet PH. A strategy towards the generation of testable adverse outcome pathways for nanomaterials. *ALTEX* 38.4 (Oct. 2021). pp. 580-594. DOI: 10.14573/al-
tex.2102191
 9. Ostaszewski M, Niarakis A, Mazein A, Kuperstein I, Phair R, Orta-Resendiz A, Singh V, Aghamiri SS, Acencio ML, Glaab E, Ruepp A, Fobo G, Montrone C, Brauner B, Frishman G, Monraz Gómez LC, Somers J, Hoch M, Kumar Gupta S, Scheel J, Borlinghaus H, Czauderna T, Schreiber F, Montagud A, Ponce de

-
- Leon M, Funahashi A, Hiki Y, Hiroi N, Yamada TG, Dräger A, Renz A, Naveez M, Bocskei Z, Messina F, Börnigen D, Fergusson L, Conti M, Rameil M, Nakonecnij V, Vanhoefer J, Schmiester L, Wang M, Ackerman EE, Shoemaker JE, Zucker J, Oxford K, Teuton J, Kocakaya E, Summak GY, Hanspers K, Kutmon M, Coort S, Eijssen L, Ehrhart F, Rex DAB, Slenter D, **Martens M**, Pham N, Haw R, Jassal B, Matthews L, Orlic-Milacic M, Senff-Ribeiro A, Rothfels K, Shamovsky V, Stephan R, Sevilla C, Varusai T, Ravel JM, Fraser R, Ortseifen V, Marchesi S, Gawron P, Smula E, Heirendt L, Satagopam V, Wu G, Riutta A, Golebiewski M, Owen S, Goble C, Hu X, Overall RW, Maier D, Bauch A, Gyori BM, Bachman JA, Vega C, Grouès V, Vazquez M, Porras P, Licata L, Iannuccelli M, Sacco F, Nesterova A, Yuryev A, de Waard A, Turei D, Luna A, Babur O, Soliman S, Valdeolivas A, Esteban-Medina M, Peña-Chilet M, Rian K, Helikar T, Puniya BL, Modos D, Treveil A, Olbei M, De Meulder B, Ballereau S, Dugourd A, Naldi A, Noël V, Calzone L, Sander C, Demir E, Korcsmaros T, Freeman TC, Augé F, Beckmann JS, Hasenauer J, Wolkenhauer O, Willighagen EL, Pico AR, Evelo CT, Gillespie ME, Stein LD, Hermjakob H, D'Eustachio P, Saez-Rodriguez J, Dopazo J, Valencia A, Kitano H, Barillot E, Auffray C, Balling R, Schneider R; COVID-19 Disease Map Community. COVID-19 Disease Map, a computational knowledge repository of virus-host interaction mechanisms. *Molecular Systems Biology* 17.10 (Dec. 2021). e10851. DOI: 10.15252/msb.202110851
10. Mortensen HM, **Martens M**, Senn J, Levey T, Evelo CT, Willighagen EL, Exner T. The AOP-DB RDF: Applying FAIR Principles to the Semantic Integration of AOP Data Using the Research Description Framework. *Frontiers in Toxicology* 4 (Feb. 2022). pp. 1-6. DOI: 10.3389/ftox.2022.803983
 11. Pains A, Campia I, Cronin MTD, Asturiol D, Ceriani L, Exner TE, Gao W, Gomes C, Kruisselbrink J, **Martens M**, Meek MEB, Pamies D, Pletz J, Scholz S, Schüttler A, Spînu N, Villeneuve DL, Wittwehr C, Worth A, Luijten M. Towards a

- qAOP framework for predictive toxicology - Linking data to decisions. *Computational Toxicology* 21 (Feb. 2022). p. 100195. DOI: 10.1016/j.comtox.2021.100195
12. **Martens M**, Evelo CT, Willighagen EL. Providing Adverse Outcome Pathways from the AOP-Wiki in a Semantic Web Format to Increase Usability and Accessibility of the Content. *Applied In Vitro Toxicology* 8.1 (Mar. 2022). pp. 2-13. DOI: 10.1089/aivt.2021.0010
 13. **Martens M**, Kreidl F, Ehrhart F, Jean D, Mei M, Mortensen HM, Nash A, Nymark P, Evelo CT, Cerciello F. A Community-Driven, Openly Accessible Molecular Pathway Integrating Knowledge on Malignant Pleural Mesothelioma. *Frontiers in Oncology* 12:849640 (Apr. 2022). DOI: 10.3389/fonc.2022.849640
 14. Clerbaux LA, Amigó N, Amorim MJ, Bal-Price A, Batista Leite S, Beronius A, Bezemer GFG, Bostroem AC, Carusi A, Coecke S, Concha R, Daskalopoulos EP, De Bernardi F, Edrosa E, Edwards SW, Filipovska J, Garcia-Reyero N, Gavins FNE, Halappanavar S, Hargreaves AJ, Hogberg HT, Huynh MT, Jacobson D, Josephs-Spaulding J, Kim YJ, Kong HJ, Krebs CE, Lam A, Landesmann B, Layton A, Lee YO, Macmillan DS, Mantovani A, Margiotta-Casaluci L, **Martens M**, Masereeuw R, Mayasich SA, Mei LM, Mortensen H, Munoz Pineiro A, Nymark P, Ohayon E, Ojasi J, Paini A, Parissis N, Parvatam S, Pistollato F, Sachana M, Sørli JB, Sullivan KM, Sund J, Tanabe S, Tsaoun K, Vinken M, Viviani L, Waspe J, Willett C, Wittwehr C. COVID-19 through Adverse Outcome Pathways: Building networks to better understand the disease - 3rd CIAO AOP Design Workshop. *ALTEX* 39.2 (Apr. 2022). pp. 322–335. DOI: 10.14573/altex.2112161
 15. Wittwehr C, Clerbaux LA, Edwards S, Angrish M, Mortensen H, Carusi A, Gromelski M, Lekka E, Virvilis V, **Martens M**, Bonino da Silva Santos LO, Nymark P. Why adverse outcome pathways need to be FAIR. *ALTEX* (Aug. 2023). DOI: 10.14573/altex.2307131

-
16. Murugadoss S, Vinković Vrček I, Schaffert A, Paparella M, Pem B, Sosnowska A, Stępnik M, **Martens M**, Willighagen EL, Puzyn T, Roxana Cimpan M, Lemaire F, Mertens B, Dusinska M, Fessard V, Hoet PH. Linking nanomaterial-induced mitochondrial dysfunction to existing adverse outcome pathways for chemicals. *ALTEX* (Aug. 2023). DOI: 10.14573/altex.2305011
 17. **Martens M**, Stierum R, Schymanski EL, Evelo CT, Aalizadeh R, Aladjov H, Arturi K, Audouze K, Babica P, Berka K, Bessems J, Blaha L, Bolton EE, Cases M, Damalas DE, Dave K, Dilger M, Exner T, Geerke DP, Grafström R, Gray A, Hancock JM, Hollert H, Jeliaskova N, Jennen D, Jourdan F, Kahlem P, Klanova J, Kleinjans J, Kondic T, Kone B, Lynch I, Maran U, Martinez Cuesta S, Ménager H, Neumann S, Nymark P, Oberacher H, Ramirez N, Remy S, Rocca-Serra P, Salek RM, Sallach B, Sansone SA, Sanz F, Sarimveis H, Sarntivijai S, Schulze T, Slobodnik J, Spjuth O, Tedds J, Thomaidis N, Weber RJM, van Westen GJP, Wheelock CE, Williams AJ, Witters H, Zdrazil B, Županič A, Willighagen EL. ELIXIR and Toxicology: a community in development. *F1000Research* 10 (Oct. 2023). p. 1129. DOI: 10.12688/f1000research.74502.2
 18. Agrawal A, Balcı H, Hanspers K, Coort SL, **Martens M**, Slenter DN, Ehrhart F, Digles D, Waagmeester A, Wassink I, Abbassi-Daloui T, Lopes EN, Iyer A, Acosta JM, Willighagen LG, Nishida K, Riutta A, Basaric H, Evelo CT, Willighagen EL, Kutmon M, Pico AR. WikiPathways 2024: next generation pathway database. *Nucleic Acids Research* (Nov. 2023). gkad960. DOI: 10.1093/nar/gkad960

Preprints

1. **Martens M**, Meuleman AB, Kearns J, de Windt C, Evelo CT, Willighagen EL. Molecular Adverse Outcome Pathways: towards the implementation of transcriptomics data in risk

assessments. *bioRxiv* (Jan. 2023), p. 2023.03.02.530766. DOI: 10.1101/2023.03.02.530766

2. van Rijn J, **Martens M**, Ammar A, Cimpan MR, Fessard V, Hoet P, Jeliaskova N, Murugadoss S, Vinkovic Vrcek I, Willighagen EL. Exploring Adverse Outcome Pathways for Nanomaterials with semantic web technologies. *ChemRxiv* (July 2023). DOI: 10.26434/chemrxiv-2023-kjcl6236

