Linking exposome and health using computational approaches

By

Nikhil C

LIFE10202104001

The Institute of Mathematical Sciences
Chennai

A thesis submitted to the
Board of Studies in Life Sciences
In partial fulfillment of requirements
for the Degree of

DOCTOR OF PHILOSOPHY

of

HOMI BHABHA NATIONAL INSTITUTE



October 2026

Homi Bhabha National Institute

Recommendations of the Viva Voce Committee

As members of the Viva Voce Committee, we certify that we have read the dissertation prepared by **Nikhil C** entitled: "*Linking exposome and health using computational approaches*" and recommend that it may be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

Date: 27/10/2026

Chair - Prof. Sitabhra Sinha (In-person)

Date: 27/10/2026

Supervisor/Convener - Prof. Areejit Samal (In-person)

Date: 27/10/2026

Member 1 - Prof. Rahul Siddharthan (In-person)

Date: 27/10/2026

Member 2 - Prof. Sandeep Choubey (In-person)

Date: 27/10/2026

Member 3 - Prof. Anshu Bhardwaj (Online)

Date: 27/10/2026

External Examiner - Prof. Dhiraj Kumar (In-person)

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to HBNI.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it may be accepted as fulfilling the dissertation requirement.

Date: 27/10/2026

Place: Chennai

Supervisor

Statement by Author

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at Homi Bhabha National Institute (HBNI) and is deposited in the Library to be made available to borrowers under rules of the HBNI.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgement of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the Competent Authority of HBNI when in his or her judgement the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.



Nikhil C

Declaration

I, hereby declare that the investigation presented in this thesis has been carried out by me. The work is original and has not been submitted earlier as a whole or in part for a degree or diploma at this or any other Institution or University.



Nikhil C

CERTIFICATION ON ACADEMIC INTEGRITY

Undertaking by the Student

1. I **Nikhil C**, HBNI Enrolment No. **LIFE10202104001** hereby undertake that the Thesis, titled "*Linking exposome and health using computational approaches*" is prepared by me and is the original work undertaken by me.

2. I also hereby undertake that this document has been duly checked through a plagiarism detection tool and the document is found to be plagiarism free as per the guidelines of the Institute/ UGC.

3. I am aware and undertake that if plagiarism is detected in my thesis at any stage in the future, suitable penalty will be imposed as per the guidelines of the Institute/ UGC.

Nikhil. C

(27/10/2026)
Signature of the Student with date

Endorsed by the Thesis Supervisor:

I certify that the thesis written by the researcher is plagiarism free as mentioned above by the student.

Areejit Samal

(27/10/2026)

Signature of the Thesis Supervisor with Date and Name: Areejit Samal
Designation: Professor G
Department/ Centre: Computational Biology group
Name of the CI/ OCC: The Institute of Mathematical Sciences (IMSc), Chennai

प्रोफेसर
PROFESSOR

List of Publications arising from the thesis

Journals

Published

1. *ViCEKb: Vitiligo-linked Chemical Exposome Knowledgebase*, **N. Chivukula**, K. Ramesh, A. Subbaroyan, A.K. Sahoo, G.B. Dhanakoti, J. Ravichandran and A. Samal*, Science of the Total Environment, 913: 169711 (2024).
   https://doi.org/10.1016/j.scitotenv.2023.169711
2. *Network toxicology focused investigation on the impacts of inorganic arsenic and cadmium on human and ecosystem health*, **N. Chivukula**†, S.R. Madgaonkar†, K. Ramesh, S. Mangot, P. Karthikeyan, S.R. Marigoudar, K.V. Sharma & A. Samal*, Environmental Toxicology and Chemistry, 45: 468-489 (2026).
   https://doi.org/10.1093/etojnl/vgaf277
3. *A FAIR-compliant resource and toxicology knowledge graph for mechanistic exploration of endocrine disrupting chemicals associated toxicity*, **N. Chivukula**, S. Vashishth, P. Kandasamy, S.R. Madgaonkar and A. Samal*, Journal of Hazardous Materials Advances, 24: 101660 (2026).
   https://doi.org/10.1016/j.hazadv.2026.101660

[† Joint-first authors; * Corresponding author(s)]



Nikhil C

List of Publications not included in the thesis

Journals

Published

1. *EPEK: Creation and analysis of an Ectopic Pregnancy Expression Knowledgebase*, A. Natarajan[†], **N. Chivukula**[†], G.B. Dhanakoti, A.K. Sahoo, J. Ravichandran[⋆] and A. Samal[⋆], Computational Biology and Chemistry, 104: 107866 (2023).
   https://doi.org/10.1016/j.compbiolchem.2023.107866
2. *Analysis of structure-activity and structure-mechanism relationships among thyroid stimulating hormone receptor binding chemicals by leveraging the ToxCast library*, A.K. Sahoo[†], S.P. Baskaran[†], **N. Chivukula**, K. Kumar and A. Samal[⋆], RSC Advances, 13: 23461-23471 (2023).
   https://doi.org/10.1039/D3RA04452A
3. *T9GPred: A Comprehensive Computational Tool for the Prediction of Type 9 Secretion System, Gliding Motility, and the Associated Secreted Proteins*, A.K. Sahoo, R.P. Vivek-Ananth, **N. Chivukula**, S.V. Rajaram, K. Mohanraj, D. Khare, C. Acharya and A. Samal[⋆], ACS Omega, 8: 34091-34102 (2023).
   https://doi.org/10.1021/acsomega.3c05155
4. *Cheminformatics analysis of multi-target structure-activity landscape of environmental chemicals binding to human endocrine receptors*, S.P. Baskaran[†], A.K. Sahoo[†], **N. Chivukula**, K. Kumar and A. Samal[⋆], ACS Omega, 8: 49383-49395 (2023).
   https://doi.org/10.1021/acsomega.3c07920
5. *An integrative data-centric approach to derivation and characterization of an adverse outcome pathway network for cadmium-induced toxicity*, A.K. Sahoo[†], **N. Chivukula**[†], K. Ramesh, J. Singha, S.R. Marigoudar, K.V. Sharma and A. Samal[⋆], Science of The Total Environment, 920: 170968 (2024).
   https://doi.org/10.1016/j.scitotenv.2024.170968
6. *Leveraging integrative toxicogenomic approach towards development of stressor-centric adverse outcome pathway networks for plastic additives*, A.K. Sahoo[†], **N.**

Chivukula[†], S.R. Madgaonkar, K. Ramesh, S.R. Marigoudar, K.V. Sharma and A. Samal[⋆], Archives of Toxicology, 98: 3299–3321 (2024).

https://doi.org/10.1007/s00204-024-03825-z

7. *Network-based investigation of petroleum hydrocarbons-induced ecotoxicological effects and their risk assessment*, A.K. Sahoo[†], S.R. Madgaonkar[†], **N. Chivukula**, P. Karthikeyan, K. Ramesh, S.R. Marigoudar, K.V. Sharma and A. Samal[⋆], Environment International, 194: 109163 (2024).

   https://doi.org/10.1016/j.envint.2024.109163

8. *The dawn of a new air pollutant: inhalable microplastics as emerging vectors of hazardous contaminants and their implications for human health*, A. Biswas, N. Saini, **N. Chivukula**, A. Samal, M. R. Jansari, P. Bhadury and G. K. Darbha[⋆], Environment International, 205: 109897 (2025).

   https://doi.org/10.1016/j.envint.2025.109897

9. *Integrative network-based construction of ecologically relevant computational adverse outcome pathways for organic mercury-induced toxicity*, S.R. Madgaonkar, **N. Chivukula**[#], V. Garisetti[#], S.R. Marigoudar, K.V. Sharma and A. Samal[⋆], Environmental Toxicology and Chemistry, 45: 731-747 (2026).

   https://doi.org/10.1093/etojnl/vgag026

10. *A comprehensive resource on chemicals used in aquaculture and their ecotoxicity*, S.R. Madgaonkar, S. Vashishth, **N. Chivukula**, V. Garisetti, S.R. Marigoudar, K.V. Sharma and A. Samal[⋆], Cell Reports Sustainability, 8: 100794 (2026).

    https://doi.org/10.1016/j.crsus.2026.100794

11. *Benchmarking Artificial Intelligence Models for Predicting Nuclear Receptor Activity from Tox21 Assays*, **N. Chivukula**[†], J. Karthikeyan[†], H. Thangavel[†], S.R. Madgaonkar[†] and A. Samal[⋆], Computational Toxicology, 39: 100439 (2026).

    https://doi.org/10.1016/j.comtox.2026.100439

12. *An integrated resource for systems-level analysis of aging hallmarks and associated genes*, R. Tiwari, M. Balaji, **N. Chivukula**, P. Sil and A. Samal[⋆], Biogerontology,

27: 173 (2026).

https://doi.org/10.1007/s10522-026-10520-2

Submitted

13. *TICToK: A comprehensive knowledgebase of tattoo ink chemicals and investigation of their associated toxicities and regulations*, **N. Chivukula**, S.R. Madgaonkar, S.R. Marigoudar, K.V. Sharma, V. Kishore, A.V. Singh* and A. Samal*, bioRxiv, 2025.08.02.668261 (2025).

https://doi.org/10.1101/2025.08.02.668261

[† Joint-first authors; $^{\#}$ Equal contribution; * Corresponding author(s)]



Nikhil C

Oral or Poster presentations

1. Poster presentation titled '*An enhanced and expanded FAIR-compliant resource and toxicology knowledge graph for endocrine disrupting chemicals*' at Modelling and Tackling Complex Biological Systems 2.0, held at IMSc, Chennai during February 16–18, 2026.
2. Poster presentation titled '*Network toxicology focused investigation on the impacts of inorganic arsenic and cadmium on human and ecosystem health in India*' at Computational Biology Conference 2.0 (CBC2.0) held at Indian Biological Data Center, Regional Centre for Biotechnology, Faridabad, India during 8 – 10 December, 2025.
3. Flash presentation titled '*Computational Biology Meets the Exposome: Transforming Toxicology Without Animal Testing*', at the JRC Summer School on Non-Animal Approaches in Science - Changing the paradigm held at the European Commission's Joint Research Centre (JRC) in Ispra, Italy during 19 - 23 May, 2025.
4. Poster presentation titled '*Data-Driven Insights into Plastic Additives-Induced Toxicities Using Stressor-centric Adverse Outcome Pathway (AOP) Networks*', at the Society of Environmental Toxicology and Chemistry (SETAC) Europe 35^{th} Annual Meeting held in Vienna, Austria during 11 - 15 May, 2025.
5. Oral presentation titled '*Toxicogenomics-driven Insights into Plastic Additives-induced Toxicity through Stressor-centric Adverse Outcome Pathway (AOP) Networks*' at Aspects of Gene and Cellular Regulation (AOGCR 2025), held at IMSc, Chennai during March 27-28, 2025.
6. Poster presentation titled '*Leveraging Integrative Toxicogenomic Approach Towards Development of Stressor-centric Adverse Outcome Pathway Networks for Plastic Additives*', at Micro-Plastics Research Innovations & Sustainable Management (M-PRISM 2025), held at IISER Kolkata, during 15-16 February, 2025. **Awarded the Best Poster Award.**
7. Poster presentation titled '*ViCEKb: Creation and network-based analysis of a*

curated knowledgebase on vitiligo-triggering chemicals to link exposome and health', at NetSciX 2025, held at IIT Indore, during 15-17 January, 2025.

8. Oral presentation titled '*Linking Exposome and Health by Leveraging Data Science*' at Leveraging Data Science to Characterize Plastic Exposome held at IMSc, Chennai on 23 November, 2024.
9. Poster presentation titled '*ViCEKb: Creation and analysis of a curated knowledge-base on vitiligo-triggering chemicals to link exposome and health*' at Big Data Algorithms for Biology (BDBIO 2024), held at IISc Bengaluru during May 31 - June 1 2024.
10. Poster presentation titled '*ViCEKb: Creation and analysis of a curated knowledge-base on vitiligo-triggering chemicals to link exposome and health*', at International Conference on Bioinformatics in Health and Food Security 2024 (ICB-HFS 2024), held at Pondicherry University during March 14-16, 2024.
11. Poster presentation titled '*ViCEKb: Creation and analysis of a curated knowledge-base on vitiligo-triggering chemicals to link exposome and health*' at 17^{th} Annual International Biocuration Conference (AIBC 2024), held at Indian Biological Data Centre, Regional Centre for Biotechnology, Faridabad, during March 6-8, 2024. **Awarded the Best Poster Award.**
12. Oral presentation titled '*Beyond observations, Towards Solutions: Computational Biology as a key to unraveling environmental health impacts*' at the Contemporary Perspectives in Computational Biology, held at IMSc, Chennai during February 19-20, 2024.
13. Poster presentation titled '*EPEK: creation and analysis of an Ectopic Pregnancy Expression Knowledgebase*' and '*ViCEKb: Creation and analysis of a curated knowledgebase on vitiligo-triggering chemicals to link exposome and health*' at Modelling and Tackling Complex Biological Systems, held at IMSc, Chennai during October 13–14, 2023.
14. Poster presentation titled '*T9GPred: A Predictor for Bacterial Type 9 Secretion*

System, Associated Gliding Motility and Secreted Proteins', at Machine Learning for Health and Disease (MLHD 2023), held at ICTS Bengaluru during July 24 - August 4, 2023.

15. Poster presentation titled '*Uncovering Biological Insights of Ectopic Pregnancy through Systems-level Analysis using Expression Data*', at Network Biology Day, held at IMSc, Chennai on July 20, 2023.
16. Poster presentation titled '*T9GPred: A Predictor for Bacterial Type 9 Secretion System, Associated Gliding Motility and Secreted Proteins*', at Big Data Algorithms for Biology (BDBIO 2023), held at IISc Bengaluru during June 2-3, 2023.
17. Poster presentation titled '*Uncovering Biological Insights of Ectopic Pregnancy through Systems-level Analysis using Expression Data*', at IMSc - Celebrating 60 Years of Creative Science, held at IMSc, Chennai during January 2-5, 2023.



Nikhil C

Acknowledgements

First and foremost, I want to thank my advisor Prof. Areejit Samal. His support and guidance have provided a constant source of strength throughout my PhD journey. He always pushed me to strive for the best and provided me with immense opportunities to collaborate both within and outside the lab. I am greatly thankful to him for suggesting ideas that have been developed into various projects during my PhD, some of which form the core of this thesis. I also appreciate the scientific foundation he provided for my growth as a researcher. Beyond science, the many life lessons he imparted will always remain at the core of my identity as a scientist. These lessons will continue to push me forward throughout my career.

Next, I thank my coauthors by specifically listing their contributions to the chapters in this thesis. First and foremost, I am thankful to Dr. Janani Ravichandran. She is a coauthor on the published manuscript included in Chapter 2 and helped with discussions in executing the work in chapter 3. Her extensive work on characterizing the chemical exposome provided the basis for further exposome research in our lab. Although my time at IMSc did not completely overlap with hers our constant discussions helped shape the projects on ViCEKb and the DEDuCT update. I am thankful to Dr. Ajaya Kumar Sahoo for being a great mentor and coauthor across multiple publications during my PhD. He patiently guided me through my initial years. He helped me learn the essential soft skills required to perform and publish research. He also taught me the relevant computational skills that form the basis of the analysis in this thesis. Importantly, he guided me in performing the cheminformatics-based exploration of ViCEKb chemical space and in writing the manuscript included in Chapter 2. I thank Dr. Ajay Subbaroyan for his initial help in collecting data for the ViCEKb work in Chapter 2. I am greatly thankful to Kundhanathan Ramesh. He diligently curated data for the ViCEKb work in Chapter 2 and provided immense help in organizing the evidence to support the toxicity mechanisms in Chapter 4. I also thank Gokul Balaji for helping me design and develop the user interface

for ViCEKb, and Kishan Kumar for designing the logo and icons for the webserver. I am thankful to Shrish Vashishth and Pavithra Kandasamy. They read through the papers and curated data that went into the construction of DEDuCT. The update presented in Chapter 3 would not have been possible without their patient and dedicated help. I also thank Swetha Mangot. She was a summer intern who worked on the initial data curation for the work presented in Chapter 4. I am greatly thankful to my coauthor and friend Shreyes Rajan Madgaonkar, who significantly contributed to the work presented in Chapter 4. The environmental risk assessment was possible entirely because of his contributions, and he was also instrumental in setting up the website for that project. I would like to thank Panneerselvam Karthikeyan from NCCR for his initial help in understanding the species sensitivity distributions presented in Chapter 4. I am also deeply thankful to our collaborators from NCCR including Mr. K.V. Sharma and Dr. S.R. Marigoudar. They are coauthors on Chapter 4. Beyond the work presented in this thesis they have also been involved in a range of research through our ongoing collaboration. They have been particularly supportive of our scientific approaches. Their critical discussions and suggestions have greatly helped shape the various projects.

I thank my doctoral committee members, Prof. Anshu Bhardwaj, Prof. Sandeep Choubey, Prof. Rahul Siddharthan, and Prof. Sitabhra Sinha, whose critical suggestions have helped shape this thesis. I am also thankful to the faculty members who taught me during my coursework, including Prof. Sandeep Choubey, Prof. Vasudharani Devanathan, Prof. Prabuddha Ganguli, Prof. S. Krishnaswamy, Prof. Basant Kumar Nayak, Prof. Areejit Samal, Prof. Rahul Siddharthan, Prof. Sitabhra Sinha, Prof. Vasudeva Siruguri, Prof. B. S. Tomar, and Prof. Satyavani Vemparala.

I am also greatly thankful to the many scientists and faculty who visited IMSc and generously gave me time to discuss my work. Their suggestions and critical comments helped me refine my projects significantly. I am extremely grateful to the systems administrators Mr. Ravindra Reddy, Mr. U. Nagamalleswara Rao, and Mr. G. Srinivasan. They provided immense help in setting up and maintaining the online webservers. I am

also thankful to the IMSc Administration and specifically Ms. Indra R. for her support through various administrative processes and hurdles over the years. I also thank Ms. Prema P. and Ms. Usha Otheeswaran for always helping me with a smile while processing my paperwork. I would also like to thank the Registrar and the administrative office for seamlessly processing my papers. This support was especially helpful for the paperwork involved during my registration and international travel. I thank the librarian and the library staff for swiftly accepting my book procurement suggestions and making these resources available in time for my research. I would like to thank the mess staff for always preparing delicious and healthy food throughout the years. I am also grateful to the housekeeping staff for keeping the campus spotless. Their hard work made the institute a welcoming place to return to time and again. I thank the staff of the civil and electrical departments. They were always helpful in accommodating my requests regarding office furniture or electrical issues. Finally I want to thank the security staff for their warm smiles and for welcoming me every morning. Finally I would like to thank IMSc and the Department of Atomic Energy for their generous funding. This support ensured my research progressed without any hurdles and allowed me to procure the materials required to conduct my studies, and present my research at various national conferences.

I would like to thank the past members of the lab Ajay, Ajaya, BS Karthikeyan, Geetha, Gokul, Ishwarya, Janani, Jayalakshmi, Kaustav, Kishan, Kundhanathan, Pavithra, Pritam, Priya Darshini, Priya Shree, Sabrina, Shashank, Shreelekshmi, Soumitra, Suchetana, Sudharsan, Vinayak, Vivek, and Yasharth for the many shared experiences, and moments that made life in the lab vibrant. I also thank the current members of the lab including Camelia, Harshini, Jaisubasri, Madhumita, Priya, Priyotosh, Rahul, Shreyes, Shrish, and Vasavi. I am grateful for the joyous atmosphere they created in the office. I am specifically thankful to Camelia, Madhumita, Priyotosh, Shreyes, and Shrish for helping me with proof-reading this thesis, in midst of their busy schedules.

On a personal note, I want to thank my initial guiding light, Ms. Santhi Sathiapalan, whose guidance during my high school years inspired me to choose biological research as

a career path and motivated me to strive to become both a researcher and a great teacher. I am also thankful to my lifelong friends, Abhijit, Aishwarya, Kausthub, Nandan and Rishi, who are now more like family. Their support during the low points of my college years and my PhD, and our frequent chats, always relieved my stress and helped me refocus on my work.

Finally, I come to my family. The two strongest pillars in my life are my mom, Lakshmi and my dad, Sashikanth, who have always stood by all my decisions and supported me completely throughout this journey. I was fortunate enough to conduct my research while staying at home with them, which made this journey all the more meaningful. My younger brother, Chinmaya, has been another pillar in my life, always offering a different perspective on my situations and helping me get through tough times. I would also like to thank all my grandparents, Dharma Rao, Saraswathi, Swarajyalakshmi, for their unwavering support. I know my late grandfather, Seshagiri, would be greatly delighted to see me reach this stage of my career as he had always encouraged my pursuit of knowledge and would have been proud to see this goal fulfilled.



Nikhil C

Contents

List of Figures i

List of Tables vi

Abstract vii

1 Introduction 1

1.1 Background and motivation 1

1.2 Thesis organization 13

2 ViCEKb: Vitiligo-linked Chemical Exposome Knowledgebase 17

2.1 Methods 18

2.1.1 Workflow for the compilation of chemical triggers of vitiligo with supporting evidence in published literature 18

2.1.2 Curation and compilation of data from shortlisted literature 20

2.1.3 Classification of vitiligo-triggering chemicals 21

2.1.4 Physicochemical properties, molecular descriptors and predicted ADMET properties of chemicals 22

2.1.5 Comparison of vitiligo-triggering chemicals with consumer product database and chemical regulations 23

2.1.6 Identification of target genes of vitiligo-triggering chemicals based on ToxCast skin sensitization assays 23

2.1.7 Exploration of the space of vitiligo-triggering chemicals 24

2.1.8 Analysis of transcriptomics signatures of vitiligo-triggering chemicals 25

2.1.9 Web interface and database management system 25
2.2 Results 26
2.2.1 ViCEKb - Vitiligo-linked Chemical Exposome Knowledgebase . . 26
2.2.2 ViCEKb chemicals are present in consumer products 29
2.2.3 ViCEKb shows low overlaps with skin sensitizing chemical lists and exposome databases 32
2.2.4 Cheminformatics-based analysis of ViCEKb chemical space revealed its uniqueness 34
2.2.5 Vitiligo-triggering chemicals are associated with similar biological responses 37
2.3 Discussion 40

3 DEDuCT 3.0: An enhanced and expanded FAIR-compliant resource and toxicology knowledge graph for endocrine disrupting chemicals 43
3.1 Methods 45
3.1.1 Workflow to identify endocrine disrupting chemicals 45
3.1.2 Enriching DEDuCTv3.0 with additional chemical information for EDCs 49
3.1.3 Exploration of chemical space within DEDuCTv3.0 50
3.1.4 Construction of stressor-centric adverse outcome pathway (AOP) networks for chemicals in DEDuCTv3.0 52
3.1.5 Construction of a toxicology knowledge graph for chemicals in DEDuCTv3.0 58
3.1.6 Exploration of potential EDCs across consumer products and chemical lists 64
3.1.7 Web interface and database management system 65
3.2 Results 66
3.2.1 DEDuCTv3.0 - an enhanced resource on potential endocrine disrupting chemicals 66

3.2.2 Exploration of EDC-relevant toxicity mechanisms through stressor-AOP network 70
3.2.3 DEDuCT-KG: A toxicology knowledge graph on EDCs 77
3.2.4 Potential exposure to EDCs through consumer products 86
3.2.5 Distribution of potential EDCs in chemical lists 87
3.2.6 Enhancements to DEDuCT webserver 89
3.3 Discussion 94

4 Network toxicology focused investigation on the impacts of inorganic arsenic and cadmium on human and ecosystem health in India 97
4.1 Methods 99
4.1.1 Compilation and curation of inorganic compounds of arsenic and cadmium 99
4.1.2 Construction of aggregate exposure pathways (AEPs) for inorganic arsenic and cadmium in India 99
4.1.3 Compilation and curation of AOPs within AOP-Wiki 100
4.1.4 Construction of stressor-AOP network for inorganic arsenic and cadmium 105
4.1.5 Construction of stressor-species networks by utilizing ECOTOX data 112
4.1.6 Construction of species sensitivity distributions (SSDs) for inorganic arsenic and cadmium compounds by utilizing ECOTOX data 113
4.1.7 Construction of toxicity-normalized SSD (SSDn) for inorganic arsenic and cadmium 115
4.1.8 Calculation of risk quotient (RQ) based on the environmental concentrations of arsenic and cadmium 117
4.2 Results 119
4.2.1 Inorganic arsenic and cadmium compounds curated from six exposome-relevant datasets 119
4.2.2 AEP networks for inorganic arsenic and cadmium in India 121

4.2.3 Stressor-AOP network for inorganic arsenic and cadmium 124
4.2.4 Taxonomic applicability of arsenic- and cadmium-AOPs 135
4.2.5 Exploring AEP-AOP constructs associated with inorganic arsenic and cadmium 135
4.2.6 Stressor-species networks associating impacts of inorganic arsenic and cadmium compounds with ecologically relevant species . . . 140
4.2.7 SSDs for inorganic arsenic and cadmium compounds 143
4.2.8 RQ-based analysis revealed ecotoxicity of arsenic and cadmium in Indian rivers 166
4.3 Discussion 169

5 Summary and future outlook **173**
5.1 Summary 173
5.2 Future outlook 178

References **181**

List of Figures

1.1 Schematic representation of network toxicology frameworks 8

2.1 Detailed workflow to identify and curate chemical triggers of vitiligo from published literature 19

2.2 Overview of chemical information present in ViCEKb 27

2.3 ViCEKb webserver screenshots for HOME and BROWSE pages 29

2.4 ViCEKb webserver screenshots for SEARCH and DOWNLOADS pages . 31

2.5 ViCEKb webserver screenshots for chemical information pages 31

2.6 Comparison of ViCEKb chemicals with consumer products, skin-specific chemical regulations and other exposome-relevant databases 33

2.7 Chemical similarity network (CSN) of chemicals in ViCEKb 35

2.8 Diversity and uniqueness of chemicals in ViCEKb 36

2.9 Transcriptomics-based analysis of chemicals in ViCEKb 38

3.1 Workflow to identify published literature containing potential endocrine disrupting chemicals and their associated endocrine-mediated endpoints . 46

3.2 Network representation of different associations between seven types of nodes present in DEDuCT Knowledge Graph (DEDuCT-KG) 59

3.3 Distribution of endocrine-mediated endpoints and chemicals across systems-level perturbation categories, and chronological analysis of published literature and chemicals in DEDuCTv3.0 68

3.4 Comparison of chemicals within DEDuCTv3.0 and other EDC-relevant resources 69

3.5 Chemical similarity network (CSN) of chemicals within DEDuCTv3.0 . . 71

3.6 Scaffold cloud representation of chemicals within DEDuCTv3.0 72
3.7 Target similarity network (TSN) of chemicals within DEDuCTv3.0 73
3.8 Target similarity vs structural similarity between pairs of chemicals within DEDuCTv3.0 74
3.9 Distribution of 250 chemicals in DEDuCTv3.0 across 28 disease categories based on the AO classification of their associated AOPs at Level 5 relevance 76
3.10 Screenshot of interactive user interface of DEDuCT-KG 78
3.11 Chemical-gene-disease-endpoint subgraphs for obesogen case study . . . 80
3.12 Chemical-gene-disease-endpoint subgraphs for Alzheimer's disease case study 84
3.13 Chemical-gene-disease-endpoint subgraphs for Parkinson's disease case study 85
3.14 Distribution of 2049 unique potential EDCs from DEDuCTv3.0, TEDx, EDCs Databank, and WHO report across 54 chemical lists 88
3.15 Screenshot of the updated chemical information page in DEDuCT webserver 90
3.16 Screenshot from DEDuCT webserver of the updated chemical information pages associated genes, phenotypes and diseases 91
3.17 Screenshot from DEDuCT webserver of the updated chemical information page on associated ToxCast endpoints 92
3.18 Screenshot from DEDuCT webserver of the updated chemical information page on associated data from the Chemical and Products Database (CPDat) 92
3.19 Screenshots from DEDuCT webserver on the updated search options . . . 93

4.1 Workflow to curate AEP-relevant information associated with arsenic and cadmium in India 102
4.2 Workflow to curate complete and connected AOPs from AOP-Wiki 103
4.3 Summary of the network toxicology workflow utilized to investigate toxicological impacts of arsenic and cadmium 120
4.4 AEP network corresponding to arsenic in the Indian context 122

4.5 AEP network corresponding to cadmium in the Indian context 123
4.6 Undirected AOP network for 51 arsenic-AOPs 126
4.7 Undirected AOP network for 78 cadmium-AOPs 127
4.8 Directed network corresponding to the LCC in the arsenic-AOP network . 129
4.9 Directed network corresponding to the LCC in the arsenic-AOP network, where the KEs are colored based on their betweenness centrality values . 130
4.10 Directed network corresponding to the LCC in the arsenic-AOP network, where the KEs are colored based on their eccentricity values 131
4.11 Directed network corresponding to the LCC in the cadmium-AOP network 132
4.12 Directed network corresponding to the LCC in the cadmium-AOP network, where the KEs are colored based on their betweenness centrality values . 133
4.13 Directed network corresponding to the LCC in the cadmium-AOP network, where the KEs are colored based on their eccentricity values 134
4.14 AEP-AOP constructs explored in case studies of arsenic- and cadmium-induced toxicities 137
4.15 Sankey plot depicting associations between inorganic arsenic and cadmium compounds, and ECOTOX species groups through their acute and chronic toxicity data 141
4.16 Sankey plot depicting associations between inorganic arsenic and cadmium compounds, and ECOTOX species groups through their bioconcentration factors (BCFs) 142
4.17 Stressor-species network constructed for inorganic arsenic compounds using the chronic toxicity concentrations as edge weights 145
4.18 Plots of SSD for arsenic oxide (CAS:1327-53-3) based on acute toxicity data 146
4.19 Plots of SSD for disodium arsenate (CAS:7778-43-0) based on acute toxicity data 147
4.20 Plots of SSD for sodium arsenite (CAS:7784-46-5) based on acute toxicity data 148

4.21 Plots of SSD for trisodium arsenate (CAS:13464-38-5) based on acute toxicity data 149
4.22 Plots of SSD for cadmium (CAS:7440-43-9) based on acute toxicity data . 150
4.23 Plots of SSD for cadmium chloride (CAS:10108-64-2) based on acute toxicity data 151
4.24 Plots of SSD for cadmium sulfate (CAS:10124-36-4) based on acute toxicity data 152
4.25 Plots of SSD for cadmium nitrate (CAS:10325-94-7) based on acute toxicity data 153
4.26 Plots of SSDn for inorganic arsenic based on acute toxicity data 154
4.27 Plots of SSDn for inorganic cadmium based on acute toxicity data 155
4.28 Plots of SSD for arsenic oxide (CAS:1327-53-3) based on chronic toxicity data 157
4.29 Plots of SSD for disodium arsenate (CAS:7778-43-0) based on chronic toxicity data 158
4.30 Plots of SSD for sodium arsenite (CAS:7784-46-5) based on chronic toxicity data 159
4.31 Plots of SSD for cadmium (CAS:7440-43-9) based on chronic toxicity data 160
4.32 Plots of SSD for cadmium chloride (CAS:10108-64-2) based on chronic toxicity data 161
4.33 Plots of SSD for cadmium sulfate (CAS:10124-36-4) based on chronic toxicity data 162
4.34 Plots of SSD for cadmium nitrate (CAS:10325-94-7) based on chronic toxicity data 163
4.35 Plots of SSDn for inorganic arsenic based on chronic toxicity data 164
4.36 Plots of SSDn for inorganic cadmium based on chronic toxicity data . . . 165
4.37 Comparative analysis of risk quotients (RQs) of arsenic and cadmium . . 168

5.1 Summary of work done in characterizing chemical exposome of vitiligo . 174

5.2 Summary of work done in updating and gaining mechanistic insights from DEDuCTv3.0 175
5.3 Summary of work done in network toxicology-based investigation of inorganic arsenic- and cadmium-induced adverse effects on human and ecosystem health 177

List of Tables

2.1 Comparative analysis of the scaffold diversity of the chemicals in ViCEKb with other chemical lists 37
2.2 Overview of L1000 signatures of 18 ViCEKb chemicals 39

4.1 Summary of resources and criteria for endpoint identification with chemical counts and mapped key events 110
4.2 Comparison of risk quotient (RQ) data for arsenic and cadmium in the Indian rivers 168

Abstract

Human health is shaped by both genetics and lifetime environmental exposures. The exposome framework systematically captures such environmental exposures and links them to health outcomes. The chemical exposome is of special concern, as thousands of new chemicals are introduced into the global market annually, with many of them inevitably finding their way into the environment, yet their effects on human health remain poorly understood. Traditional toxicity testing relied heavily on slow and expensive animal studies, resulting in only a fraction of these chemicals being evaluated. In recent years, computational approaches have emerged as vital alternatives. They help organize toxicity data to prioritize chemicals for testing and provide mechanistic insights into their adverse effects. Therefore, in this thesis, we aim to link the chemical exposome to human health by systematically characterizing the chemical exposome and gaining mechanistic insights into chemical-induced adverse effects through diverse computational approaches.

As the first objective, we focus on characterizing the chemical exposome associated with specific human health outcomes. We investigate the chemical exposome of vitiligo by curating a dedicated knowledgebase of its environmental chemical triggers. We next explore the chemical space to assess regulatory coverage, and utilize skin-specific transcriptomic signatures to gain mechanistic insights into the otherwise less understood etiology of this disease. Building on this, we extend our focus to endocrine disrupting chemicals (EDCs), a broader class of environmental chemicals associated with a wide range of adverse health effects. We update the Database of Endocrine Disrupting Chemicals and their Toxicity profiles (DEDuCT) by integrating data from diverse toxicological resources, providing a single window access to all EDC-relevant information. We further map these chemicals to Adverse Outcome Pathways (AOPs), and organize all compiled data into a large-scale toxicology knowledge graph, termed DEDuCT-KG, which we utilize to gain mechanistic insights into EDC-associated obesity and neurodegenerative disorders.

As the second objective, we expand the scope of our computational investigation

beyond human health, employing diverse network toxicology frameworks to assess the impacts of inorganic arsenic and cadmium on human and ecosystem health from a One Health perspective. We construct Aggregate Exposure Pathway (AEP) networks for the first time in the Indian context, and integrate them with stressor-AOP networks to trace complete source-to-outcome pathways. We compute Species Sensitivity Distributions (SSDs) from aquatic toxicity data to estimate hazard concentrations, and integrate them into stressor-species networks to identify particularly vulnerable species and aid in ecological risk assessment. Finally, we calculate Risk Quotients (RQs) to gauge the ecological risks associated with arsenic and cadmium in Indian rivers.

Overall, the resources developed and frameworks utilized in this thesis provide results and insights that are easily accessible to the scientific community, laying the groundwork for future efforts towards a more complete characterization of the chemical exposome. The structured network toxicology investigation provides a basis for exploring the links between environmental chemical stressors and human and ecosystem health from a One Health perspective.

Chapter 1

Introduction

Genetics loads the gun, but the environment pulls the trigger.

- Judith Stern

1.1 Background and motivation

Human health is a multifactorial state, and its well-being is influenced by more than just the diseases it faces. The World Health Organization (WHO) has defined human health as 'a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity', emphasizing the role of various external factors in influencing health. Different individuals are exposed to different sets of external factors with varying quantities and durations of exposure. In this context, exposome was introduced to enable prioritization and identification of such exposures towards understanding health [1]. Exposome is a dynamic, rapidly evolving concept that refers to the life-course of an individual's environmental exposures and how these exposures relate to their health [1–3]. Thus, it is important to elucidate the link between the exposome and human health.

The exposome can broadly be categorized into the internal exposome, specific external exposome, and general external exposome [1]. The internal exposome encompasses individual's internal biological environment, including their physiology and age. The

specific external exposome encompasses the diet, occupational exposures, physical and biological exposures, while the general external exposome encompasses the broader societal and environmental factors, such as the socioeconomic status and education levels among others [1,2,4,5]. These domains are not independent of each other, but rather overlap and interact in complex ways [1,2,4,5]. Beyond these broad domains, other external factors such as climate change, urbanization, and loss of biodiversity act indirectly by modifying the factors to which we are exposed, thereby altering our exposome [4]. Moreover, there is growing evidence that environmental exposures can influence the epigenome, altering gene activity and expression, with potential transgenerational effects [4, 6]. In this fast-changing world, with a rising burden of lifestyle- and environment-linked diseases, it becomes increasingly important to characterize the complete exposome to better understand its role in human health and design preventive strategies. Recognizing this need, global efforts such as the Human Exposome Project have been conceptualized, in a similar spirit to the Human Genome Project, to systematically chart the totality of environmental exposures and their biological consequences [7, 8].

Among the different components of specific external exposome, the chemical exposome garners special interest among researchers and regulatory bodies alike. Importantly, it is one of the fastest expanding components of the exposome, with around 2000 new chemicals being introduced in the consumer space each year [9]. A systematic review of global chemical inventories has revealed that at least 350000 chemicals have been registered in the market [10], while nearly 119 million chemical structures have already been documented in PubChem [11], one of the largest openly accessible resources on chemical structures. A significant fraction of these chemicals inevitably enter the environment through various anthropogenic activities, yet their effects on human health remain poorly understood [10]. This can be largely attributed to the fact that traditional toxicity testing is heavily reliant on animal studies, which are expensive, time-consuming, and low-throughput, resulting in only a fraction of these chemicals being experimentally evaluated for their biological activity [12]. Moreover, such studies relied on understanding apical toxicities derived from

extremely high doses of chemicals, which were not representative of real-life exposures, and thus not very informative [13].

Recognizing such limitations, the United States National Research Council (US NRC), through its landmark vision report titled '*Toxicity Testing in the 21st Century: A Vision and a Strategy*', motivated a shift away from conventional animal-based testing strategies towards computational and high-throughput *in vitro* approaches [13]. This report also emphasized the importance of understanding toxicity mechanisms over apical endpoints, which inspired the development of new approach methodologies (NAMs) that integrate different *in vitro* and *in silico* methods to enhance the efficiency and relevance of toxicity assessment [14]. Building on this, a follow-up report in 2017 further emphasized the importance of integrating data within *in silico* prediction tools, and highlighted that toxicity data scattered across literature leads to inefficiencies and redundancy [15].

This report further suggested that establishing open-access databases using standardized formats helps enhance data sharing and model development, underscoring the importance of curated knowledgebases in characterizing the chemical exposome and linking it to human health. In this direction, the Findable, Accessible, Interoperable, and Reusable (FAIR) guiding principles have emerged as an important framework to ensure that such resources are structured in a manner that enables their seamless integration and reuse by both humans and machines [16]. Specifically, the Findable principle ensures that data and metadata have unique identifiers to facilitate their discovery and citation. The Accessible principle ensures that data and metadata are retrievable through their identifiers, and that metadata remains available even when the data itself cannot be shared. The Interoperable principle requires that data and metadata use formal and broadly applicable language to enable knowledge sharing across different resources. Finally, the Reusable principle emphasizes that the extent of standardization and its endorsement by other resources should be clearly stated, along with the data usage license, to support reuse across related domains [16]. Together, these principles provide a standardized framework for data sharing and resource development, and have been increasingly adopted by toxicological databases

and knowledgebases to enhance their utility and accessibility to the research community.

The United States Environmental Protection Agency (USEPA) has been at the forefront of this global effort of characterizing chemical toxicants, with efforts such as the Tox21 program to test diverse environmental chemicals against different human receptors [17]. This created the first large-scale effort of high throughput toxicity testing, which later got incorporated under a broader program called ToxCast that integrated data from diverse sources [18]. Over the years, ToxCast has analyzed assay outputs and provided data through an online publicly accessible resource, ToxCast invitrodb [19], with the latest version comprising 9800 chemicals tested across different biological targets, providing a high confidence dataset to support further *in silico* developments [20]. The USEPA also hosts the CompTox Chemicals Dashboard [21], a web-based application that provides access to a range of chemical datasets and tools, including chemical properties, toxicity data, exposure information, and bioassay results, to support computational toxicology research and aid further characterization of chemicals [22, 23]. The Comparative Toxicogenomics Database (CTD) [24, 25] is the largest publicly accessible resource curated from published literature, systematically compiling chemical-gene, chemical-disease, chemical-phenotype, and gene-disease associations across multiple species, to offer insights into the biological processes underlying chemical-induced health effects. The Library of Integrated Network-based Cellular Signatures (LINCS) [26] is another important resource that cataloged cellular responses to chemical and genetic perturbations across different human cell types, providing experimentally derived chemical-induced transcriptomics signatures that can aid in understanding the biological effects of chemicals. While LINCS is no longer actively updated, its data remains freely accessible through iLINCS [27], an interactive platform for integrative analysis and visualization of LINCS data. Together, these resources provide a wealth of chemical toxicological information, but are broad in scope and may not specifically capture the unique mechanisms and endpoints pertaining to specific adverse effects. Therefore, there is a need for focused characterization of the chemical exposome specific to adverse effects, which, when integrated with omics studies, can help decipher

the underlying mechanisms involved in chemical-induced adverse effects that are otherwise not fully captured by these broad scope resources [28].

Among the NAMs, *in silico* techniques have emerged as powerful tools that can help organize the disparate chemical data across the literature and databases, aid in prioritizing chemicals for further testing, and provide valuable insights into underlying mechanisms, thereby refining testing strategies [29]. Cheminformatics-based approaches can further aid in characterizing the chemical space of the exposome, enabling analysis of chemical diversity, identification of structure-activity relationships, and prioritization of chemicals based on their structural and physicochemical properties [12]. In addition, chemical-induced omics signatures can be used to uncover the underlying mechanisms of action of chemicals, thereby providing insights into the biological pathways perturbed by chemical exposure [28]. While such targeted approaches can help uncover adverse effect-specific toxicological insights, understanding the broader adverse effects of chemicals across multiple biological systems requires a more integrative strategy.

Beyond human health, chemical exposures can also affect animal and ecosystem health, highlighting the need for a broader perspective that goes beyond human-centric assessments. The One Health concept recognizes that human, animal, and environmental health are deeply intertwined, and advocates for an integrated approach to address health challenges that span across these domains [30, 31]. In particular, chemical stressors have been identified as a key concern within the One Health framework, as their effects are not limited to humans alone but extend across multiple species and ecosystems [30]. Therefore, characterizing the chemical exposome and its impacts from a One Health perspective requires approaches that can integrate diverse chemical, biological, and ecological data in a unified framework, which in turn can provide a more comprehensive basis for risk assessment and support better regulatory and mitigation strategies.

Network toxicology is an emerging computational framework that integrates toxicology, network biology, computational science, and environmental sciences to study the impact of toxic substances on biological systems [32]. In the context of NAMs, it

provides a novel strategy for integrating diverse environmental and biological data into network-based frameworks [32, 33]. Importantly, network toxicology also facilitates the compilation and organization of chemical effects across multiple species, thereby supporting the One Health perspective by capturing the interconnectedness of human, animal, and environmental health [34]. In this thesis, we employ diverse network toxicology-based computational approaches to systematically characterize the chemical exposome and link it to human and ecosystem health outcomes.

Knowledge graphs are one such network-based computational framework that has emerged for organizing diverse information and effectively representing the complex interconnections among disparate data sources [35–37]. In knowledge graphs, nodes represent entities of interest, and the edges denote the semantic relations that characterize the connections between the entities [35–37]. Such structured representations have provided a basis for efficiently processing heterogeneous knowledge and have eventually supported the training and reasoning of modern artificial intelligence (AI) systems [35–37]. In particular, toxicology knowledge graphs have emerged as a tool to systematically integrate existing chemical-associated toxicity and biological information, enabling the discovery and expansion of knowledge on chemical-induced toxicities [38–41]. Such representations enable the exploration of toxicity mechanisms that are otherwise overlooked within individual data sources, and moreover, organizing toxicological data into a knowledge graph structure makes it readily usable for AI-based research, thereby expanding the scope of data-driven investigations into chemical-induced toxicities.

The Aggregate Exposure Pathway (AEP) is another conceptual network toxicology framework that integrates existing knowledge on the plausible transport or transformation processes, linking the point of introduction of a chemical or stressor into the environment, and its concentration at the biological site of action [42–45]. AEPs are considered Next Generation Risk Assessment (NGRA) tools as they can assist in identifying potential exposure routes, assessing human and ecological risks, and informing regulatory decisions for better risk management and mitigation strategies [14, 42, 43, 46]. The AEP is represented as

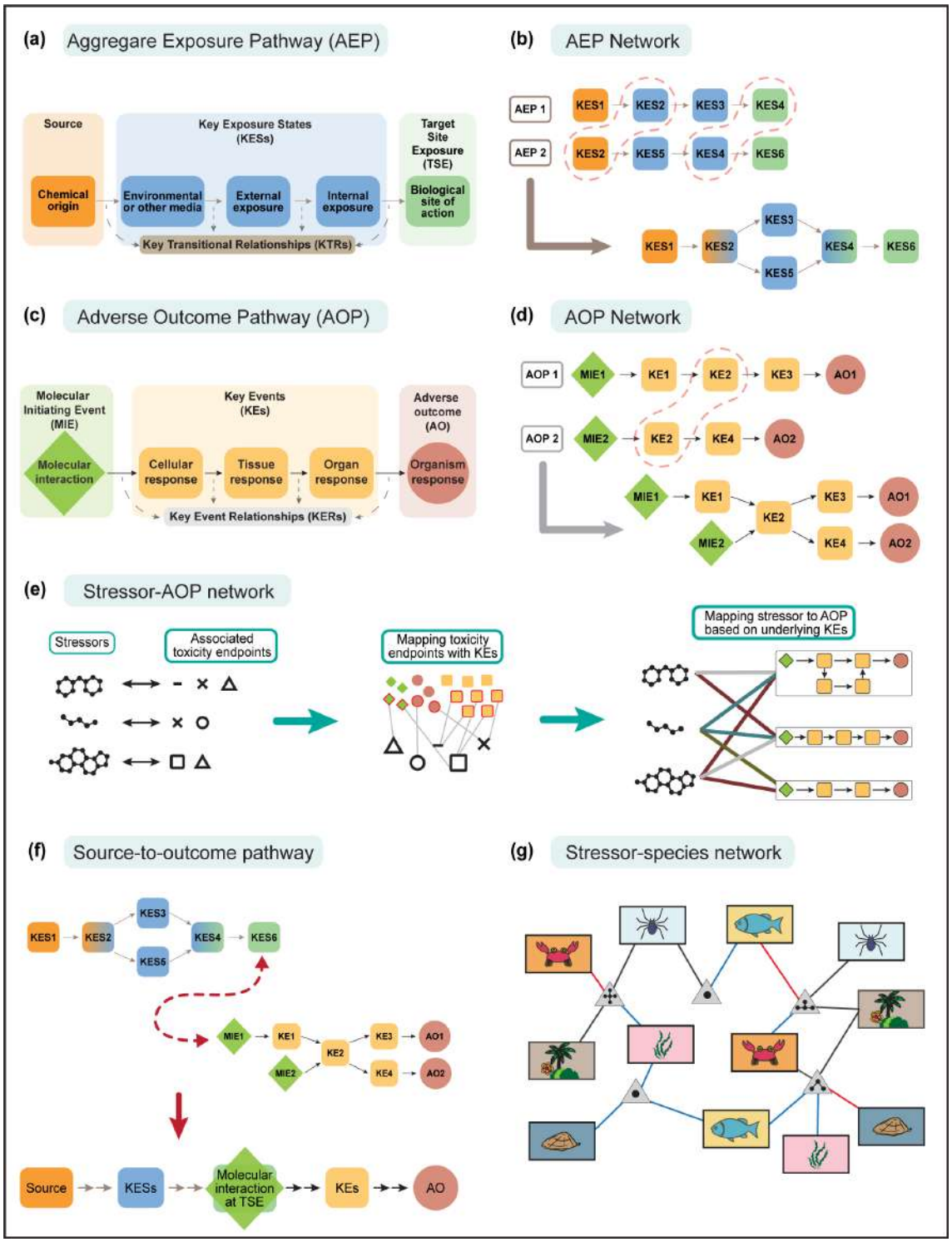

(a) Aggregare Exposure Pathway (AEP)
Source
Chemical origin
Key Exposure States (KESs)
Environmental or other media
External exposure
Internal exposure
Target Site Exposure (TSE)
Biological site of action
Key Transitional Relationships (KTRs)
(b) AEP Network
AEP 1
AEP 2
KES1
KES2
KES3
KES4
KES5
KES6
(c) Adverse Outcome Pathway (AOP)
Molecular Initiating Event (MIE)
Molecular interaction
Key Events (KEs)
Cellular response
Tissue response
Organ response
Adverse outcome (AO)
Organism response
Key Event Relationships (KERs)
(d) AOP Network
AOP 1
AOP 2
MIE1
MIE2
KE1
KE2
KE3
KE4
AO1
AO2
(e) Stressor-AOP network
Stressors
Associated toxicity endpoints
Mapping toxicity endpoints with KEs
Mapping stressor to AOP based on underlying KEs
(f) Source-to-outcome pathway
Source
KESs
Molecular interaction at TSE
KEs
AO
(g) Stressor-species network

Figure 1.1 ***(previous page)*:** Schematic representation of network toxicology frameworks utilized in this thesis. (a) Aggregate Exposure Pathway (AEP) organizing chemical exposure data into Key Exposure States (KESs), that are directionally linked through Key Transitional Relationships (KTRs). The origin of chemical is termed Source, and the biological site of action is termed Target Site Exposure (TSE). (b) AEP network constructed by assembling two different AEPs via shared KESs. (c) Adverse Outcome Pathway (AOP) organizing toxicity mechanism into Key Events (KEs), that are directionally linked through Key Event Relationships (KERs). The KEs include Molecular Initiating Event (MIE) and Adverse Outcome (AO) spanning different biological levels of organization. (d) AOP network constructed by assembling two different AOPs via shared KEs. (e) Stressor-AOP network constructed by mapping stressor to KEs, and subsequently assembling AOPs that comprise the mapped KEs. (f) Source-to-outcome pathways constructed based on the association between TSE of an AEP and MIE of an AOP. (g) Stressor-species network linking stressor to species.

a directed graph, where nodes correspond to the presence of a chemical in an environmental state, termed Key Exposure State (KES), and directional links represent the transport or transformation from one state to another, termed Key Transitional Relationship (KTR; Figure 1.1a) [42, 43]. The KESs are organized into different levels, including the source, environmental media, external exposure, internal exposure, and target site exposure (TSE), which corresponds to the site of molecular interaction or interaction with a biological entity, thereby clearly capturing the source-to-outcome pathways [42, 43]. An AEP network, constructed by joining different pathways through common KESs, provides a holistic view of exposures pertaining to a chemical and supports mechanistic understanding of its movement through the environment (Figure 1.1b) [44, 45].

The Adverse Outcome Pathway (AOP) framework is a network toxicology framework to understand mechanisms underlying stressor-induced toxicities [47–49]. AOP is a modular framework that describes stressor-induced toxicity as a linear network of causally linked biological events, known as Key Events (KEs; Figure 1.1c). This network originates at a molecular level, termed Molecular Initiating Event (MIE) and progresses to an Adverse Outcome (AO) of regulatory significance [47–49]. The causal connections between KEs are termed Key Event Relationships (KERs), and both KEs and KERs are supported by scientific evidence (Figure 1.1d) [47–49]. AOPs are therefore valuable tools in toxicity testing, as they support the development of targeted testing strategies, aid in chemical

prioritization, and provide a mechanistic basis for risk assessment and regulatory decision-making [47–49]. Importantly, the chemically agnostic nature of AOPs allows mechanistic knowledge from well-studied chemicals to be leveraged for the assessment of less studied ones, thereby enhancing the efficiency of chemical safety evaluations [48, 49].

The AOP-Wiki [50] is the largest open-source repository for AOPs developed globally, and is maintained by the Society for the Advancement of Adverse Outcome Pathways (SAAOP) in collaboration with Organisation for Economic Co-operation and Development (OECD), to facilitate global and collaborative AOP development. Within AOP-Wiki, information is organized across distinct pages for KEs, KERs, AOPs, and associated stressors, with each page capturing relevant biological and chemical information, including biological organization levels, domain of applicability, ontological terms, supporting evidence, and references [51]. The contributed data is reviewed and endorsed by the Working Party on Hazard Assessment (WPHA) and the Working Group of the National Coordinators of the Test Guidelines Programme (WNT) under the OECD AOP development program [51]. The entire AOP-Wiki dataset is freely accessible online, with quarterly updates available for download, making it a valuable resource for the research community to explore and expand knowledge on stressor-induced toxicity mechanisms.

The AOPs compiled within AOP-Wiki can further be connected through shared biological events to construct AOP networks, which provide a more comprehensive understanding of complex interactions and cumulative risks from various stressors, enhancing predictive toxicology, risk assessment, and regulatory decision-making [52–54]. In our earlier work, we demonstrated the utility of integrating mammalian-centric biological endpoint data within the AOP framework to explore toxicity mechanisms associated with cadmium [55] and plastic additives [56], and conducted targeted case studies to gain mechanistic insights into their associated adverse effects. In parallel, we extended this approach to incorporate ecotoxicologically-relevant data from the ECOTOX knowledgebase to investigate toxicity mechanisms underlying petroleum hydrocarbon-induced toxicity through the AOP framework [57]. Importantly, stressor-AOP networks were constructed by linking plastic

additives and petroleum hydrocarbons to relevant AOPs as bipartite networks (Figure 1.1e), providing a broader perspective on the diversity of biological processes impacted by these chemical stressors [56,57]. Together, these efforts demonstrate that combining mammalian-centric and ecotoxicologically-relevant data sources within the AOP framework can provide a more comprehensive and One Health-focused understanding of chemical-induced toxicity mechanisms across species.

Furthermore, an AEP-AOP construction offers a unified source-to-outcome framework that mechanistically links contaminant release, environmental transport, exposure, and toxicological effects [43,58,59]. By bridging exposure science and toxicology, it enables the integration of external exposures with internal target site concentrations and subsequent biological responses [43,58,59]. Importantly, the TSE defined in the AEP corresponds directly to the MIE in the AOP, thereby providing the point of connection between the two frameworks (Figure 1.1f) [42,43]. In this way, the AEP and AOP complement each other, with the AEP tracing how contaminants reach the site of action and the AOP describing the subsequent biological cascade leading to adverse outcomes [42,43]. This combined framework organizes complex mechanistic data, highlights knowledge gaps for risk assessment, and facilitates broader evaluations within the One Health perspective [43,58,59].

The stressor-species network is another network toxicology framework that examines the complex interactions between multiple environmental stressors and various species within ecosystems (Figure 1.1g) [57,60–62]. Such networks help identify the impacts of chemical stressors on diverse species, elucidating potential risks to ecosystems and aiding in the prioritization of regulatory measures [57,60–62]. Stressor-species networks constructed using acute and chronic toxicity data can reveal the range of species affected by a given chemical stressor, while those constructed using bioconcentration factor (BCF) data can highlight the potential for these chemicals to accumulate within different species, providing indirect evidence of their potential for biomagnification across the food chain [57]. Such networks can be constructed using data from the ECOTOX knowledgebase [63], which is

one of the largest openly accessible repositories of ecotoxicological data, compiled from published literature across diverse species and chemical stressors. The aquatic toxicity data compiled within such networks can be further utilized to support environmental risk assessment strategies, enabling the estimation of potential hazardous concentrations and the identification of particularly vulnerable species and ecosystems.

Species sensitivity distribution (SSD) is a widely used statistical method in ecological risk assessment that models the variation in sensitivity of different species to a given chemical stressor, and derives a hazard concentration affecting a defined percentage of species [64–66]. Most commonly, the HC05, which represents the concentration at which 5% of the tested species are affected, is derived from SSDs and is widely used to inform regulatory guidelines and environmental quality standards, particularly for the derivation of water quality guidelines [67]. However, a key limitation of the traditional SSD approach is that it requires sufficient toxicity data across multiple species for each chemical, which is often not available for many environmental contaminants. To address this, the toxicity-normalized SSD (SSDn) approach has been proposed as a refined method that normalizes toxicity data of multiple chemicals based on a sensitive reference species, thereby allowing the pooling of data from toxicologically similar chemicals and enabling the derivation of hazard concentration estimates even for chemicals with limited species-specific toxicity data [68, 69]. Together, SSDs and SSDns provide a robust framework for estimating the potential hazardous concentrations of chemical stressors in aquatic environments. The derived HC05 values can further be utilized to compute the predicted no-effect concentration (PNEC) by dividing the HC05 by an appropriate assessment factor [70]. Risk quotients (RQs) can then be derived by expressing the measured environmental concentration (MEC) of a chemical as a fraction of its computed PNEC, thereby providing a quantitative measure of its associated ecological risk [71, 72]. RQ values greater than one indicate a potential ecological risk, and such assessments can provide valuable inputs to guide policy decisions and support the development of better regulatory guidelines for the protection of ecosystem health.

In this thesis, we aimed to utilize diverse computational approaches to link the chemical exposome to human and ecosystem health. Towards this, the first objective of this thesis is to utilize diverse computational approaches to gain mechanistic insights into chemically induced adverse effects. To achieve this, in Chapter 2, we initially focus on vitiligo, a complex autoimmune disease of the skin, where chemical exposures have been observed as potential environmental triggers, yet remain less explored. We systematically curate a knowledgebase of chemical triggers of vitiligo, and explore their chemical space to assess existing regulatory coverage. We further utilize skin tissue-specific transcriptomics signatures to gain mechanistic insights into the otherwise less understood etiology of this disease. In Chapter 3, we extend this focus to a broader class of chemicals associated with a wide range of adverse health effects, such as endocrine disrupting chemicals (EDCs). We update DEDuCT by integrating data from diverse toxicological resources, map EDCs to AOPs to explore their associated toxicity mechanisms, and organize all compiled data into a large-scale toxicology knowledge graph. We then demonstrate the utility of this knowledge graph by gaining mechanistic insights into EDC-associated obesity and neurodegenerative disorders.

The second objective of this thesis is to systematically characterize the environmental distribution of chemical contaminants, extend the network toxicology framework to trace their complete source-to-outcome pathways, and assess the broader impacts of chemical exposure on human and ecosystem health through a One Health perspective. To achieve this, in Chapter 4, we focus on two prevalent heavy metal pollutants in India, namely, inorganic arsenic and cadmium. We construct AEP networks to trace their movement from environmental sources to biological target sites, and integrate them with stressor-AOP networks to obtain complete source-to-outcome pathways. We further construct stressor-species networks and compute SSDs to assess the ecological effects of these chemicals across diverse species and estimate hazard concentrations. Finally, we calculate RQs to gauge the ecological risks associated with inorganic arsenic and cadmium in Indian rivers.

1.2 Thesis organization

The remaining chapters of this thesis are organized as follows:

Chapter 2 presents a detailed workflow designed to characterize the chemical exposome of vitiligo. We first systematically screen published literature to curate chemical triggers of vitiligo, along with their sources of exposure and associated published evidence. We then cross-reference these chemicals with the USEPA's Chemical and Products Database (CPDat) to explore their presence in daily-use consumer products. Next, we check for the presence of these chemicals in skin-specific regulation lists across various jurisdictions, and compare them against other exposome databases, to identify the uniqueness of the chemicals in this dataset. We further conduct a systematic cheminformatics-based analysis to explore the chemical space of these triggers, and compute their scaffolds to identify regulatory gaps in skin-specific regulations. Finally, we utilize iLINCS to obtain skin tissue-specific transcriptomics signatures for the vitiligo-triggering chemicals, and explore the commonly perturbed biological pathways to gain mechanistic insights into the etiology of the disease. To enable further research, we compile all curated data into an online knowledgebase, the Vitiligo-linked Chemical Exposome Knowledgebase (ViCEKb), accessible at https://cb.imsc.res.in/vicekb/. **The work reported in this chapter is contained in the published manuscript [73].**

Chapter 3 presents an updated FAIR-compliant knowledgebase DEDuCTv3.0, and the first large-scale toxicology knowledge graph associated with EDCs, termed DEDuCT-KG. We first follow an established four-stage manual screening workflow to curate new EDCs and their associated endocrine-mediated endpoints from more than 14000 newly published research articles, and merge them with the previously curated data to obtain a comprehensive and updated list of EDCs and their associated endocrine-mediated effects. Next, we curate chemical-gene data from ToxCast invitrodb v4.3 and CTD, and chemical-phenotype and chemical-disease data from CTD, and integrate it with the chemical-specific data within the webserver to provide a single point of access to all EDC-related information.

We also integrate exposure data from CPDat and explore potential exposure sources for these chemicals in daily-use consumer products. Next, we curate the latest AOP data from AOP-Wiki, and systematically map the EDCs to these AOPs to construct a stressor-AOP network for EDCs, and utilize it to explore diverse toxicity mechanisms. Importantly, in addition to the EDC-specific endocrine-mediated endpoints, we also integrate EDC-specific human- and rodent-relevant toxicity data curated from ToxCast invitrodb v4.3, CTD, and NeurotoxKb, and map them to AOPs from AOP-Wiki, thereby broadening the mechanistic coverage of the EDC-AOP network. Finally, we construct a large-scale toxicology knowledge graph, DEDuCT-KG, comprising EDCs, associated endocrine-mediated endpoints, genes, phenotypes, diseases, key events, and AOPs, and make it accessible through an interactive user interface. We further demonstrate the utility of the knowledge graph by identifying potential mechanisms of action underlying EDC-associated obesity and neurodegenerative disorders. The DEDuCT webserver is accessible at https://cb.imsc.res.in/deduct/. **The work reported in this chapter is contained in the published manuscript [74].**

Chapter 4 presents a comprehensive network toxicology-focused investigation on the impacts of inorganic arsenic and cadmium from an India-specific One Health perspective. We first curate inorganic arsenic and cadmium compounds and their toxicological endpoints from six exposome-relevant resources. Next, we construct AEP networks by systematically curating India-specific data on their presence across different environmental and biological states, capturing the complete source-to-biological site of action pathways for these heavy metals across different environmental media and exposure routes. We further curate complete and connected AOPs from AOP-Wiki, and systematically map the extracted toxicological endpoints to identify arsenic- and cadmium-relevant AOPs, and construct stressor-AOP networks for both arsenic and cadmium. We further explore the taxonomic applicability of the highly relevant AOPs based on compiled evidence. Subsequently, we merge the constructed AEP and AOP networks to obtain complete source-to-outcome pathways for inorganic arsenic and cadmium, and conduct case studies

supported by published literature to explore key exposure and toxicity mechanisms, and identify knowledge gaps in terms of human exposures and tissue-level bioaccumulation. Next, we utilize acute and chronic toxicity and BCF data from the ECOTOX knowledgebase to construct stressor-species networks to assess the ecological effects of these chemicals across diverse species. We further construct SSDs based on aquatic toxicity data to estimate HCs for these chemicals in aquatic environments, and integrate the hazard concentration data into the stressor-species network, enabling the identification of highly sensitive species and contributing to ecological risk assessment strategies. Finally, we compute RQs associated with arsenic and cadmium in Indian rivers to assess their ecological risks. We have made the constructed networks and related data openly accessible at https://cb.imsc.res.in/heavymetaltox/. **The work reported in this chapter is contained in the published manuscript [75].**

Chapter 5 concludes with a brief summary of the work presented across different chapters, and discusses the future scope of utilizing computational approaches to investigate environmental chemicals in the exposome, and link them to human and ecosystem health.

Chapter 2

ViCEKb: Vitiligo-linked Chemical Exposome Knowledgebase

Vitiligo (also known as leukoderma) is a chronic, often long-lasting autoimmune disease, where the melanocytes are pathologically killed, ultimately resulting in depigmented patches on the skin [76]. The characteristic vitiligo-linked skin depigmentation recurs after the cessation of conventional treatments, due to the presence of resident memory T cells in the lesion [77]. Epidemiological studies have shown that people diagnosed with vitiligo are also susceptible to other autoimmune disorders like type 1 diabetes, rheumatoid arthritis, pernicious anemia, and Addison's disease [78,79]. Apart from health effects, its visible and early onset in children and young adults negatively affects their psychosocial development, thereby leading to a decreased quality of life [80,81]. Therefore, understanding the etiology is important to identify novel treatments for vitiligo.

Vitiligo is a complex disease, wherein environmental factors trigger autoimmune pathways depending on the underlying genetic predispositions [82]. Multiple genome-wide association studies (GWAS) have extensively examined the genetic loci and variations associated with vitiligo [83, 84]. There have been simultaneous efforts to generate comprehensive gene set databases, such as the Gene and Autoimmune Disease Association Database (GAAD) [85], and VitiVar [86], that were compiled from the vast resource of

published literature. Meanwhile, the data on environmental triggers of vitiligo remains sparse and is poorly understood. Specifically, environmental chemical exposures through various sources, including occupational and daily use products, have been identified as potential risk factors for developing the disease [87–90], yet the specific details regarding these associations remain buried and fragmented across published literature.

In this chapter, we aim to bridge this existing knowledge gap by systematically curating information on chemical triggers of vitiligo from published literature. Subsequently, we compare this chemical list against existing chemical regulations and consumer products to identify regulatory oversights. Thereafter, we employ different cheminformatics-based approaches to explore the diversity in the chemical space of ViCEKb. Finally, we utilize skin-specific transcriptomics signatures of these chemicals to gain mechanistic insights underlying disease pathogenesis. **The work reported in this chapter is contained in the published manuscript [73].**

2.1 Methods

2.1.1 Workflow for the compilation of chemical triggers of vitiligo with supporting evidence in published literature

In this chapter, we first aimed to compile chemical triggers of vitiligo in the human exposome that have reported evidence in the published literature. To achieve this, we designed a workflow based on mining published literature to identify reported evidence on chemical exposure leading to vitiligo pathogenesis, and thereafter, compiled a curated list of chemicals that can trigger vitiligo in humans (Figure 2.1).

First, we mined PubMed [91] using the following query: '(leukoderma OR leucoderma OR vitiligo OR hypopigmentation OR depigmentation) AND (chemical* OR exposure* OR expose* OR exposing) AND (human[ORGN] OR "Homo sapiens"[ORGN])'. This PubMed search, last performed on 10 August 2022, retrieved 1756 articles likely to contain

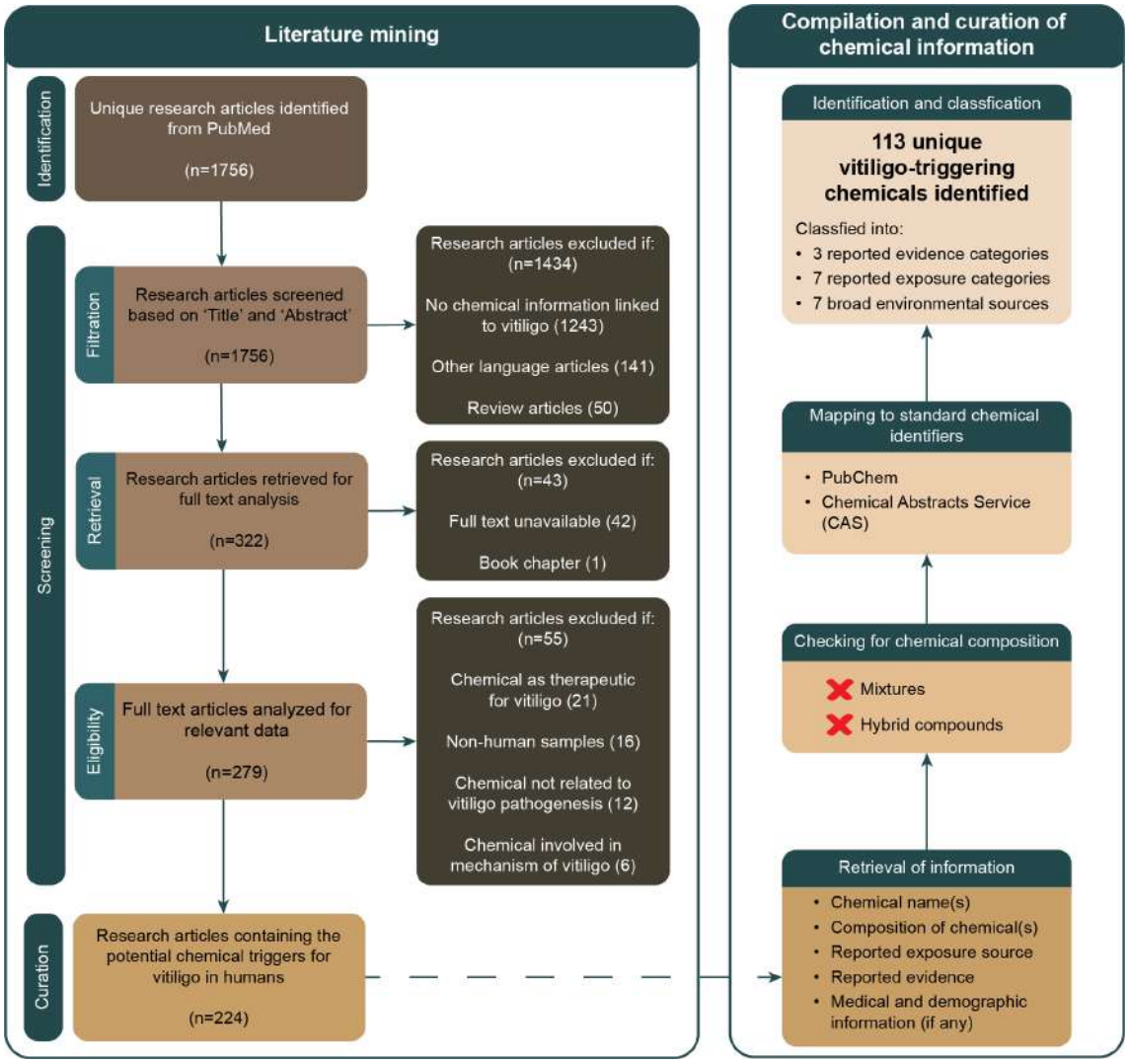


Figure 2.1: The workflow to identify published literature containing chemical triggers of vitiligo is presented as a PRISMA statement. From the shortlisted articles, various chemical information was retrieved and standardized to curate a list of 113 unique chemical triggers of vitiligo.

information on vitiligo-triggering chemicals in humans. Next, we manually screened the 1756 articles based on their titles and abstracts and excluded the article if: (i) it did not contain any chemical information about vitiligo pathogenesis; (ii) it was written in a language other than English; (iii) it was a review article (Figure 2.1). Consequently, we identified 322 articles at the end of the screening process, and thereafter, proceeded to collect their full texts. Here, we excluded an article if it was a book chapter or if its full text was unavailable. Accordingly, we retrieved 279 full text articles, following which we manually screened through every article and excluded them if: (i) it contained

evidence from non-human organisms; (ii) it reported chemicals involved in the downstream mechanism of vitiligo; (iii) it reported chemical therapeutics for vitiligo; (iv) it reported chemicals not involved in pathogenesis of vitiligo (Figure 2.1).

Finally, we curated a list of 224 published articles that contained evidence on chemical exposure leading to the pathogenesis of vitiligo in humans. Figure 2.1 details the workflow to compile and curate the literature evidence for chemical triggers of vitiligo in humans and their corresponding information as a PRISMA statement [92]. In the next section, we describe the workflow for extracting chemical triggers of vitiligo from 224 published articles.

2.1.2 Curation and compilation of data from shortlisted literature

We manually screened the curated list of 224 published articles to extract information about chemical triggers of vitiligo. From each article, we retrieved the reported chemical name(s), their chemical composition, corresponding chemical exposure source, and the type of evidence asserting the chemical to be a trigger for vitiligo. We excluded chemical triggers reported as multicomponent chemical mixtures because it is difficult to attribute the cause of vitiligo to any of its individual chemical components. Thereafter, we mapped the retrieved chemical names to standard chemical information from PubChem [11] and Chemical Abstracts Service (CAS) [93], and subsequently, compiled a list of 113 unique chemicals (Supplementary Table S2.1). For each of the 113 identified chemicals, we also obtained their identifiers from the Distributed Structure-Searchable Toxicity (DSSTox) database [94, 95] to aid in toxicity predictions (Supplementary Table S2.1). Furthermore, we obtained chemical categorizations such as Kingdom, Superclass, and Class for the identified chemicals using ClassyFire [96, 97] (Supplementary Table S2.1).

2.1.3 Classification of vitiligo-triggering chemicals

Based on the type of evidence from published literature

Clinicians and researchers have employed different techniques to associate potential chemical triggers in patients diagnosed with vitiligo [98–100]. Based on the supporting evidence used to establish certain chemicals as triggers for vitiligo, we broadly classified the identified chemicals into three distinct categories. The first category, with the highest level of evidence, is 'Patch test done', which consists of chemicals that have been tested in small doses on human patient skin and depigmentation was clinically verified. The second category is 'No patch test done, Observation' and has a lower level of evidence compared to 'Patch test done'. This category consists of chemicals inferred from the history of chemical exposure among patients clinically diagnosed with vitiligo. The third category, with the lowest level of evidence, is '*in vitro* studies', which consists of evidence from enzymatic assays and cell line studies.

Based on the type of chemical exposure

In addition to the type of evidence, we have compiled the sources of chemical exposures reported in published literature. We observed that there were several types of reported sources [101–103], which we organized into the following seven broad categories:

i. Cohort study: population-wide studies comprising various exposure sources
ii. Consumable: chemical exposure through consumption
iii. Cosmetics: chemical exposure through skin care and skin modifying products
iv. Daily-use product: chemical exposure through products used on a daily basis
v. Laboratory: *in vitro* chemical studies
vi. Medical: chemical exposure through medical treatments
vii. Occupational: chemical exposure in the workplace

Here, we note that there may be an overlap of chemicals across different categories.

Based on environmental source

PubChem catalogs extensive chemical information and annotations, including information on environmental sources from over 900 data sources [104]. We used this environmental annotation data within PubChem, and classified the chemicals into seven broad categories, namely, 'Agriculture and farming', 'Consumer products', 'Industry', 'Intermediates', 'Medicine and healthcare', 'Natural Sources', and 'Pollutant'. We further divided these broad categories into 42 sub-categories [105]. We annotated the environmental source of a chemical as 'Unclassified' if the corresponding annotation was not available in PubChem. Here, we note that a single chemical can belong to multiple categories within this classification system.

2.1.4 Physicochemical properties, molecular descriptors and predicted ADMET properties of chemicals

For the identified 113 chemicals, we obtained the respective two-dimensional (2D) and three-dimensional (3D) molecular structures from PubChem. If the 3D structures were not available in PubChem, we employed Open Babel [106] to generate them from the obtained 2D structures. We employed RDKit [107] to compute the basic physicochemical properties and the molecular scaffolds of any chemical based on the Bemis-Murcko definition [108]. Additionally, we also computed the 1D, 2D and 3D descriptors using PaDEL [109, 110]. We have made all this information readily available on our webserver to further aid the development of chemical toxicity predictors by exploring structure-activity relationships.

Absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties can aid in assessing chemical toxicity [111]. We employed several computational tools, including admetSAR 2.0 [112], SwissADME [113], pkCSM [114], and vNN server [115] to predict the ADMET properties for the 113 identified chemicals. The absorption properties of chemicals have been estimated based on Caco-2 permeability and skin permeation (log Kp). The potential distribution of the chemical upon absorption has been estimated based

on the fraction unbound in human and subcellular localizations. The toxic properties of the chemicals have been estimated based on their cytotoxicity, maximum recommended tolerated dose (MRTD), micronucleus, and skin sensitization. For each chemical, we have made these predicted properties readily available on our webserver.

2.1.5 Comparison of vitiligo-triggering chemicals with consumer product database and chemical regulations

Chemical and Products Database (CPDat) [116, 117] contains qualitative and quantitative information on the chemical composition of consumer products reported in the public domain. Additionally, it provides information on: (i) Organisation for Economic Co-operation and Development (OECD) [118] categorization of functional uses of chemicals; (ii) product compositions and their uses; (iii) regulation in health hazard evaluations. Notably, the product use categorization (PUC) provided by CPDat has enabled harmonization of product categories and aided in the assessment of consumer exposure [119]. Therefore, we leveraged CPDat to identify the presence of the curated chemical triggers of vitiligo in consumer products.

The CompTox Chemicals Dashboard [21–23] is one of the largest public repositories that aggregates various chemical lists related to toxic effects on humans. Here, we used the skin-specific chemical lists to check whether the identified chemicals were evaluated for their skin sensitizing effects. Additionally, we checked whether the identified chemicals were documented for other adverse effects using databases such as DEDuCT [105, 120, 121], FCCP [122, 123], ExHuMId [124, 125], and NeurotoxKb [126, 127].

2.1.6 Identification of target genes of vitiligo-triggering chemicals based on ToxCast skin sensitization assays

The high-throughput screening assays undertaken as part of the ToxCast project have helped identify perturbations in biological processes induced by chemical exposure [128].

Therefore, we leveraged data from ToxCast version 3.5 [129] to identify the potential gene targets of the identified chemicals in human skin samples as follows. First, we checked the ToxCast assays for skin tissue and identified 58 relevant assay endpoints (belonging to assay source id 4) from the assay summary file. Then, we filtered the chemical data from level 5 and 6 processing file corresponding to assay source id 4, to consider the biologically active (hit_c = 1) representative chemicals (gsid_rep = 1), and identified the necessary chemicals based on their CAS identifiers. Subsequently, we retrieved the target gene, target bioprocess and assay endpoint information for the identified chemicals.

2.1.7 Exploration of the space of vitiligo-triggering chemicals

Chemical similarity network

We constructed a chemical similarity network (CSN) to assess the similarity of the identified vitiligo-triggering chemicals. First, we used RDKit to obtain the ECFP4 fingerprints of the identified chemicals. Second, we calculated the Tanimoto coefficients [130] between all pairs of chemical fingerprints to check the similarity between chemicals [105, 131]. Third, we constructed a network in which nodes represent chemicals and edges connect all pairs of nodes with weights given by the computed Tanimoto coefficient between the corresponding pair of chemicals. Next, we filtered out edges with Tanimoto coefficient $<$ 0.5 and visualized the resulting high similarity backbone of the CSN using Cytoscape [132]. Finally, we employed RDKit to compute the maximum common substructure for the connected components identified in the CSN.

Scaffold diversity

In addition to CSN, we assessed the similarity among identified chemicals based on their chemical scaffolds. We employed RDKit to compute the scaffolds of the identified vitiligo-triggering chemicals according to the Bemis-Murcko definition [108]. To visualize the computed scaffolds, we used the Scopy package [133] in Python to generate a scaffold cloud representation. Then, we used various skin-specific chemical regulatory lists to

visualize the scaffold diversity in the identified vitiligo-triggering chemicals. We used RDKit to compute the Bemis-Murcko scaffolds in each of the chemical lists, and analyzed the scaffold diversity via cyclic system retrieval (CSR) curves [134]. Finally, we visualized the scaffold overlaps as an UpSet plot [135] using R.

2.1.8 Analysis of transcriptomics signatures of vitiligo-triggering chemicals

The Library of Integrated Network-Based Cellular Signatures (LINCS) project has cataloged an extensive dataset of cellular signatures from different human cell types upon exposure to various chemical perturbations [26]. To analyze the transcriptomics signatures of chemicals triggering vitiligo, we leveraged skin tissue-specific L1000 chemical perturbation signatures from the iLINCS [27] web-based platform [136]. We observed that the maximum number of chemicals were tested on cell line A375 for 6h at a concentration of 10 μM, and therefore we shortlisted such chemical signatures to ensure maximum chemical coverage. If a chemical had multiple signatures, we selected the latest one containing significantly ($p < 0.05$) differentially expressed genes (DEGs). Next, we generated the significant DEGs set for each signature by setting the constraints of a minimum 2-fold change in expression ($|\text{logfoldchange}| \geq 2$) and significance value $p < 0.05$. Thereafter, we employed Enrichr [137, 138] to perform gene set enrichment for each of the generated gene sets, and retrieved the enriched pathways from 'Reactome 2022' [139] and enriched biological process ontology terms from 'GO Biological Process 2023' [140]. We additionally set a constraint of an adjusted p-value < 0.05 to identify significantly enriched terms from each of these sets.

2.1.9 Web interface and database management system

We have created an online resource, Vitiligo-linked Chemical Exposome Knowledgebase (ViCEKb), which compiles all the detailed information on the 113 unique chemical triggers

of vitiligo identified from 224 published articles. Importantly, ViCEKb provides easy access to various chemical information such as the type of supporting evidence, environmental sources, chemical structures, physicochemical properties, molecular descriptors, predicted ADMET properties, chemical gene interactions from ToxCast assays, and presence in various skin-specific regulatory lists and chemical exposome databases. ViCEKb is accessible at: https://cb.imsc.res.in/vicekb/.

The data in ViCEKb is stored in MariaDB [141] and retrieved using Structure Query Language (SQL). The web interface of ViCEKb was created using custom HTML, CSS, PHP [142], jQuery [143], and Bootstrap 5 [144]. The interactive charts were created using Google Charts [145]. ViCEKb is hosted on Apache [146] webserver running on Debian 9.4 Linux Operating System.

2.2 Results

2.2.1 ViCEKb - Vitiligo-linked Chemical Exposome Knowledgebase

Our primary goal in this chapter was to compile a resource on vitiligo-triggering chemicals within the human exposome. We followed a detailed workflow (Figure 2.1) to shortlist 224 published articles, from which we compiled and curated the different chemical names, reported evidence, and exposure types. Through this extensive process, we identified 113 unique chemicals (Supplementary Table S2.1) which we categorized under three broad evidence categories: 'Patch test done' (48 chemicals), 'No patch test done, Observation' (53 chemicals), and '*in vitro* studies' (12 chemicals) (Figure 2.2a). We further classified the reported sources of chemical exposure into seven categories, of which medical exposure was observed to be the most reported source (85 chemicals) (Figure 2.2b). Additionally, we categorized the chemicals into seven broad environmental sources (Figure 2.2c), which we further subdivided into 42 sub-categories (Supplementary Table S2.2). We observed that the majority of chemicals belonged to the environmental source 'Intermediates' (80 chemicals) (Figure 2.2c), which consists of industrial intermediates and human metabolites.

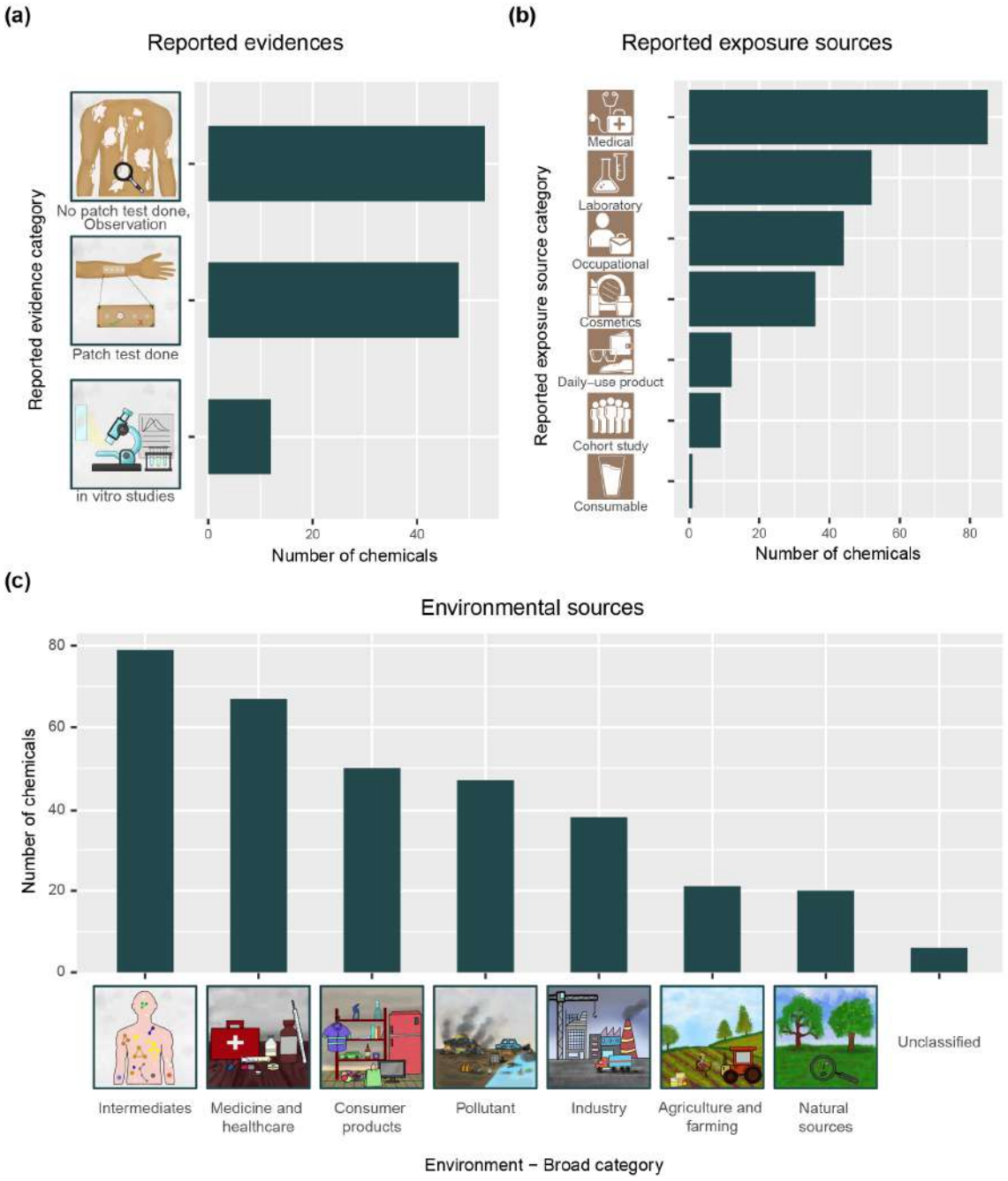


Figure 2.2: Overview of various chemical information retrieved from published literature. (a) The chemicals have been grouped into three categories based on their reported type of evidence. (b) The chemicals have been grouped under seven categories based on their reported source of exposure. (c) The chemicals have been grouped under seven broad environmental categories based on data available in PubChem. Note that six chemicals have no reported environmental sources and are therefore unclassified.

We compiled the identified 113 chemicals and associated information in a user friendly, FAIR (Findable, Accessible, Interoperable, Reusable) [16] and TRUST (Transparency, Responsibility, User focus, Sustainability and Technology) [147] compliant webserver named Vitiligo-linked Chemical Exposome Knowledgebase (ViCEKb), which is openly accessible at: https://cb.imsc.res.in/vicekb/.

The ViCEKb webserver includes various chemical information such as: (i) chemical identifiers; (ii) chemical structures and classification; (iii) chemical properties and descriptors; (iv) literature evidence with reported exposure sources; (v) environmental sources; (vi) chemical-gene interactions from ToxCast; (vii) presence of identified chemicals in various regulations including skin-specific regulations and chemical toxicity regulations. Additionally, it also provides easy access to chemical information compiled from other chemical toxicity and exposome databases like DSSTox [94, 95], DEDuCT 2.0 [120, 121], ExHuMId [124, 125], FCCP [122, 123], and NeurotoxKb [126, 127].

ViCEKb webserver allows easy navigation through the database with options like 'BROWSE', 'SEARCH', and 'DOWNLOAD' (Figure 2.3a). The 'BROWSE' option enables users to search for chemicals by their: (i) Evidence level (Figure 2.3b); (ii) Environmental source (Figure 2.3c); (iii) Reported exposure type (Figure 2.3d); (iv) ToxCast skin sensitization target process (Figure 2.3e). The 'SEARCH' option allows users to query the chemicals in this database with their names or identifiers (Figure 2.4a), and returns their structure, corresponding evidence from published literature, and a link to the chemical information page (Figure 2.4b). Additionally, the 'ADVANCED SEARCH' option provides users with a cheminformatics platform to search chemicals based on physicochemical feature similarity (Figure 2.4c) or chemical similarity (Figure 2.4d). The datasets in ViCEKb can be downloaded as flat files using the 'DOWNLOAD' option (Figure 2.4e). The webserver also has a 'HELP' option that provides a tutorial for using the database.

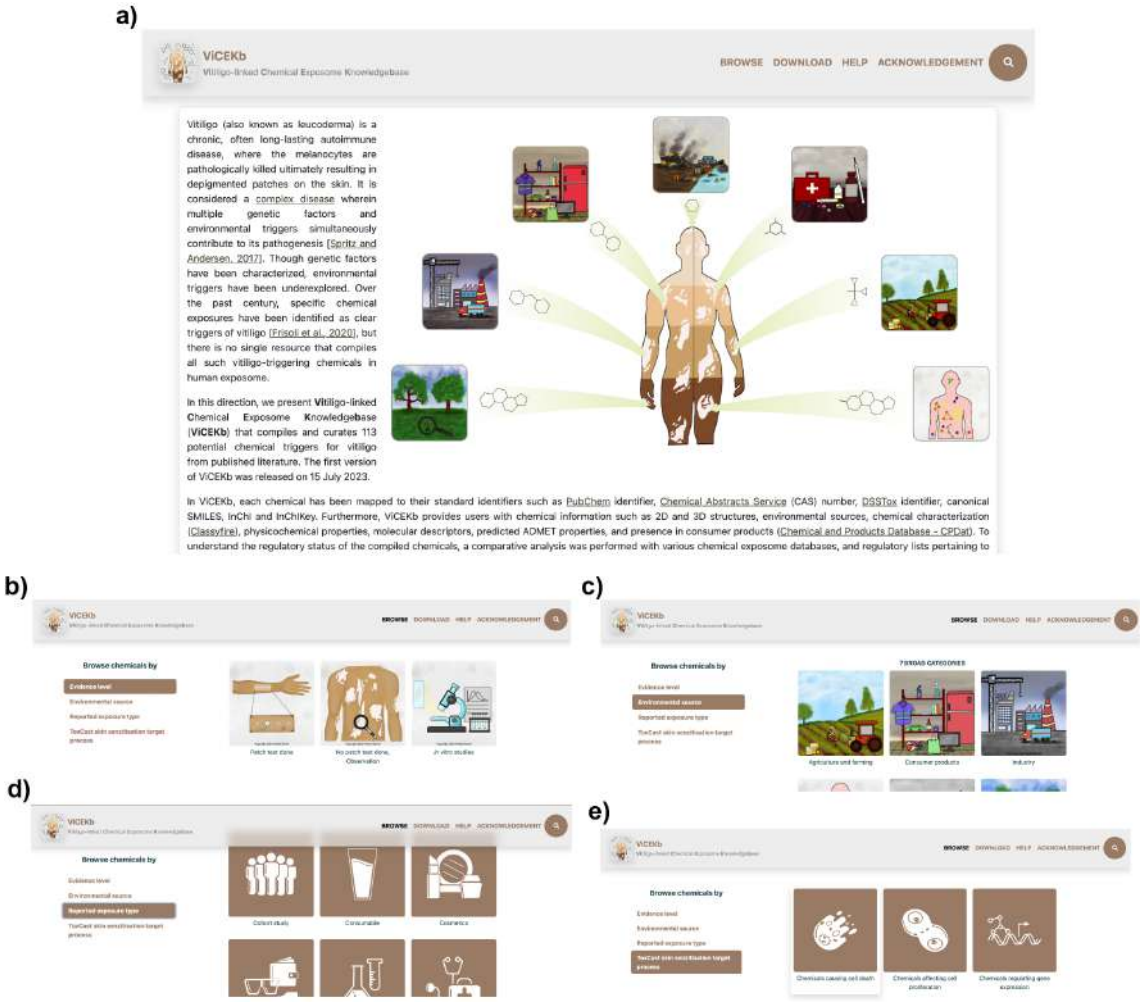
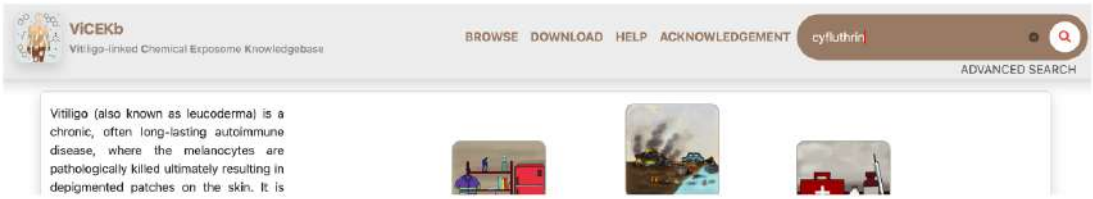


Figure 2.3: ViCEKb webserver images. (a) The home page of the ViCEKb webserver has an easy navigation bar. (b) Browse chemicals by evidence level. (c) Browse chemicals by environmental category. (d) Browse chemicals by reported exposure sources. (e) Browse for chemicals based on the ToxCast skin sensitization target.

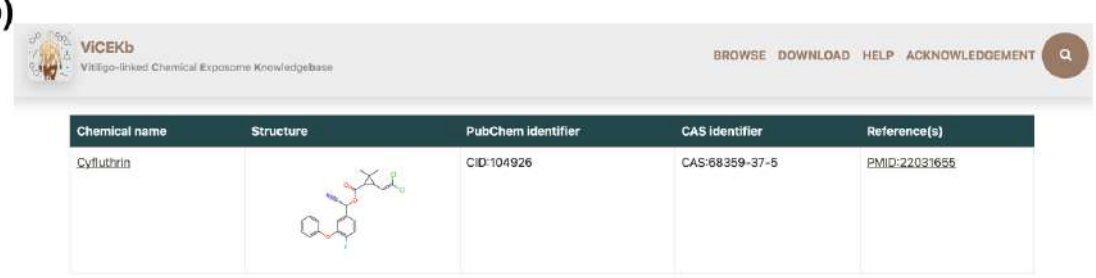
2.2.2 ViCEKb chemicals are present in consumer products

The pathogenesis of vitiligo has been observed to be linked with exposure to regular domestic products in developing countries [98]. In general, humans depend on a wide variety of consumer products, most of which come in direct contact with the skin. Therefore, we explored the presence and functions of chemicals from ViCEKb in consumer products by leveraging CPDat [116]. We identified 91 of 113 chemicals with information on their functional uses, product compositions, and their uses, and presence in occupational health hazard reports, all of which we have made available in our webserver under the

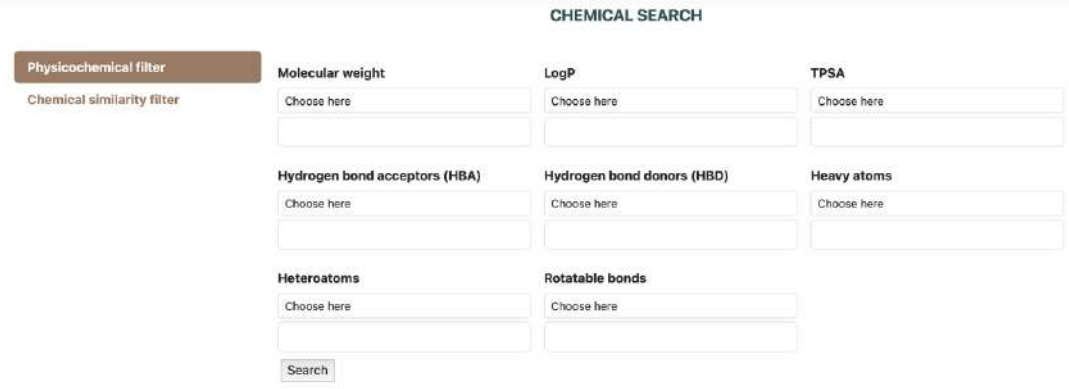
a)



b)



c)



d)



e)

BATCH DOWNLOAD

S. No.	File description	File size	Download	Release date
1	Names and identifiers of chemicals List of 113 chemicals with their PubChem identifier, CAS Number, DTSXID, Name, IUPAC Name, SMILES, InChI and InChIKey	31K		10/10/2023
2	Classification and categorization of chemicals List of 113 chemicals with their chemical Kingdom, cheimcal SuperClas, chemical Class, Environmental source - broad category and Environmental source - sub-category	23K		10/10/2023
3	Literature evidence, reported exposure, exposure category, evidence levels List of literature evidences for 113 chemicals along with reported exposures, standardized exposure category and standardized evidence levels	24K		10/10/2023
4	Physicochemical properties List of 113 chemicals and their Physicochemical properties, namely, Molecular weight, Log P, TPSA, Hydrogen bond acceptors, Hydrogen bond donors, Carbon atoms, Heavy atoms, Heteroatoms, Chiral centers and Rotatable bonds	6.8K		10/10/2023
5	Predicted ADMET properties List of 113 chemicals with their predicted ADMET (Absorption, Distribution, Metabolism, Excretion and Toxicity) properties	24K		10/10/2023
6	ToxCast skin assays List of chemicals with reported ToxCast skin assays, and the corresponding end point values	46K		10/10/2023

Figure 2.4 ***(previous page)*****:** ViCEKb webserver images of the 'SEARCH' and 'DOWNLOAD' sections. (a) The chemicals can be searched by their chemical identifiers in the top-right corner of the page. (b) The search results are displayed on a tabulated page that contains information pertaining to the search term. (c) ADVANCED SEARCH option based on physicochemical properties. (d) ADVANCED SEARCH option based on chemical similarity. (e) The 'DOWNLOAD' section contains the batch download files compiled on the webserver.

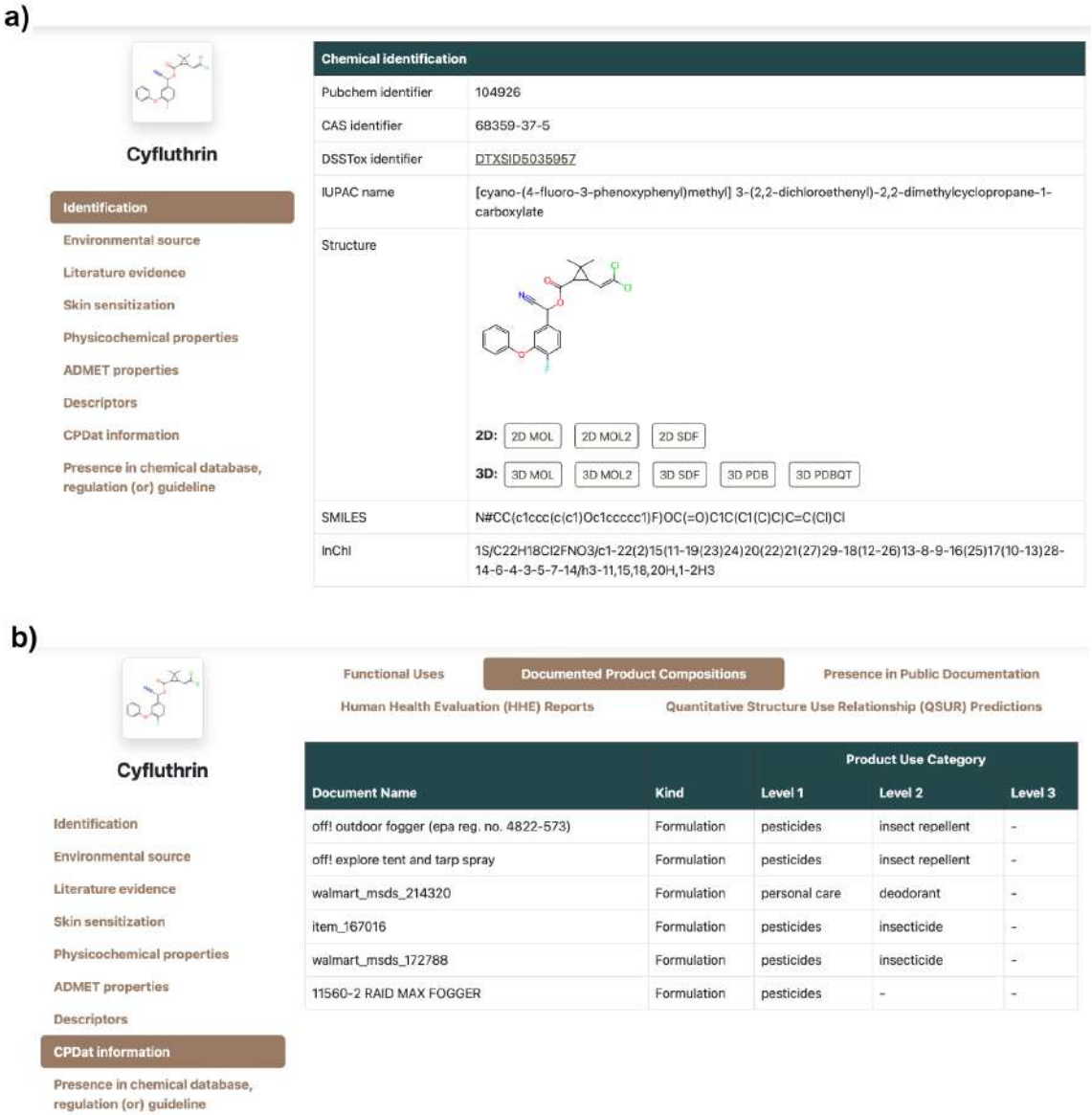


Figure 2.5: ViCEKb webserver images of CHEMICAL INFORMATION page. (a) Chemical identification page containing various chemical information and structure download files. (b) The CPDat information section contains the various CPDat information divided into different sections.

'CPDat information' section in the chemical information page (Figure 2.5). Among the 91 chemicals, 35 of them were found in various products, including personal care and

home hygiene (Figure 2.6a; Supplementary Table S2.3). Moreover, 24 of the 91 chemicals had standardized functional uses wherein fragrance was the most reported (Figure 2.6b; Supplementary Table S2.4). Additionally, five chemicals were documented in health hazard evaluation reports from different industries (Supplementary Table S2.5). Furthermore, we observed that 50 of 113 chemicals are present in a variety of consumer products, based on the reported environmental sources in PubChem [11] (Supplementary Tables S2.1-S2.2).

For instance, chemicals found in consumer products such as isopropyl alcohol (CID: 3776), hydroquinone (CID:785) and methyl methacrylate (CID:6658), have the highest evidence of 'Patch test done' in ViCEKb (Supplementary Table S2.1), and have also been documented in several health hazard evaluation reports (Supplementary Table S2.5). Moreover, methyl methacrylate (CID:6658) is classified in DEDuCT as a Category III endocrine disrupting chemical (EDC) and as a neurotoxicant in NeurotoxKb. Another chemical, 4-tert-butylphenol (CID:7393), which is found in household maintenance products and children's toys (according to FCCP), has the highest evidence of 'Patch test done' in ViCEKb (Supplementary Table S2.1), and is classified as a Category II EDC in DEDuCT. In sum, chemicals in ViCEKb are found in a variety of consumer products but are not regulated for their skin sensitizing effects.

2.2.3 ViCEKb shows low overlaps with skin sensitizing chemical lists and exposome databases

The CompTox Chemicals Dashboard [21–23] is one of the largest publicly available resources that aggregates lists of chemicals evaluated for their toxic effects. The dashboard provides four chemical lists pertaining to skin sensitizing chemicals (Supplementary Table S2.6), and we observed a low overlap with the chemicals in ViCEKb (Figure 2.6c). Only two chemicals with 'Patch test done' evidence, namely p-phenylenediamine (CID:7814) and 2-mercaptobenzothiazole (CID:697993), have been documented for their skin sensitizing effects in an occupational setup. Cinnamaldehyde (CID:637511) was

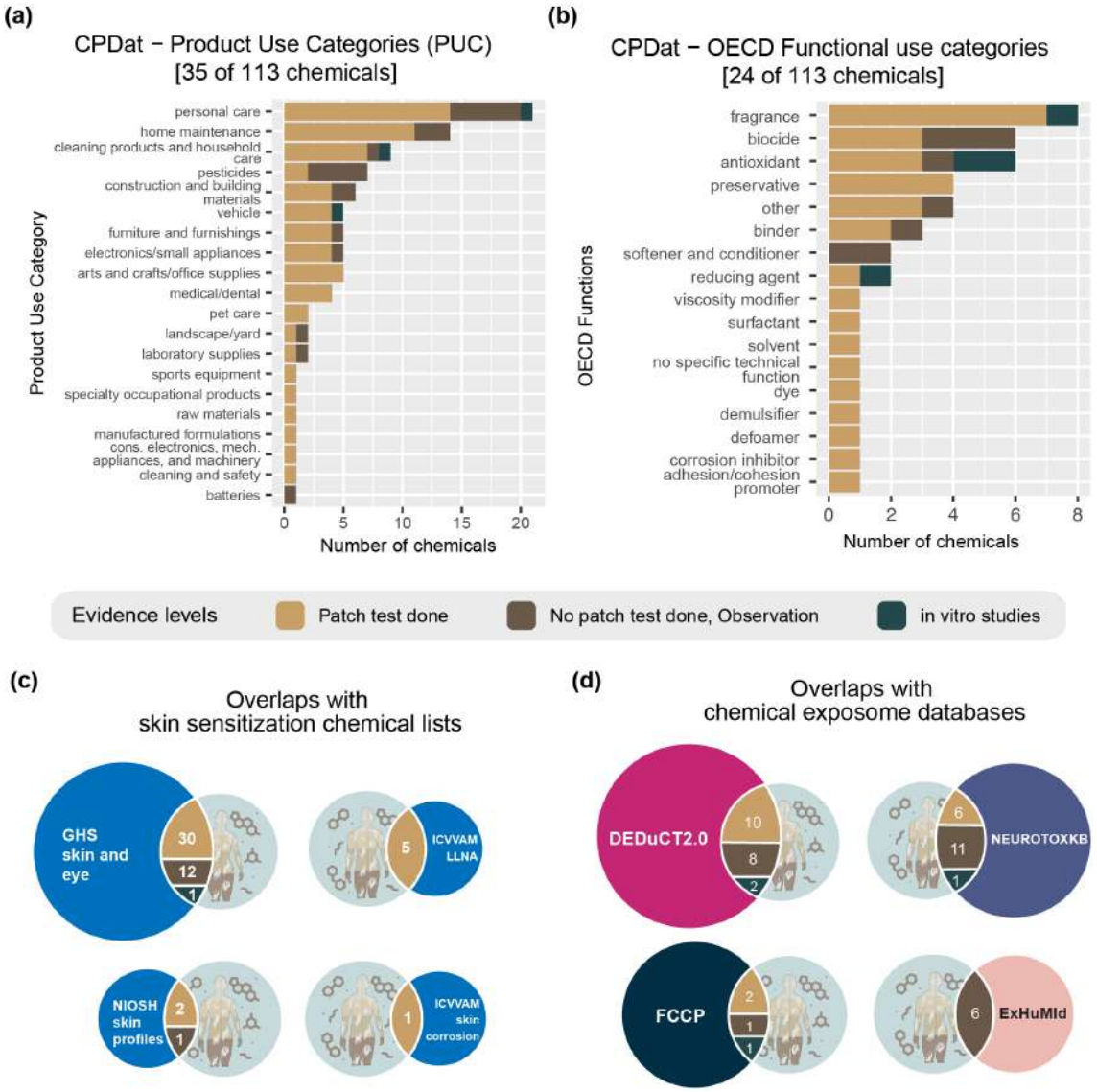


Figure 2.6: The presence of vitiligo-triggering chemicals in consumer products, skin sensitizing chemical lists and other chemical exposome databases. The Chemical and Products Database (CPDat) was leveraged to obtain the consumer product-specific data. (a) 35 of 113 chemicals from ViCEKb are categorized into 20 product use categories (PUC) in CPDat. (b) 24 of 113 chemicals in ViCEKb are categorized into 17 Organisation for Economic Co-operation and Development (OECD) standardized functions. (c) Number of ViCEKb chemicals overlapping with skin sensitizing chemical lists from CompTox Chemicals Dashboard. (d) Number of ViCEKb chemicals overlapping with chemical exposome databases such as DEDuCT 2.0, NeurotoxKb, FCCP and ExHuMId. In each plot, the chemicals have also been classified by evidence level in ViCEKb.

the only chemical with 'Patch test done' evidence from ViCEKb prioritized for further laboratory testing for skin corrosion.

Alternatively, to assess other adverse effects caused by chemicals in ViCEKb, we leveraged multiple chemical exposome databases (Supplementary Table S2.6) and observed low overlap (Figure 2.6d). Among the few overlaps, hexachlorophene (CID:3598), having 'Patch test done' evidence in ViCEKb, is also documented in NeurotoxKb as a neurotoxicant. Hexachlorophene mainly functions as a preservative and is found in disinfectants and chalkboard cleaners. Another chemical having 'Patch test done' evidence in ViCEKb, 4-tert-butylcatechol (CID:7381), is documented as Category III endocrine disrupting chemical (EDC) in DEDuCT 2.0. 4-tert-butylcatechol has a variety of functional uses, including fragrance component in cosmetic products and aiding the production of epoxy resins.

2.2.4 Cheminformatics-based analysis of ViCEKb chemical space revealed its uniqueness

We performed different cheminformatics-based analyses to gain better insights into the chemical space in ViCEKb. First, we characterized the chemicals by structure and found that most are organic compounds, with benzenoids being the most prevalent (Supplementary Table S2.1). Next, we visualized the similarity among the chemicals in ViCEKb through a CSN and observed that the chemicals are highly diverse, with only 28 of the 113 chemicals showing $\geq 50\%$ similarity (Figure 2.7; Supplementary Table S2.7). The high similarity backbone of the CSN shows 12 distinct chemical clusters, with the largest clusters containing six chemicals each (Figure 2.7).

Finally, we computed the chemical scaffolds and observed 56 distinct scaffolds across 94 of the 113 chemicals present in ViCEKb (Figure 2.8a), where the benzene scaffold was the most represented scaffold (26 of 94) (Supplementary Table S2.1). Notably, we observed that the scaffolds in ViCEKb are the most diverse (AUC – 0.69)

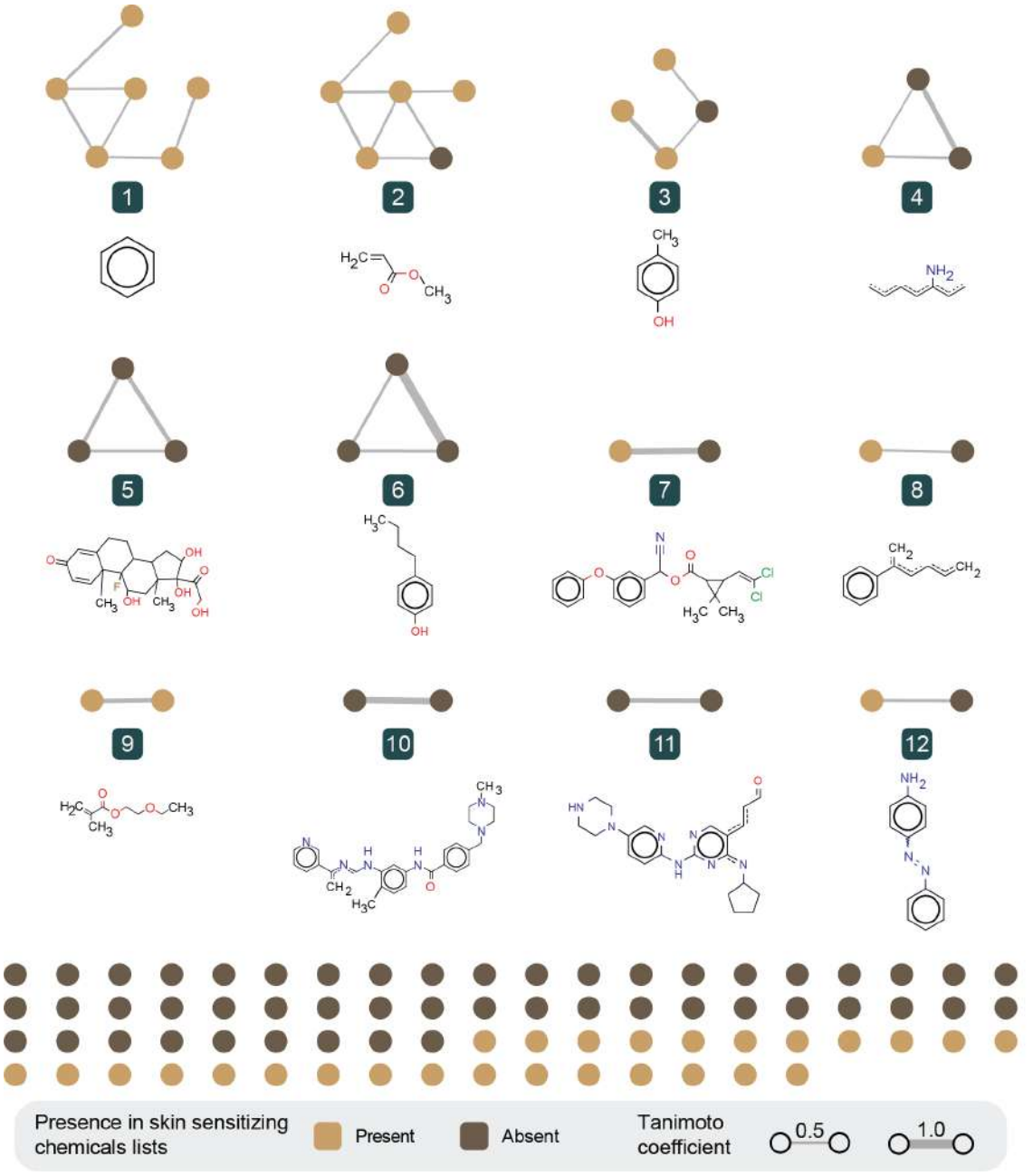


Figure 2.7: Chemical similarity network (CSN) of 113 chemicals in ViCEKb. The nodes of the network are chemicals, denoted by their presence in the skin sensitizing chemical lists obtained from the CompTox Chemicals Dashboard. The edges denote the Tanimoto coefficient computed between the ECFP4 fingerprints of the respective chemicals. The CSN is visualized using Cytoscape software. RDKit was employed to compute the maximum common substructure for each of the 12 connected components identified in the CSN.

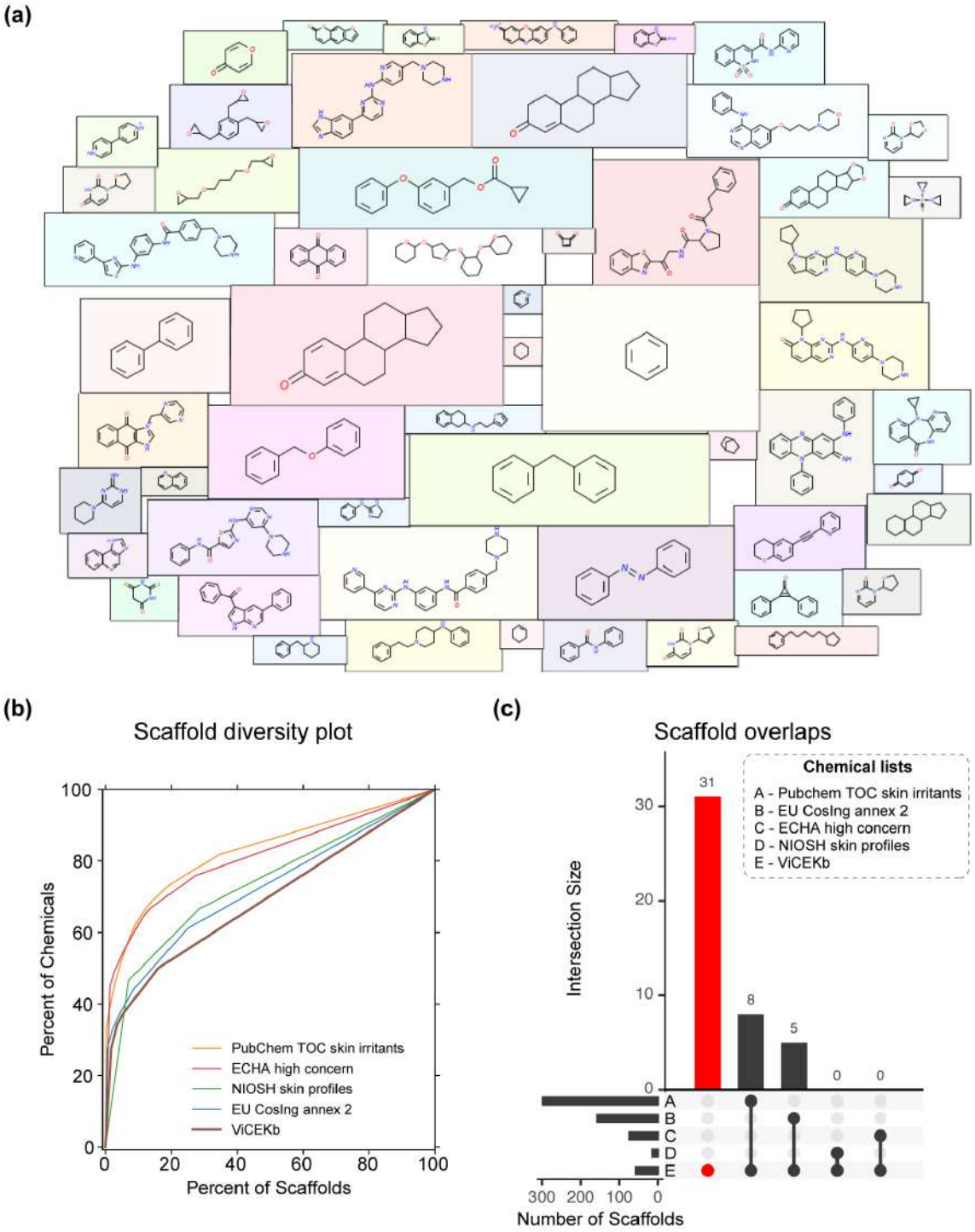


Figure 2.8: The chemicals in ViCEKb are diverse and unique. (a) A scaffold cloud representation of the 56 unique scaffolds identified from 94 of the 113 chemicals in ViCEKb. The scaffolds are computed using RDKit based on the Bemis-Murcko definition. (b) Cyclic System Retrieval (CSR) plot of chemicals in VICEKb, PubChem TOC: skin irritants, ECHA high concern chemicals, NIOSH skin profiles, and EU CosIng annex 2. This plot shows that the ViCEKb is the most diverse in its scaffold set. (c) UpSet plot representation of unique overlaps between chemicals in ViCEKb and the other chemical lists, plotted using the UpSetR package in R.

compared to those present in cosmetics and skin-related regulation lists (Table 2.1; Figure 2.8b). Additionally, we observed that 31 of 56 distinct scaffolds in ViCEKb are unique and not captured in these cosmetics and skin-related regulations (Figure 2.8c). These 31 scaffolds span across 32 chemicals, of which six chemicals (1,4-butanediol diglycidyl ether (CID:17046), neomycin (CID:8378), 2,3,5-triglycidyl-4-aminophenol (CID:22058861), thiotepa (CID:5453), monobenzone (CID:7638), 3-(diethylamino)-7-(phenylamino)phenoxazin-5-ium (CID:18711204)) have 'Patch test done' evidence. Of these chemicals, neomycin (CID:8378), 2,3,5-triglycidyl-4-aminophenol (CID:22058861), and 3-(diethylamino)-7-(phenylamino)phenoxazin-5-ium (CID:18711204) have not been documented in skin sensitizing chemical lists.

	M	**N**	**N_{sing}**	**N/M**	**N_{sing}/M**	**N_{sing}/N**	**AUC**	**P_{50} (%)**
ViCEKb	94	56	47	0.596	0.500	0.839	0.691	16.071
NIOSH skin profiles	30	14	10	0.467	0.333	0.714	0.733	14.286
EU CosIng annex 2	301	156	117	0.518	0.389	0.750	0.721	14.744
ECHA high concern chemicals	220	73	53	0.332	0.241	0.726	0.810	4.110
PubChem TOC: skin irritants	1069	298	195	0.279	0.182	0.654	0.827	4.362

Table 2.1: Scaffold diversity of chemicals in ViCEKb compared against other chemical lists. For each list, M represents the total number of chemicals, N represents the count of scaffolds, N_{sing} represents the total number of scaffolds that appear in only one chemical (singleton), AUC represents the area under the corresponding CSR curve, and P_{50} represents the percentage of scaffolds required to retrieve 50% of the chemicals.

2.2.5 Vitiligo-triggering chemicals are associated with similar biological responses

As the cheminformatics-based analysis of ViCEKb revealed a high structural diversity among chemicals, we were further interested to explore the nature of their biological responses. A transcriptomics-based approach has been used in pharmacology to explore biological similarity between drugs in the context of drug repurposing [27, 148]. By employing a similar approach to identify the similarity among chemicals triggering vitiligo, we obtained skin tissue specific chemical signatures for 18 chemicals, and identified their significant DEGs (Table 2.2; Figure 2.9a). We performed pathway enrichment analysis

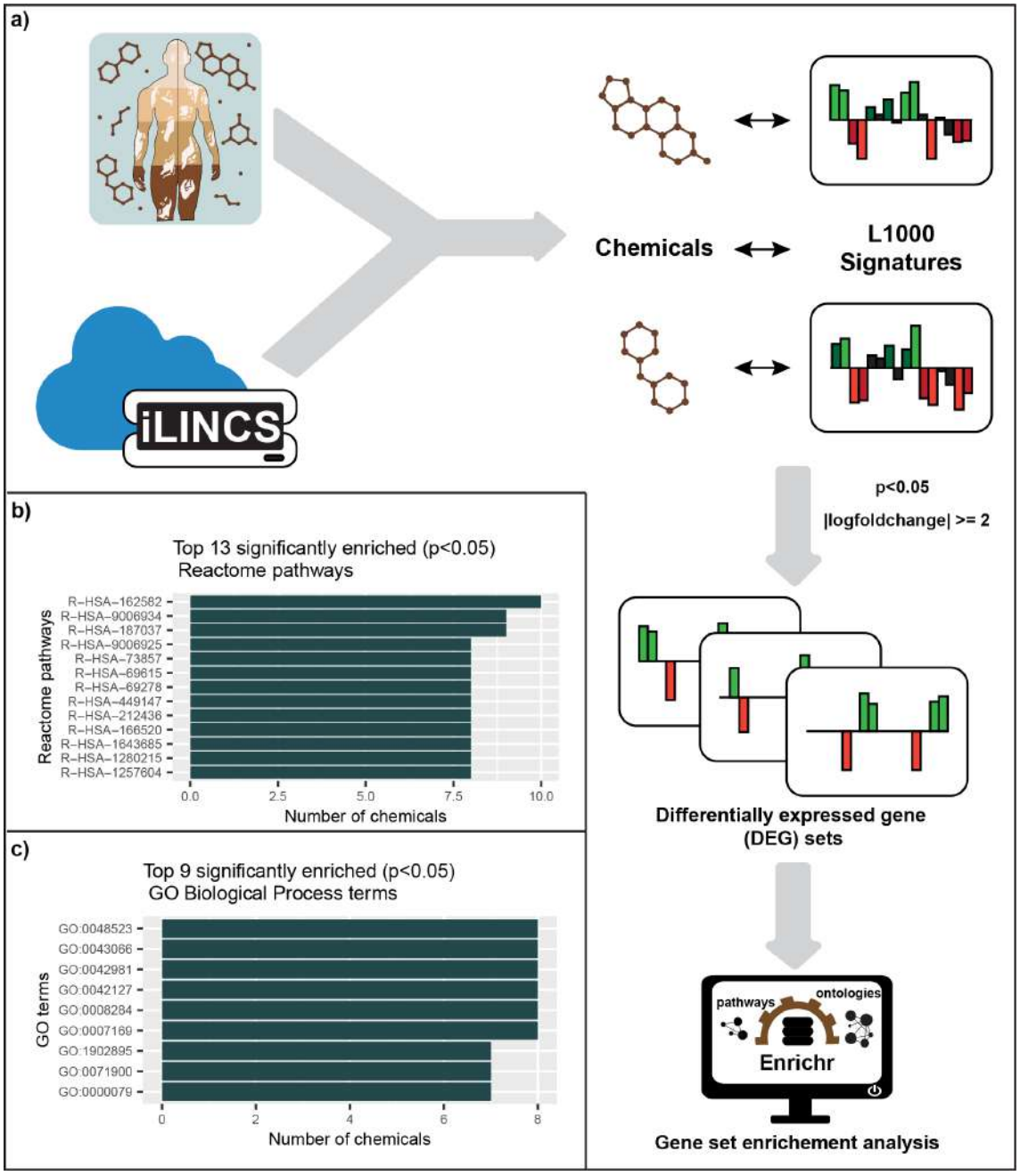


Figure 2.9: Analysis of transcriptomics signatures of ViCEKb chemicals. (a) The workflow to analyze the significantly enriched pathways and gene ontology (GO) terms from transcriptomics signatures of ViCEKb chemicals. (b) Top 13 significantly enriched Reactome pathways present in ViCEKb chemical sets. (c) Top nine significantly enriched GO biological process terms present in ViCEKb chemical sets.

for the different sets of DEGs, and identified that signal transduction pathway (R-HSA-162582) is significantly enriched in 10 of the 18 chemicals (Figure 2.9b; Supplementary Table S2.8). The Gene Ontology (GO) enrichment also shows various cellular signaling processes significantly enriched across chemicals (Figure 2.9c; Supplementary Table S2.9). In particular, receptor tyrosine kinase signaling pathway is significantly enriched across 9 of 18 chemicals. Tyrosine kinase signaling is known to play a key role in autoimmune and inflammatory skin disorders [149], thereby highlighting a biological connection between these nine chemicals. Notably, these nine chemicals are categorized under medical exposure, and most of them have been predicted to be highly permeable in skin [114].

Chemical	**PubChem identifier**	**LINCS signature identifier**	**Number of upregulated genes**	**Number of downregulated genes**	**Number of DEGs**
Mitoxantrone	CID:4212	LINCSCP_486	124	145	269
Latanoprost	CID:5311221	LINCSCP_178719	132	103	235
Sepantronium	CID:10126189	LINCSCP_2144	73	110	183
Diphencyprone	CID:65057	LINCSCP_1913	50	90	140
Palbociclib	CID:5330286	LINCSCP_2376	58	61	119
Stavudine	CID:18283	LINCSCP_1948	90	22	112
Dasatinib	CID:3062316	LINCSCP_2169	59	50	109
Vemurafenib	CID:42611257	LINCSCP_2385	36	40	76
Triamcinolone acetonide	CID:6436	LINCSCP_3096	14	20	34
Triamcinolone	CID:31307	LINCSCP_2916	11	11	22
Imatinib	CID:5291	LINCSCP_179309	9	7	16
Nicotinamide	CID:936	LINCSCP_1802	7	8	15
Atomoxetine	CID:54841	LINCSCP_2866	6	9	15
Methylprednisolone	CID:6741	LINCSCP_178224	5	8	13
Nevirapine	CID:4463	LINCSCP_3080	4	3	7
Thiotepa	CID:5453	LINCSCP_2587	5	1	6
Clofazimine	CID:2794	LINCSCP_1917	0	1	1
Isotretinoin	CID:5282379	LINCSCP_2428	1	0	1

Table 2.2: L1000 signatures of 18 ViCEKb chemicals obtained from iLINCS web-based platform. The upregulated and downregulated genes are shortlisted using constraints $p < 0.05$ and $|\text{logfoldchange}| \geq 2$.

2.3 Discussion

Vitiligo is a complex disease, where both the genetic predispositions and environmental triggers contribute to its pathogenesis [82]. While the genetic factors underlying vitiligo have been extensively studied [84, 86], their environmental factors have been less explored. Therefore, in this chapter, we developed a first-of-its-kind resource that comprehensively catalogs chemical triggers of vitiligo with published evidence, namely, Vitiligo-linked Chemical Exposome Knowledgebase (ViCEKb), and made it accessible for further research at: https://cb.imsc.res.in/vicekb/.

Developing ViCEKb involved an extensive manual effort in systematically curating published literature to identify 113 unique chemical triggers of vitiligo, and standardizing their reported evidence and sources of exposure (Figure 2.1). Importantly, we noted that only one of these 113 chemicals, 2-benzyl-4-chlorophenol (CID:8425), was previously reported to be associated with vitiligo in Haz-Map [150], while none of these chemicals were reported to be associated with vitiligo in Comparative Toxicogenomics Database (CTD) [151]. This highlights the importance of ViCEKb in bridging the existing knowledge gap regarding the environmental triggers of vitiligo.

A cheminformatics-based exploration of the chemical space of ViCEKb revealed its diversity in comparison with different chemical regulation lists (Figure 2.8). Notably, many of the chemicals are found in a diverse range of consumer products, including cosmetics (Figure 2.6). A direct comparison with other chemical exposome databases revealed a low overlap of chemicals, emphasizing the novelty and uniqueness of the data curated in ViCEKb. Finally, integration of skin-specific chemical-induced transcriptomics signatures of structurally diverse chemicals from ViCEKb, highlighted a significant commonality in their linked biological processes.

Importantly, this knowledgebase allows for the identification and prioritization of vitiligo triggering chemicals. For instance, propyl gallate (CID:4947) is documented in IFRA transparency list [152] as a chemical used in fragrance formulations used in

consumer goods. Flutamide (CID:3397) is documented as a safe chemical product in cosmetics [153]. Both these chemicals have the highest level of evidence ('Patch test done') in ViCEKb and are predicted to have high permeability through skin [114], but are not considered as skin sensitizing chemicals. Additionally, ToxCast skin sensitization assays for both propyl gallate and flutamide reveal downregulation of genes identified as biomarkers [86] in vitiligo pathogenesis.

Overall, the creation and subsequent computational characterization of a vitiligo-triggering chemical space has helped identify significant regulatory oversight. The integration of biological data has further aided in understanding potential mechanisms through which environmental triggers can lead to vitiligo, providing a brief overview of the underlying pathology of the disease. Therefore, we expect the observations from this chapter will motivate the efforts to decipher the complex etiology of vitiligo, which can in turn aid in the proper diagnosis and treatment of the disease.

Supplementary Information

Supplementary Tables S2.1-S2.9 associated with this chapter are available for download from the GitHub repository: https://github.com/asamallab/PhDThesis-Nikhil_C/blob/main/SI/ST_Chapter2.xlsx.

Chapter 3

DEDuCT 3.0: An enhanced and expanded FAIR-compliant resource and toxicology knowledge graph for endocrine disrupting chemicals

Endocrine disrupting chemicals (EDCs) are of particular concern to human and environmental health, as they can perturb the endocrine system, resulting in adverse developmental, physiological, and reproductive outcomes, among others [154–156]. Growing evidence of declining fertility rates and reproductive health problems has increased concern about EDCs and has led to more studies and regulations across global bodies to mitigate their effects [155, 156]. In response to this need, the Database of Endocrine Disrupting Chemicals and their Toxicity Profiles (DEDuCT) has emerged as a resource that organizes published evidence and makes scientific data on EDCs easily accessible for better understanding [105, 120]. The French Agency for Food, Environmental and Occupational Health & Safety (ANSES) has recognized the DEDuCT's four-stage criteria for extracting information from published literature as an important tool to gain knowledge on potential EDCs from published studies, and moreover, had used the data for designing their Second

French National Strategy for Endocrine Disruptors (SNPE2) [157]. Furthermore, DEDuCT has also been instrumental in providing a curated dataset of chemicals for the development of *in silico* methods for the prediction of EDCs [158, 159]. Additionally, the curated endocrine-mediated endpoint data within DEDuCT has been extensively integrated with the Adverse Outcome Pathway (AOP) framework to explore chemical-induced toxicity mechanisms across different classes of chemicals [55, 56]. In short, DEDuCT is an important resource not only for academic purposes but also one that has gained regulatory relevance.

Since the release of the previous version of DEDuCT in 2020, a large body of literature on EDCs has been generated. In parallel, while toxicologically-relevant databases, such as the United States Environmental Protection Agency's ToxCast [18, 20] and CompTox Chemicals Dashboard [22], the Comparative Toxicogenomics Database (CTD) [24], and AOP-Wiki [50], have been continuously updated and expanded, they do not specifically focus on the unique mechanisms and endpoints associated with EDCs. Although the previous version of DEDuCT included ToxCast data, it lacked integration with information from other resources and did not capture chemical structure information from the CompTox Chemicals Dashboard. Therefore, to provide a single point of access to all EDC-related data and align the DEDuCT knowledgebase with Findable, Accessible, Interoperable and Reusable (FAIR) principles [16], there is a need to curate data from recently published literature and integrate it with data from these external resources. Moreover, network toxicology frameworks such as the AOP and toxicology knowledge graphs help organize this diverse data and provide a strong basis for gaining novel mechanistic insights.

In this chapter, we describe an updated FAIR-compliant version of DEDuCT that compiles EDC-relevant data from the latest published evidence and other toxicological databases. Next, we explore the toxicity mechanisms of these chemicals by constructing EDC-AOP networks that integrate toxicological data from diverse sources. We further construct a large-scale toxicology knowledge graph and use it to gain mechanistic insights into the endocrine-mediated effects of chemicals curated in this resource. We further

utilize different cheminformatics techniques to explore the chemical space of DEDuCT and independently examine the existing regulatory coverage of these chemicals. **The work reported in this chapter is contained in the published manuscript [74].**

3.1 Methods

3.1.1 Workflow to identify endocrine disrupting chemicals

The World Health Organization (WHO) and the United Nations Environment Programme (UNEP) report titled '*State of the Science of Endocrine Disrupting Chemicals 2012*' states that 'A potential endocrine disruptor is an exogenous substance or mixture that possesses properties that might be expected to lead to endocrine disruption in an intact organism, or its progeny or (sub) populations' [160]. In line with this definition, Karthikeyan *et al*. [105] had earlier developed a four-stage workflow to systematically compile and curate potential EDCs and their observed adverse effects specific to endocrine disruption in supporting published evidence. This effort resulted in the development of the Database of Endocrine Disrupting Chemicals and their Toxicity profiles (DEDuCT), which is one of the largest resources on endocrine disrupting chemicals (EDCs), that systematically compiles chemical-associated toxicities via extensive manual curation from published literature [105, 120]. Importantly, a working group of 18 experts convened by the French Agency for Food, Environmental and Occupational Health and Safety (ANSES) have independently verified and recognized this four-stage workflow as an important tool for curating potential EDCs from published literature [157].

The previous version, DEDuCT version 2.0 (released on 2 October 2020) compiled information on 792 potential EDCs with supporting evidence from 2218 articles [120]. Since then, there has been a continued worldwide effort in testing and identification of EDCs, and this has resulted in an additional large corpus of published literature. Therefore, in this chapter, we present an updated version, DEDuCTv3.0, that compiles information on 1043 potential EDCs from 3269 research articles published until January 2025 (Supplementary

Table S3.1). To achieve this, we implemented the systematic four-stage workflow involving comprehensive manual curation of chemicals and their associated endocrine-mediated effects in humans or rodents, as reported in published literature (Figure 3.1) [105]. This four-stage workflow is summarized in the following subsections.

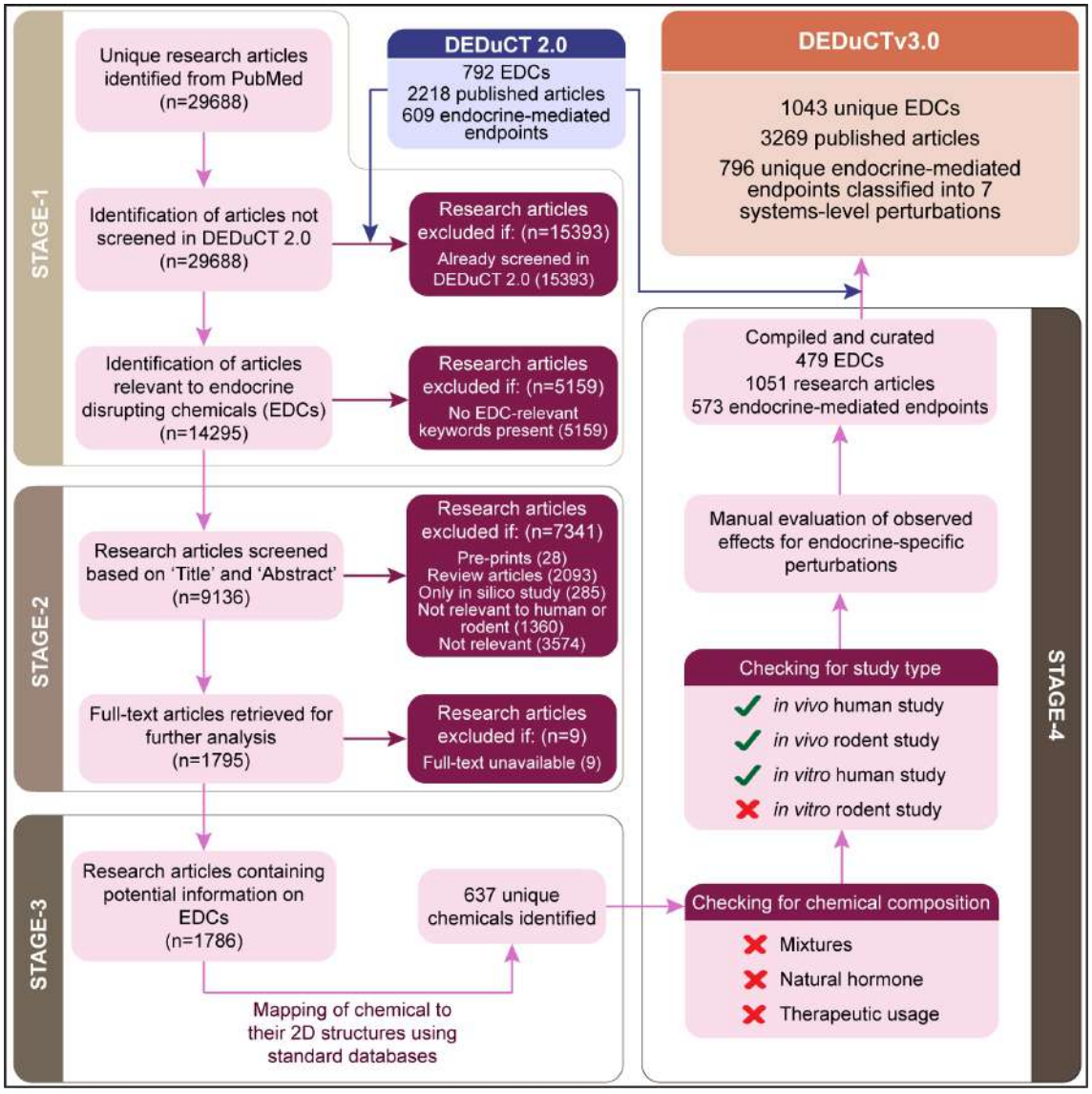


Figure 3.1: The workflow to identify published literature containing potential endocrine disrupting chemicals and their associated endpoints is presented as a PRISMA statement.

Stage 1: Mining EDC-relevant literature

The first stage involves mining EDC-relevant published literature from various sources. In comparison to the previous version, we observed that the sources for published literature,

such as the World Health Organization (WHO) report [160], the Endocrine Disruption Exchange (TEDx) [161], and EDCs Databank version 2015 [162], were no longer updated or accessible. Therefore, we relied only on published literature retrieved from PubMed [91] using the following query:

"EDCs" OR "EDC" OR ("endocrine" AND "disrupt") OR ("disrupt" AND "endocrine") OR "endocrine disruptors" OR "endocrine-disruptors" OR "endocrine disruptor" OR "endocrine-disruptor" OR "endocrine disrupters" OR "endocrine-disrupters" OR "endocrine disruption" OR "endocrine-disruption" OR "endocrine disruptive" OR "endocrine-disruptive" OR "endocrine disrupting" OR "endocrine-disrupting" OR "endocrine disrupter"

We last conducted this search on 21 January 2025, without any additional time range filters, and retrieved a list of 29688 published research articles. Among these, we identified 14295 articles that had not been screened in the previous two versions of DEDuCT. Thereafter, we manually screened the titles and abstracts of these 14295 new articles to verify the relevance of the article, ensuring that the identified keywords were specifically associated with EDCs and their related toxicities. Through this process, we identified 9136 new EDC-relevant articles at the end of stage 1 (Figure 3.1).

Stage 2: Literature filter based on experimental design and test organism

In stage 2 of the workflow, we manually screened 9136 articles filtered from stage 1 to identify studies based on *in vivo* evidence in humans, or *in vivo* experimental evidence in rodents, or *in vitro* experimental evidence in human or rodent cell-line studies. During this stage, we filtered out the studies comprising receptor-based binding assays or *in silico* methods to infer endocrine disrupting potential, as the evidence was insufficient to determine if the chemical can lead to adverse effects based on environmental exposures. Additionally, we excluded review articles and correlation-based studies as they lack experimental evidence directly linking chemical exposure to the observed effects. Through this screening process, we identified 1795 new research articles relevant to EDC-specific

experiments in humans or rodents at the end of stage 2 (Figure 3.1).

Stage 3: Compilation of reported chemicals from filtered research articles

In stage 3 of the workflow, we systematically curate information on the tested chemicals within the published literature. To achieve this, we first obtain the full texts of the articles for conducting a comprehensive screening. We obtained full texts of 1786 of 1795 articles identified from stage 2, and thereafter curated the chemical data, and mapped the chemicals to their two-dimensional (2D) structure identifiers using standard chemical databases such as PubChem [11] and Chemical Abstracts Service (CAS) [93]. This process enabled the consolidation of synonymous chemical names and standardization of chemical representations across studies. Next, we excluded chemicals that lacked standardized structural information due to insufficient chemical composition data. Through this systematic process, we identified 637 unique chemicals that had been tested for endocrine disruption in humans or rodents across 1786 research articles (Figure 3.1).

Stage 4: Identification of potential EDCs with supporting evidence on endocrine-mediated perturbations

In stage 4 of the workflow, we assessed the tested chemicals for the presence of significant endocrine-associated adverse effects in the published literature. First, we discarded chemicals if they were a natural hormone, had associated evidence where they were solely tested as part of a mixture, or were studied only for their therapeutic effects without mentioning endocrine-relevant adverse effects. Next, we discarded studies that relied solely on *in vitro* rodent experiments. Thereafter, we checked the observed adverse effects or endpoints for significance in the associated literature and then evaluated their endocrine disruption relevance, such as changes in morphology, physiology, reproduction, growth, development, and lifespan. Finally, we considered a chemical as a potential EDC if it contained at least one significant supporting evidence for endocrine-relevant effects from associated literature. Through this extensive manual curation, we identified 479 chemicals with strong supporting evidence of endocrine disruption from 1051 published articles. Notably, 251 of

these 479 chemicals were not compiled in previous versions of DEDuCT.

In earlier versions of DEDuCT, the reported endocrine-mediated adverse effects were standardized into a unified vocabulary of 609 endocrine-mediated endpoints describing the toxicological profiles of chemicals [105]. These endpoints were further grouped into seven systems-level perturbation categories, namely, Endocrine-mediated cancer (CT), Developmental (DT), Hepatic (HT), Immunological (IT), Metabolic (MT), Neurological (NT), and Reproductive (RT) endocrine-mediated perturbations [105, 120]. Therefore, we standardized the curated adverse effects against the existing endocrine-mediated endpoints, and defined new endpoints only where standardization was not possible. Through this extensive manual curation process, we identified 573 unique endocrine-mediated endpoints, of which 187 were newly defined. These were further categorized into seven systems-level perturbations based on their underlying biological processes (Supplementary Table S3.2).

Finally, we integrated the compiled data with DEDuCT version 2.0. This updated resource now contains 1043 unique potential EDCs and 796 unique endocrine-mediated endpoints spanning a total of 3269 publications.

3.1.2 Enriching DEDuCTv3.0 with additional chemical information for EDCs

In addition to experimental evidence curated from the published evidence, we enriched DEDuCTv3.0 with additional chemical and biological data from external sources. For the 1043 potential EDCs within DEDuCTv3.0, we compiled diverse chemical information including two-dimensional (2D) and three-dimensional (3D) chemical structures, physicochemical properties, predicted absorption, distribution, metabolism, excretion and toxicity (ADMET) properties, and molecular descriptors. We further curated EDC-associated gene, phenotype, and disease information from Comparative Toxicogenomics Database (CTD) [24, 25], active readouts from ToxCast invitrodb v4.3 [20], associated AOPs from AOP-Wiki [50], consumer product presence information from Chemical and

Products Database v4.0 (CPDatv4) [163], and presence in various regulation lists and human biospecimens.

We used ClassyFire [96] to hierarchically classify the 1043 potential EDCs based on their structure (Supplementary Table S3.1). Where ClassyFire failed to classify the chemical, we relied on the structural definitions to obtain the Kingdom of such chemicals [96]. We further classified the chemicals into four categories (I-IV) (Supplementary Table S3.1), based on the type of experimental evidence supporting endocrine disruption in humans or rodents. These categories are: (i) Category I: chemicals with *in vivo* human evidence; (ii) Category II: chemicals with both *in vivo* rodent and *in vitro* human evidence; (iii) Category III: chemicals with only *in vivo* rodent evidence; (iv) Category IV: chemicals with only *in vitro* human evidence.

3.1.3 Exploration of chemical space within DEDuCTv3.0

We explored the chemical space of 1043 chemicals within DEDuCTv3.0 using a chemical similarity network (CSN) and a target similarity network (TSN). To construct the CSN, we considered the 1043 chemicals as nodes, and the edges were based on the structural similarity between corresponding nodes. Here, we computed the structural similarity between chemicals as the Tanimoto coefficient based on their ECFP4 fingerprints [105,131]. Thereafter, we considered chemical pairs with Tanimoto coefficients > 0.5 to be structurally similar. We further utilized these Tanimoto coefficients as edge weights when visualizing the network.

To construct TSN, we curated the chemical-gene interaction information from ToxCast invitrodb v4.3 [20]. We first identified chemicals with active assay endpoint information, based on the 'hitc' value ≥ 0.9 from the '`mc5-6_winning_model_fits-flags_invitrodbv4_3_AUG2024.csv`' file. In some cases, these active assay endpoints could be non-specific, as the chemicals show activity only in a narrow concentration range that overlaps with that of cell stress or cytotoxicity, a phenomenon termed the cytotoxicity-

associated burst [164]. Therefore, to ensure specific assay endpoint associations for chemicals, we applied the following Z-score statistic proposed by Judson *et al.* [164]:

$$Z(\text{chemical, assay}) = \frac{-\log AC_{50}(\text{chemical, assay}) - \text{median}[-\log AC_{50}(\text{chemical, cytotox})]}{\text{global cytotoxicity MAD}}$$

where, '$\log AC_{50}$(chemical, assay)' is the logarithm of the AC_{50} value of the chemical in the assay, '$\log AC_{50}$(chemical, cytotox)' is the logarithm of the AC_{50} value of the chemical in the corresponding cytotoxicity assay, and the 'global cytotoxicity MAD' is the median of the MAD (median absolute deviation) of '$\log AC_{50}$(chemical, cytotox)' distributions across all chemicals [164].

Next, we retrieved the global cytotoxicity MAD and $\log AC_{50}$(chemical, cytotox) (given by the column titled 'cytotox_median_log') from the '`cytotox_invitrodb_v4_3_AUG2024.xlsx`' [19], and discarded assays with Z-scores between +3 and -3 [164] to retain chemical-specific active assay endpoints. Subsequently, we retrieved the assay-specific gene information from the file '`assay_target_mappings_invitrodb_v4_3_AUG2024.xlsx`', and filtered out genes that were not specific to humans or rodents. Through this extensive procedure, we identified 579 of 1043 chemicals targeting 400 human- or rodent-relevant genes. Thereafter, we computed chemical similarities using the Jaccard index [165] of the corresponding target gene sets [105]. Finally, we considered chemical pairs with a Jaccard index > 0.5 to be target-wise similar. We further used these Jaccard index values as the edge weights when visualizing the network.

Finally, we probed the correlation between structural similarity and target similarity for the 579 of 1043 chemicals with both structural and target information from ToxCast. We plotted the structural similarity based on the Tanimoto coefficient on the y-axis, and the target similarity based on the Jaccard index on the x-axis [105]. Thereafter, we computed the Pearson correlation coefficient to check for correlation between structural and target similarities [105].

3.1.4 Construction of stressor-centric adverse outcome pathway (AOP) networks for chemicals in DEDuCTv3.0

The stressor-centric AOP network framework provides a holistic understanding of diverse AOPs associated with a stressor, eventually revealing the diversity in the stressor-induced toxicity mechanisms, and aiding in their risk assessment [56, 166]. Although in an earlier study, Ravichandran *et al.* [54] explored endocrine-relevant AOPs within AOP-Wiki, they did not provide much detail on the mechanisms associated with individual chemicals in DEDuCT. Therefore, in this chapter, we systematically curated AOPs from AOP Wiki, and through an integrative approach detailed in the following subsections, associated the chemicals with these AOPs and constructed an EDC-AOP network.

Compilation of AOPs within AOP-Wiki

AOP is a toxicological knowledge framework that captures biological processes underlying stressor-induced toxicity [47]. In this framework, toxicity-related biological events, referred to as key events (KEs), are organized sequentially, beginning with the interaction of the stressor with a biological target, known as the molecular initiating event (MIE), and culminating in an adverse outcome (AO) [47–49, 167]. The causal, directional relationships between these events are termed key event relationships (KERs) [47–49, 167]. AOPs are stressor-agnostic and context-specific, meaning each AOP is tailored to a particular biological scenario [47–49, 167]. Developed globally, these AOPs are deposited in the AOP-Wiki [50], which is the largest, publicly accessible repository. AOP-Wiki hosts several AOPs, where each AOP is documented in the form of KEs (including MIE and AO) and KERs, all of which are supported by scientific evidence [51]. Therefore, in this chapter, we relied on AOP-Wiki to obtain the latest available AOPs.

First, we downloaded the XML file (released 1 April 2025) from the 'Project Downloads' page in AOP-Wiki to obtain the latest information on AOPs. Then, we parsed the data to extract information associated with AOPs like AOP identifier, AOP title, associated

KEs (including MIEs and AOs) and KERs, linked stressors, handbook version followed for development of AOP, status according to Organisation for Economic Co-operation and Development (OECD), and biological applicability information such as taxonomy, sex and life-stage of the organism, and their corresponding weight of evidence. Additionally, for each KE, we extracted information like KE title, KE identifier, level of biological organization, action name, object name, object identifiers, and process name, and information associated with KERs like upstream/downstream KEs, evidence for biological plausibility of KER, adjacency, and the extent of quantitative understanding of KER.

Curating complete and connected AOPs within AOP-Wiki

AOPs are considered living documents as they are collaboratively developed globally, and are continuously updated based on new scientific evidence [51]. Consequently, many AOPs may remain incomplete or lack sufficient information for further analysis. Therefore, we systematically checked each AOP to filter high quality and complete AOPs [55–57]. First, we manually checked AOPs and removed the ones that comprised KEs with title as 'unknown' or lacked any KEs or KERs. Next, we used the NetworkX library [168] in Python to check for disconnected AOPs, and manually inspected those that contained disconnected components, and updated them before filtering out AOPs with disconnected components. Finally, we checked for the presence of MIE, AO, and a directed path between them, and filtered out those that lacked any. Through this combined manual and computational effort, we obtained 385 complete, connected and high quality AOPs from AOP-Wiki, which we then designate as 'curated AOPs' (Supplementary Table S3.3). These 385 high confidence AOPs comprised 1228 unique KEs (Supplementary Table S3.4) and 1966 unique KERs (Supplementary Table S3.5).

Subsequently, we classified the AOs of the 385 curated AOPs using the standardized disease ontology from the Disease Ontology (DO) database [169, 170]. This classification resulted in 28 disease classes across 385 AOPs (Supplementary Table S3.3). We could not map 148 out of 385 curated AOPs to any DO terms based on their respective AOs,

and we therefore designated them as 'Unclassified'. Among the 28 disease classes, we observed that 'Cancer' emerged as the most represented category, encompassing 46 of the 385 curated AOPs (Supplementary Table S3.3).

Computation of cumulative WoE of AOPs

AOP-Wiki provides a weight of evidence (WoE) for each KER in an AOP based on its biological plausibility (Supplementary Table S3.5). The WoE is a qualitative score expressed as 'High', 'Moderate', 'Low', and 'Not Specified'. Following Ravichandran *et al.* [54], we computed the fractions of KERs within an AOP with 'High' WoE [represented as F('High')], 'Moderate' WoE [represented as F('Moderate')], 'Low' WoE [represented as F('Low')], and 'Not Specified' WoE [represented as F('Not Specified')]. Subsequently, we assigned a cumulative WoE to the AOPs based on the following criteria:

i. High: If F('High') ≥ 0.5

ii. Moderate: If F('High') < 0.5, but (F('High') + F('Moderate')) ≥ 0.5

iii. Low: If (F('High') + F('Moderate')) < 0.5, but (F('High') + F('Moderate') + F('Low')) ≥ 0.5

iv. Not Specified: If none of the above criteria are satisfied

Supplementary Table S3.3 provides the cumulative WoE for the 385 curated AOPs. We then proceeded to map the chemicals in DEDuCTv3.0 to the curated AOPs.

Identification of KEs using AOP-Wiki

AOP-Wiki catalogs information on prototypical stressors for each AOP, based on well-documented associations with these stressors [51]. Based on the data obtained for the 385 curated AOPs, we identified 100 chemicals in DEDuCTv3.0 as prototypical stressors for 93 AOPs. Subsequently, we considered the 361 KEs within these AOPs to be associated with these 100 chemicals (Supplementary Table S3.6).

Identification of KEs using CTD

The CTD is a comprehensive public resource that links environmental chemicals (C), genes (G), phenotypes (P), and diseases (D) to advance understanding of their effects on health [24]. This resource facilitates the construction of CGPD-tetramers, which help identify confident associations between chemicals, phenotypes, and diseases, enabling their mapping to KEs within AOP-Wiki [55–57]. Therefore, we retrieved high confidence CGPD-tetramers associated with chemicals in DEDuCTv3.0, and subsequently leveraged them to identify KEs within AOP-Wiki.

CTD provides a chemical dictionary linking standard PubChem identifiers and CASRN to CTD chemical identifiers (IDs). Initially, we obtained a list of 804 unique CTD chemical IDs based on overlaps with chemical data compiled in DEDuCTv3.0. Upon closer inspection, we observed that some of the associations provided by CTD were incorrect. For instance, benzo(a)pyrene (DED000576) was linked to two chemical IDs within CTD based on its PubChem identifier (CID:2336). We noted that one of the chemical IDs, MESH:D001580 corresponded to the class of benzopyrenes, while MESH:D001564 corresponded to the chemical benzo(a)pyrene. Therefore, we manually screened all the associations to remove non-specific information within CTD, and identified 783 chemical IDs within CTD associated with 764 chemicals in DEDuCTv3.0.

Next, we accessed the CTD November 2025 release, and constructed the CGPD-tetramers, wherein the following were considered: (i) chemical-gene and chemical-phenotype associations that are relevant to humans and rodents, and are supported with literature evidence; (ii) chemical-disease and gene-disease associations with 'marker/mechanism' evidence, and are supported with literature evidence; (iii) gene-phenotype associations with gene ontology (GO) annotations based on only the experimental results in human or rodent-based studies [171]. Through this process, we obtained a list of 208809 CGPD-tetramers comprising 364 chemicals, 3311 genes, 1747 phenotypes, and 858 diseases.

Thereafter, we mapped the phenotypes and diseases to the KEs within AOP-Wiki

[55–57]. First, we considered the CGPD-tetramer phenotype GO terms as directly mapped to the corresponding KE when they are directly associated with it through the process identifier provided by AOP-Wiki. Next, we used the GOSim package [172] in the R programming language to generate the immediate neighboring GO terms for the CGPD-tetramer phenotype GO terms. We then overlapped these neighboring GO terms with the KE-associated process identifier to identify preliminary associations. Subsequently, we manually screened the KE titles and phenotype names to retain associations that matched the biological process outlined in the KE title and KE information page. Through this process, we linked 310 KEs to 279 phenotypes across 347 chemicals (Supplementary Table S3.6).

For disease mapping, we first obtained the disease-associated MeSH terms from the CGPD-tetramer disease terms. For each term, we screened the corresponding title and metadata information using the MeSH browser [173]. We then associated these terms with the relevant KEs based on underlying biological processes. Through this process, we linked 256 KEs to 548 diseases across 358 chemicals (Supplementary Table S3.6).

Identification of KEs using DEDuCTv3.0 endocrine-mediated endpoints and NeurotoxKb neurotoxicant endpoints

In DEDuCTv3.0, we systematically curated 796 endocrine-mediated endpoints associated with 1043 potential EDCs. We therefore relied on these endpoints to identify relevant KEs within AOP-Wiki. For each endpoint, we manually inspected the KE title and information, and checked for underlying biological processes before linking. Through this process, we linked 276 KEs with 361 endocrine-mediated endpoints across 927 chemicals (Supplementary Table S3.6).

NeurotoxKb [126, 127] is a manually curated resource focusing on mammalian neurotoxicity endpoints associated with environmental chemicals, compiled from published literature. We first extracted the neurotoxic endpoints corresponding to chemicals in DEDuCTv3.0, and considered them to identify relevant KEs within AOP Wiki. We manually

inspected the neurotoxic endpoints and KE titles in AOP Wiki, and linked 64 KEs with 55 neurotoxic endpoints across 123 chemicals (Supplementary Table S3.6).

Identification of KEs using ToxCast

ToxCast is a program by the United States Environmental Protection Agency (USEPA) designed to enhance chemical toxicity predictions through *in vitro* high-throughput screening of various environmental chemicals [18]. These *in vitro* approaches provide crucial information on the associated biological processes and gene alterations triggered by stressor interactions, which can aid in the identification of potential MIEs related to active chemicals [52, 55–57, 166]. Therefore, we relied on ToxCast invitrodb v4.3 [19, 20] to identify KEs associated with chemicals in DEDuCTv3.0.

Based on the procedure detailed in section 3.1.3, we first obtained chemical-specific active assay endpoints from ToxCast invitrodb v4.3. Next, we relied on the 'top' value of the corresponding winning model from the '`mc4_all_model_fits_invitrodbv4_3_AUG2024.Rdata`' file to determine whether these active chemicals exhibited an 'activatory' or 'inhibitory' effect [19]. Subsequently, we manually inspected these active assay endpoints, and KE titles to identify relevant KEs within AOP-Wiki. Through this extensive process, we linked 160 KEs with 318 assay endpoints across 523 chemicals (Supplementary Table S3.6).

Construction of stressor-AOP network

We finally integrated all these curated data to construct a stressor-AOP network for chemicals within DEDuCTv3.0. In order to obtain high confidence associations, we relied only on the 385 curated AOPs from AOP-Wiki (Supplementary Table S3.3). We first linked a chemical to an AOP if they shared at least one common KE. Thereafter, we characterized these chemical-AOP links based on two criteria, namely, the coverage score and the level of relevance. The coverage score of a stressor-AOP link is defined as the ratio of the number of KEs within that AOP associated with the stressor to the total number of KEs within that AOP [56, 174]. This score is a real-valued number between 0 and 1, and is denoted as the

edge weight of the linkage between a stressor and an AOP in the constructed stressor-AOP network. The level of relevance is a qualitative score used to identify the relevance of stressor-AOP association within the stressor-AOP network [56], and this score is denoted as an attribute of the edge in the constructed stressor-AOP network. Level of relevance is a five-level criterion and is defined as follows:

- Level 1: The stressor is associated with at least one KE within an AOP, where the KE is neither MIE nor AO within that AOP
- Level 2: The stressor is associated with at least one AO within an AOP, but not associated with any MIE within that AOP
- Level 3: The stressor is associated with at least one MIE within an AOP, but not associated with any AO within that AOP
- Level 4: The stressor is associated with at least one MIE and one AO within an AOP
- Level 5: The stressor is associated with at least one MIE and one AO within an AOP, and there exists a directed path between the associated MIE and AO

Supplementary Table S3.7 contains all the data on the stressor-AOP network constructed for chemicals in DEDuCTv3.0, including the coverage score and level of relevance for each of the chemical-AOP links. We have also provided these links in the corresponding chemical information pages on the DEDuCT webserver.

3.1.5 Construction of a toxicology knowledge graph for chemicals in DEDuCTv3.0

Toxicology knowledge graphs provide a framework to integrate and organize toxicological data from multiple sources, aiding in the identification of underexplored relationships, and eventually enhancing our understanding of chemical-induced toxicities [39]. Here, we constructed a toxicology knowledge graph by linking the chemicals and endocrine-mediated endpoints within DEDuCTv3.0 with gene, phenotype, disease, biological KEs, and AOPs from toxicology-relevant resources such as ToxCast, CTD, and AOP-Wiki (Figure 3.2). We

refer to this knowledge graph as 'DEDuCT-KG', and the following subsections summarize the steps involved in curation of links and nodes within the knowledge graph.

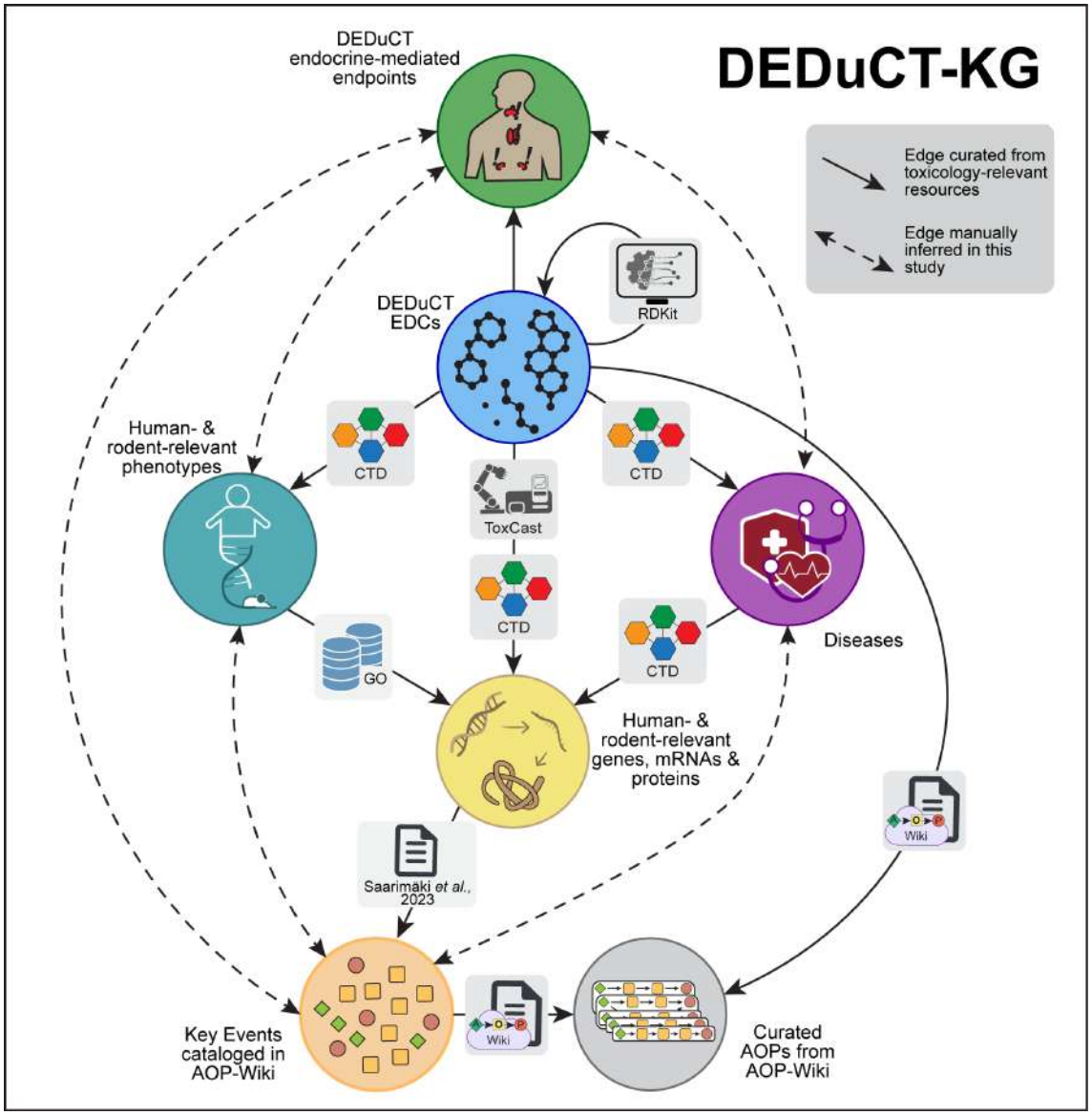


Figure 3.2: Network representation of different associations between seven types of nodes present in DEDuCT Knowledge Graph (DEDuCT-KG). The resources from where each association was obtained is mentioned on the edge connecting the corresponding nodes.

Curation of links between chemicals and endocrine-mediated endpoints

Since the endocrine-mediated endpoints associated with chemicals have been systematically curated from published literature in DEDuCTv3.0, we considered them as high confidence links between the chemicals and their associated endpoints. We assigned the

edge type as ‘reported_to_cause’ to all these chemical-endpoint links to better explain these links in the knowledge graph. Through this process, we identified 10532 links connecting 1043 chemicals with 796 endocrine-mediated endpoints in DEDuCT-KG.

Curation of links between chemicals and genes from CTD and ToxCast

CTD provides curated chemical-gene interactions obtained from published literature, where it compiles additional information on the gene form of the interacting gene, corresponding organism name and taxonomic identifier, chemical-gene interaction actions, and associated PubMed identifier (PMID) of the published literature [25]. First, we searched the NCBI taxonomy browser [175] with the corresponding organism identifiers and obtained information on the taxonomic classification such as the domain, clade, kingdom, phylum, and order. To obtain human- or rodent-relevant gene information, we selected only species such as *Homo sapiens* (organism id: 9606) and those belonging to the order ‘Rodentia’. Through this process, we identified 28 relevant species. Next, we filtered genes based on their gene form information to obtain those entries that were annotated as gene, mRNA, or protein. Further, we filtered out genes based on the associated interaction actions such as ‘abundance’, ‘synthesis’, ‘response to substance’, ‘reaction’, ‘cotreatment’, ‘transport’, ‘localization’, ‘degradation’, ‘export’, ‘import’, ‘metabolic processing’, ‘mutagenesis’, ‘secretion’, ‘uptake’, ‘stability’, and only retained those with terms such as ‘expression’ and ‘activity’. We observed that these interaction actions further included the kind of action, such as ‘increases’, ‘decreases’, or ‘affects’. For some of the chemical-gene links, we observed that there were different curated interaction actions across different studies. Therefore, to mitigate such discrepancies, we utilized the following standardization criteria: (i) if the interaction action included terms such as both ‘increases’/‘decreases’ and ‘affects’, then it was assigned to ‘increases’/‘decreases’ respectively; and (ii) in case both ‘increases’ and ‘decreases’ were present, then it was assigned as ‘affects’. Through this systematic standardization process, we identified 509996 links connecting 682 chemicals to 29560 genes. We assigned the corresponding standardized interaction actions as the edge types

within the knowledge graph.

We also obtained the chemical-gene interactions from human- or rodent-relevant data curated from ToxCast invitrodb v4.3 [20] as detailed in section 3.1.3. We identified 13205 links connecting 579 chemicals to 400 genes. We also curated the corresponding assay response, such as 'gain' or 'loss', for the obtained active assay endpoints. We further relied on the 'top' value of the corresponding winning model from the '`mc4_all_model_fits_invitrodbv4_3_AUG2024.Rdata`' file to determine whether these active chemicals exhibited an 'activatory' or 'inhibitory' effect [19]. To understand the effect of chemical on the target gene, we standardized these responses as follows: (i) responses such as 'activatory' or 'gain' were assigned as 'reported_to_activate'; (ii) responses such as 'inhibitory' or 'loss' were assigned as 'reported_to_inhibit'. We assigned these standardized response information as edge types within the knowledge graph.

Curation of links between chemicals and phenotypes or diseases from CTD

CTD provides curated chemical-phenotype interactions obtained from published literature, where it compiles additional information on the corresponding organism name and taxonomic identifier, chemical-phenotype interaction actions, and associated PMID of the published literature [25]. We followed the procedure detailed in the previous subsection to obtain human- or rodent-relevant chemical-phenotype interactions. Subsequently, we resolved discrepancies in chemical-phenotype interactions, and identified 28561 links connecting 599 chemicals to 4431 phenotypes. We assigned the standardized interaction actions as the edge types within the knowledge graph.

CTD also provides curated chemical-disease associations, where it additionally annotates the association as either 'marker/mechanism' or 'therapeutic' based on the evidence from corresponding published literature. According to the CTD terminology, 'marker/mechanism' corresponds to evidence suggesting that the chemical exposure correlates with the disease or may play a role in the etiology of the disease, while 'therapeutic' corresponds to evidence suggesting the therapeutic potential of the chemical in a dis-

ease [176]. Therefore, we obtained only associations marked as 'marker/mechanism', and identified 16219 links connecting 586 chemicals to 1805 diseases. We assigned the edge type 'causes_or_correlated_with' to better explain the chemical-disease associations in the knowledge graph.

Curation of links between genes and disease within CTD

CTD provides curated gene-disease associations, where it additionally annotates the association as either 'marker/mechanism' or 'therapeutic' based on the evidence from corresponding published literature [25]. According to CTD terminology, 'marker/mechanism' corresponds to evidence suggesting that the gene may be a biomarker for the disease or may play a role in the etiology of the disease, while 'therapeutic' corresponds to evidence suggesting the gene may be a potential therapeutic target in the treatment of the disease [176]. Therefore, we obtained only associations marked as 'marker/mechanism', and identified 27598 links connecting 7772 genes to 2466 diseases. We assigned the edge type as 'associated_with' to better explain the gene-disease associations in the knowledge graph.

Curation of links between genes and phenotypes from GO

GO provides annotated gene-phenotype associations, including the taxonomy identifier of the organism in which the association was identified, the kind of evidence, and a qualifier term to explain the association between the gene and phenotype. First, we obtained the GO annotations from the NCBI gene resource [177], and retained only gene-phenotype associations with experimental results [171]. Next, we filtered the links to retain only human- or rodent-relevant associations based on taxonomic information. We observed that each link was annotated with a qualifier term that defines the gene-phenotype association. We then filtered the links to remove qualifier terms that included 'NOT', and identified 304398 links connecting 36463 genes to 18311 phenotypes. We assigned the corresponding qualifier terms as the edge types within the knowledge graph.

Curation of links between endocrine-mediated endpoints and phenotypes or diseases in CTD

In DEDuCTv3.0, we manually curated the endocrine-mediated endpoints based on the biological endpoints reported in the supporting evidence. Mapping these endpoints to phenotypes and disease terms enables making the data more FAIR-compliant [16]. Therefore, we manually checked the definitions of the endpoints, phenotypes, and disease terms and mapped them based on underlying biological similarities. Through this process, we identified 581 links connecting 394 endocrine-mediated endpoints to 248 phenotypes and 78 diseases. We assigned the edge type 'related_to' to better explain these links in the knowledge graph.

Curation of links between chemicals, genes, phenotypes, diseases, endocrine-mediated endpoints, and key event information within AOP-Wiki

Integrating diverse toxicological data with information contained within the AOP-Wiki enables the exploration of diverse toxicity mechanisms associated with chemicals [55, 56]. Furthermore, integrating these data within a knowledge graph enables the exploration of previously missed associations and may potentially aid in elucidating the chemical-associated toxicity mechanisms that were otherwise less understood. Following the procedure detailed in section 3.1.4, we obtained all the endocrine-mediated endpoints, phenotypes, and diseases that are mapped to KE information within AOP-Wiki, and identified 2334 links connecting 411 phenotypes, 733 diseases, 320 endpoints, with 632 KEs within AOP-Wiki. We assigned the edge type 'related_to' to better explain these links in the knowledge graph. Next, we utilized the curated prototypical stressor information from AOP-Wiki and identified 175 links connecting 100 chemicals to 93 AOPs. We assigned the edge type as 'prototypical_stressor_of' to better explain these links in the knowledge graph.

Further, we obtained the gene-KE associations from data curated by Saarimäki *et al.* [178]. We first mapped the provided Ensembl gene IDs to corresponding Entrez gene

identifiers using the BioMart online tool [179], and identified 148171 links connecting 969 KEs and 14298 genes. We assigned the edge type 'mapped_to' to better explain these links in the knowledge graph.

Curation of links between chemicals based on structural similarity

To enable chemical similarity-based search across DEDuCT-KG, we computed the chemical similarities as Tanimoto coefficient based on their ECFP4 fingerprints [105, 131]. Subsequently, we considered the chemical pairs having Tanimoto coefficients > 0.5 as having structural similarity. Through this process, we identified 1609 links connecting 634 chemicals within DEDuCTv3.0. We assigned the edge type 'structurally_similar_to' to better explain these links in the knowledge graph.

Altogether, we constructed a large-scale EDC-specific toxicology knowledge graph, referred to as 'DEDuCT-KG', comprising 75106 unique nodes categorized under 7 node types, and 1065949 edges categorized under 31 edge types.

3.1.6 Exploration of potential EDCs across consumer products and chemical lists

The CPDatv4, curated and maintained by the USEPA, contains information on the presence of chemicals in consumer products [163]. Additionally, CPDatv4 provides standardized product use categorization (PUC), which aids in characterizing the real-world chemical exposures [163]. Therefore, we used CPDatv4 to explore the exposure scenarios associated with potential EDCs within DEDuCTv3.0. In addition to the PUC data, we also obtained the reported functional role of these chemicals in the corresponding products.

Next, we compiled 54 chemical lists from inventories, guidelines and regulations from public resources (Supplementary Table S3.8) to explore the extent of regulatory coverage of these potential EDCs. Based on Karthikeyan *et al.* [120], we classified 21 of the 54 lists as 'Substances in Use' (SIU), and the remaining 33 as 'Substances of Concern' (SOC). Moreover, we grouped these lists into ten broad categories based on their reported exposure

scenario (Supplementary Table S3.8). Next, we compiled the list of potential EDCs from DEDuCTv3.0, WHO report on EDCs [160], TEDx [161] and EDCs Databank version 2015 [162], and categorized them into three groups: (i) group I EDCs if they were present only in the SIU lists; (ii) group II EDCs if they were present in both SIU and SOC lists; or (iii) group III EDCs if they were present only in SOC lists. Finally, we utilized the United States High Production Volume (USHPV) chemical list (Supplementary Table S3.8) and the latest OECD High Production Volume (OECD HPV) chemical list (Supplementary Table S3.8), to identify potential EDCs that are currently being produced in high volumes across the globe.

3.1.7 Web interface and database management system

DEDuCT is an online resource accessible at: https://cb.imsc.res.in/deduct/. In this version update, we enhanced the online resource with the integration of all compiled data. In particular, we enhanced the chemical section with more chemical-specific information pages on curated chemical-gene, chemical-phenotype and chemical-disease associations from CTD [24, 25], curated ToxCast information from ToxCast invitrodb v4.3 [20], chemical-associated AOP data obtained from the constructed stressor-centric AOP network, and the consumer product presence information compiled from CPDatv4 [163]. The compiled data is stored in MariaDB [141], and is retrieved using Structured Query Language (SQL). The web interface was created using HTML, CSS, Bootstrap 4, jQuery [143] and PHP [142], and hosted on Apache webserver [146] running on Debian 9.4 operating system.

We implemented the newly constructed toxicology knowledge graph, DEDuCT-KG, using Neo4j [180]. We used an open source web-based user interface, Knowledge-Graph-UI (KG-UI) [181], to create an interactive interface for DEDuCT-KG. This user interface utilizes Cypher language [182] to query and retrieve the immediate neighbors and the shortest path between any two entities from the underlying knowledge graph [183].

3.2 Results

3.2.1 DEDuCTv3.0 - an enhanced resource on potential endocrine disrupting chemicals

The DEDuCT is a unique repository of potential EDCs and their associated endocrine-mediated effects that were systematically curated from published literature [105, 120]. Since the release of DEDuCT version 2.0 in October 2020, there has been a substantial increase in literature focusing on characterization of EDCs, with PubMed now cataloging more than 14000 EDC-associated articles that were not previously screened (Figure 3.1).

Here, we present an updated version, DEDuCTv3.0, cataloging 1043 unique potential EDCs, systematically curated from 3269 published articles (Figure 3.1; Supplementary Table S3.1). These chemicals were linked to 796 unique endocrine-mediated endpoints, which we further classified into seven functional categories (Supplementary Table S3.2). We observed that most endpoints were categorized as 'Reproductive endocrine-mediated perturbations' (Figure 3.3a). Moreover, we observed that 823 of 1043 chemicals were associated with endpoints categorized as 'Reproductive endocrine-mediated perturbations' (Figure 3.3b). Additionally, we utilized the standardized dosage information to compute the No Observed Adverse Effect Level (NOAEL) and Low Observed Adverse Effect Level (LOAEL) to enable further quantitative assessments of chemical-associated toxicities (Supplementary Table S3.9) [105].

We performed a chronological analysis of the 3269 articles and observed that nearly half of them were published between 2017 and January of 2025 (Figure 3.3c). Notably, we observed that the reliance on *in vitro* human experiments is gradually increasing over the years, suggesting the reliance on non-animal approaches towards EDC research (Figure 3.3c). In addition, we used the literature data to understand the identification of new EDCs over the years. To this end, we chronologically sorted the supporting evidence for each chemical, and considered the date of the earliest published article as the year of

identification. Through this analysis, we observed that the number of new potential EDCs identified has increased over the past few decades, aligning with the growth in research and interest regarding endocrine disruption (Figure 3.3d).

Notably, compared to previous versions, DEDuCTv3.0 introduces 251 newly curated chemicals and 187 new endocrine-mediated endpoints, demonstrating a significant expansion of coverage of chemicals and their associated biological effects. Furthermore, 150 of these 251 new chemicals have reported evidence from *in vitro* human tests, suggesting a methodological shift toward more efficient, human-relevant toxicity testing of potential EDCs (Supplementary Table S3.1). Such shifts are consistent with the new approach methodologies (NAMs) framework and underscore their utility in streamlining the identification of potential EDCs [156]. In addition, compared with other EDC-relevant data compilations, such as the WHO report [160], TEDx [161] and EDCs Databank version 2015 [162], DEDuCTv3.0 now compiles 435 chemicals not captured by any of these resources (Figure 3.4a). Importantly, among these 435 chemicals, 193 were newly compiled in DEDuCTv3.0 (Figure 3.4b), emphasizing the necessity for this effort to enable further research in EDCs.

An environmental source-based classification of EDCs in DEDuCTv3.0 revealed that 466 chemicals originated from 'Industry', while 461 chemicals originated from 'Consumer products' (Supplementary Table S3.1). Next, based on the hierarchical structural classification of chemicals provided by ClassyFire [96], we observed that 988 chemicals were classified as organic compounds, of which 242 chemicals were newly compiled in DEDuCTv3.0, while 55 were classified as inorganic compounds, of which 9 were newly compiled in DEDuCTv3.0 (Supplementary Table S3.1).

Furthermore, we visualized the chemical space within DEDuCTv3.0 as a CSN, comprising 1043 chemicals as nodes and the structural similarity > 0.5 as edges (Figure 3.5). We observed that there were 99 connected components containing at least two nodes, of which the largest connected component (LCC) comprised 248 nodes, while there were 409 isolated nodes (Figure 3.5). Notably, we observed that 90 of the 251 new chemicals in

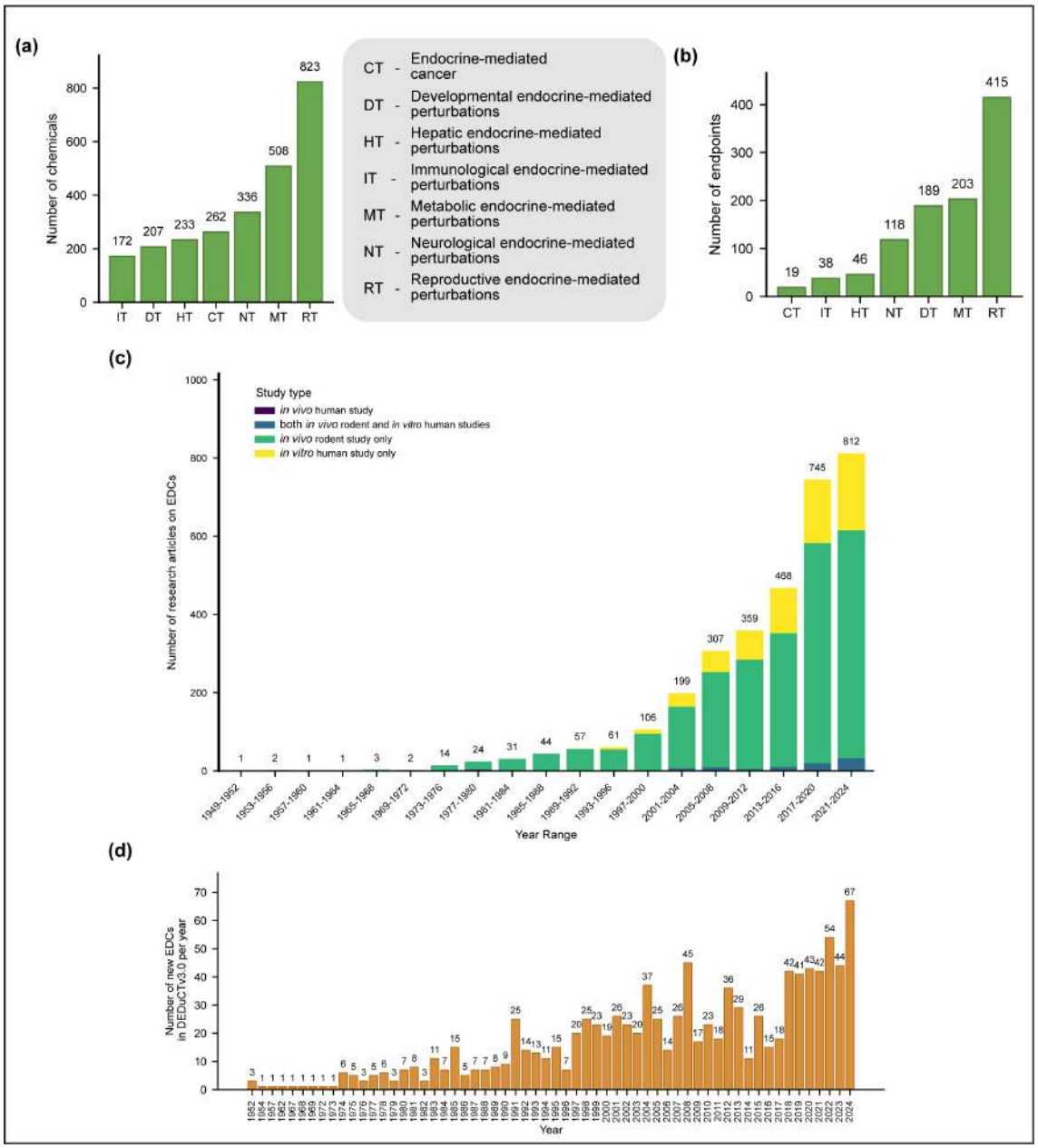


Figure 3.3: (a) The distribution of the 796 endocrine-mediated endpoints in DEDuCTv3.0 into the seven systems-level perturbation categories. (b) The distribution of chemicals within DEDuCTv3.0 into the seven systems-level perturbation events based on their associated endocrine-mediated endpoint obtained from published literature. (c) A chronological analysis of the 3269 published articles supporting 1043 potential EDCs in DEDuCTv3.0. Note, 32 articles published in January of 2025 are not included. (d) Bar plot depicting number of new EDCs identified per year based on published literature compiled in DEDuCTv3.0. Note, four new chemicals from January 2025 are not included.

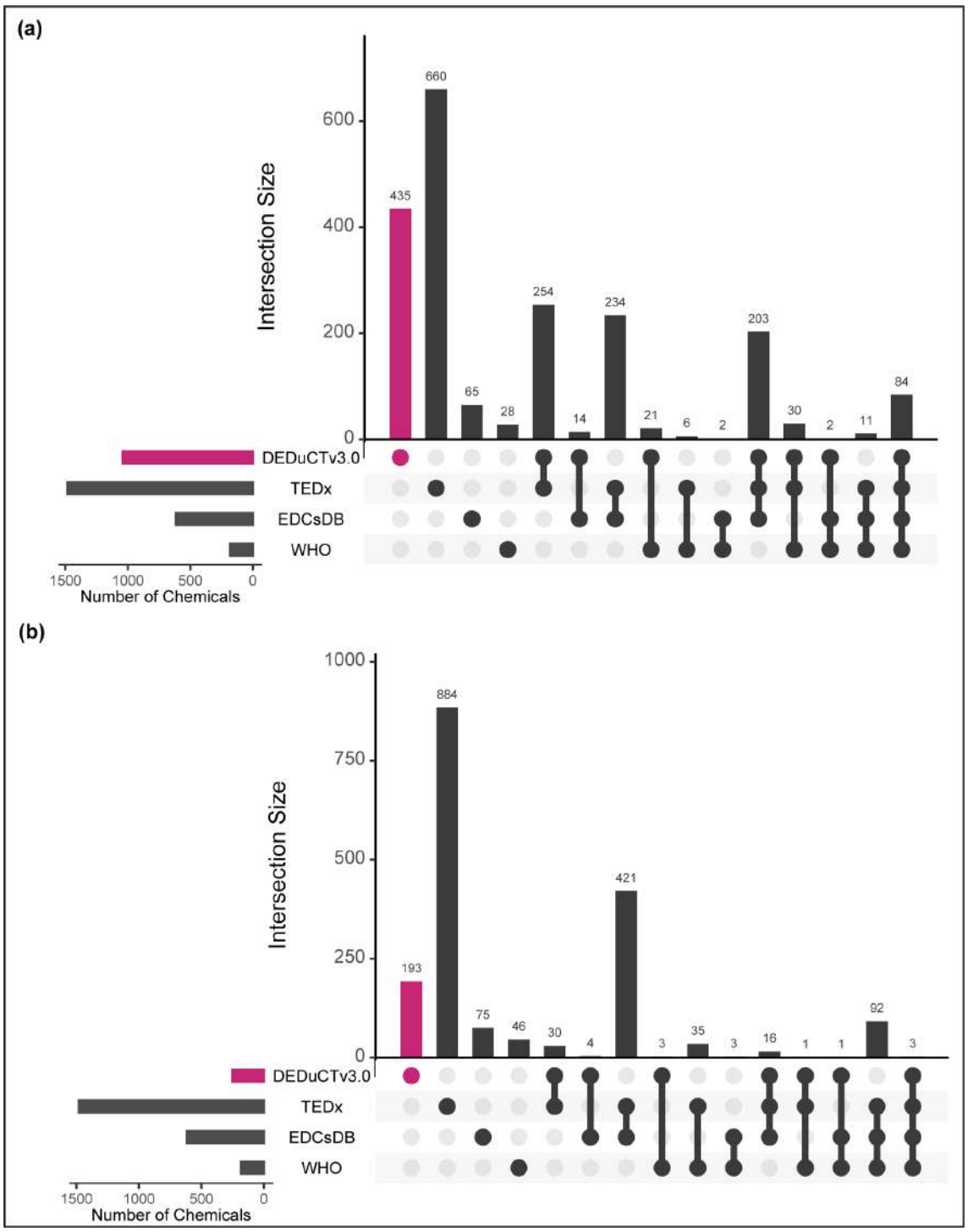


Figure 3.4: UpSet plot representation of unique overlaps between chemicals in DEDuCTv3.0 and other EDC-relevant databases such as the Endocrine Disruption Exchange (TEDx), EDCs Databank version 2015 (EDCsDB) and World Health Organization (WHO) report. (a) Comparison with all 1043 potential EDCs compiled in DEDuCTv3.0. (b) Comparison with the 251 new chemicals compiled in DEDuCTv3.0. These UpSet plots were generated using UpSetR package in R.

DEDuCTv3.0 were present as isolated nodes, suggesting they are structurally distinct from the chemicals previously compiled (Figure 3.5). Moreover, we observed that 145 of the 251 chemicals were structurally similar to chemicals compiled in DEDuCT version 2.0, while 16 others were structurally distinct (Figure 3.5). Subsequently, we used RDKit to compute chemical scaffolds according to the Bemis-Murcko definition [108]. We observed that 813 of 1043 chemicals share 323 unique scaffolds, among which 75 scaffolds are shared by at least two chemicals (Supplementary Table S3.1; Figure 3.6). Additionally, we visualized the chemical space as a TSN, comprising 579 of 1043 chemicals connected to 400 human- or rodent-relevant target genes within ToxCast invitrodb v4.3 [20]. In this TSN, we considered the 579 chemicals as nodes, and the target similarity > 0.5 as edges (Figure 3.7). We observed that there were 34 connected components containing at least two nodes, of which the LCC comprised 76 nodes, while there were 408 isolated nodes (Figure 3.7). This dissimilarity in EDC-associated gene targets suggests that these chemicals trigger diverse MIEs. Moreover, we observed a lack of correlation between the chemical and target similarity of the 579 of 1043 potential EDCs within DEDuCTv3.0 (Figure 3.8) [105].

Overall, DEDuCTv3.0 provides the most up-to-date and comprehensive information on potential EDCs and supports further research in systems toxicology and human exposome. Compared to previous versions, DEDuCTv3.0 further organizes the compiled data as a novel toxicology knowledge graph for EDCs, called DEDuCT-KG, thereby enhancing the utility of the resource for knowledge generation in EDC research.

3.2.2 Exploration of EDC-relevant toxicity mechanisms through stressor-AOP network

EDC research is increasingly transitioning towards the adoption of NAMs [156]. Within NAMs, the AOPs serve as an *in silico* toxicological framework for characterizing stressor-induced toxicity mechanisms and aiding in their risk assessment strategies [47, 156]. For

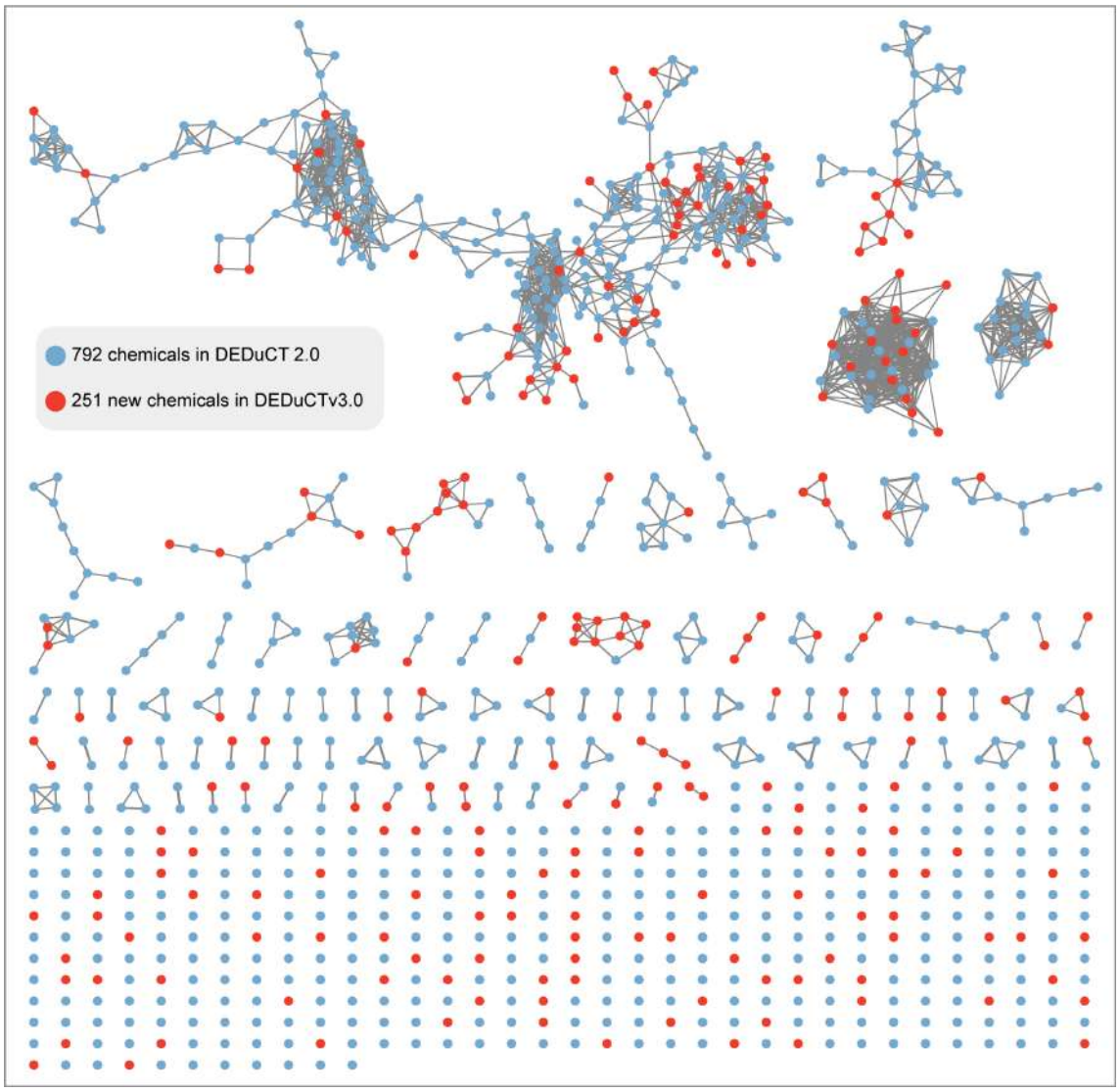


Figure 3.5: Network visualization of the chemical similarity network (CSN) constructed based on Tanimoto coefficients computed for each chemical pair using ECFP4 fingerprints. In this figure, edges are filtered to represent only those with Tanimoto coefficient > 0.5, with edge thickness scaled proportionally to the corresponding Tanimoto coefficient. The chemicals compiled in DEDuCT 2.0 are represented in blue, while the 251 new chemicals curated in DEDuCTv3.0 are shown in red.

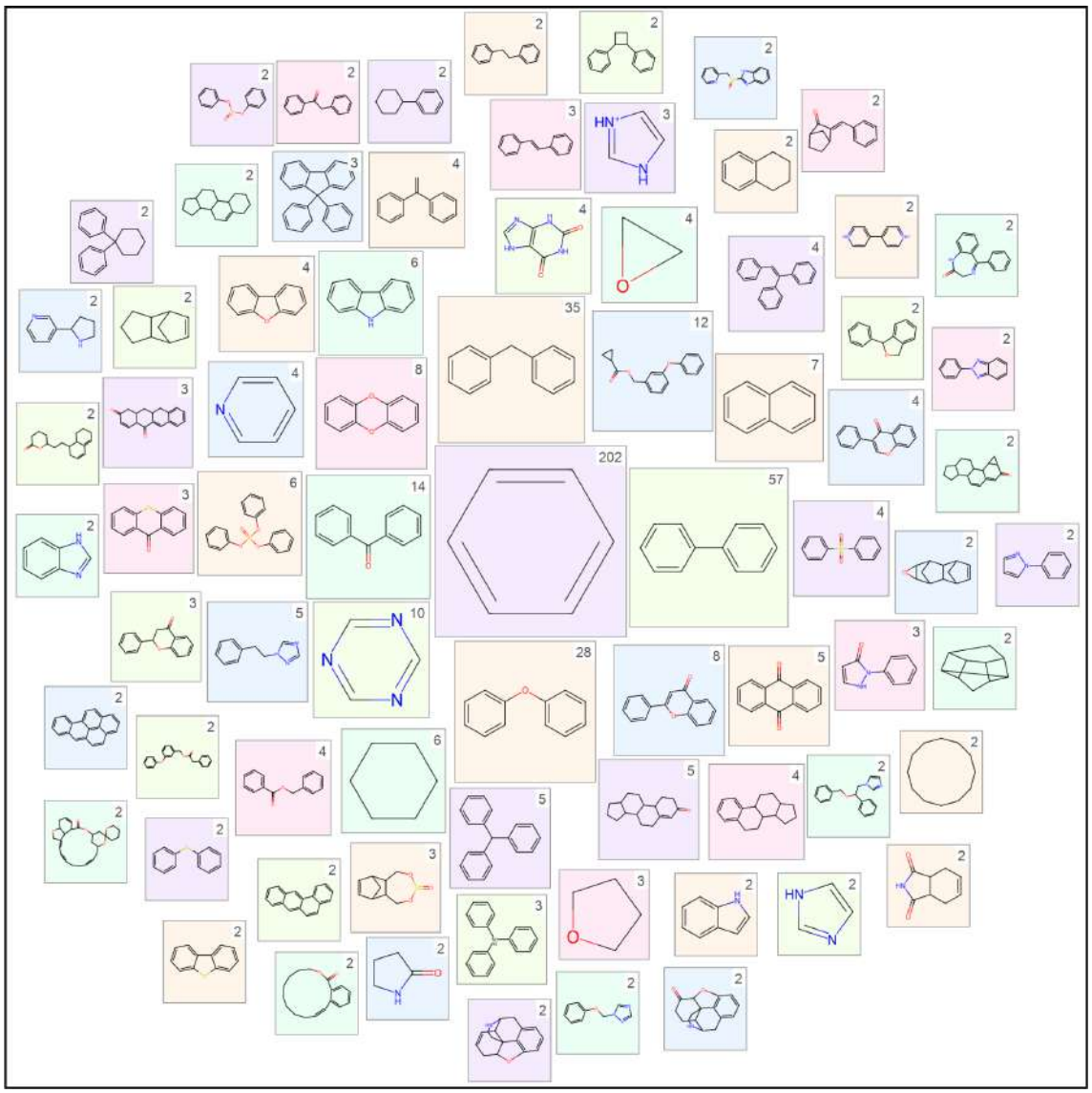


Figure 3.6: A scaffold cloud representation of the 75 unique scaffolds that are shared between at least two chemicals. The scaffolds are computed using RDKit based on the Bemis-Murcko definition. The number of chemicals sharing the scaffold is mentioned at the top-right corner of each box containing the scaffold.

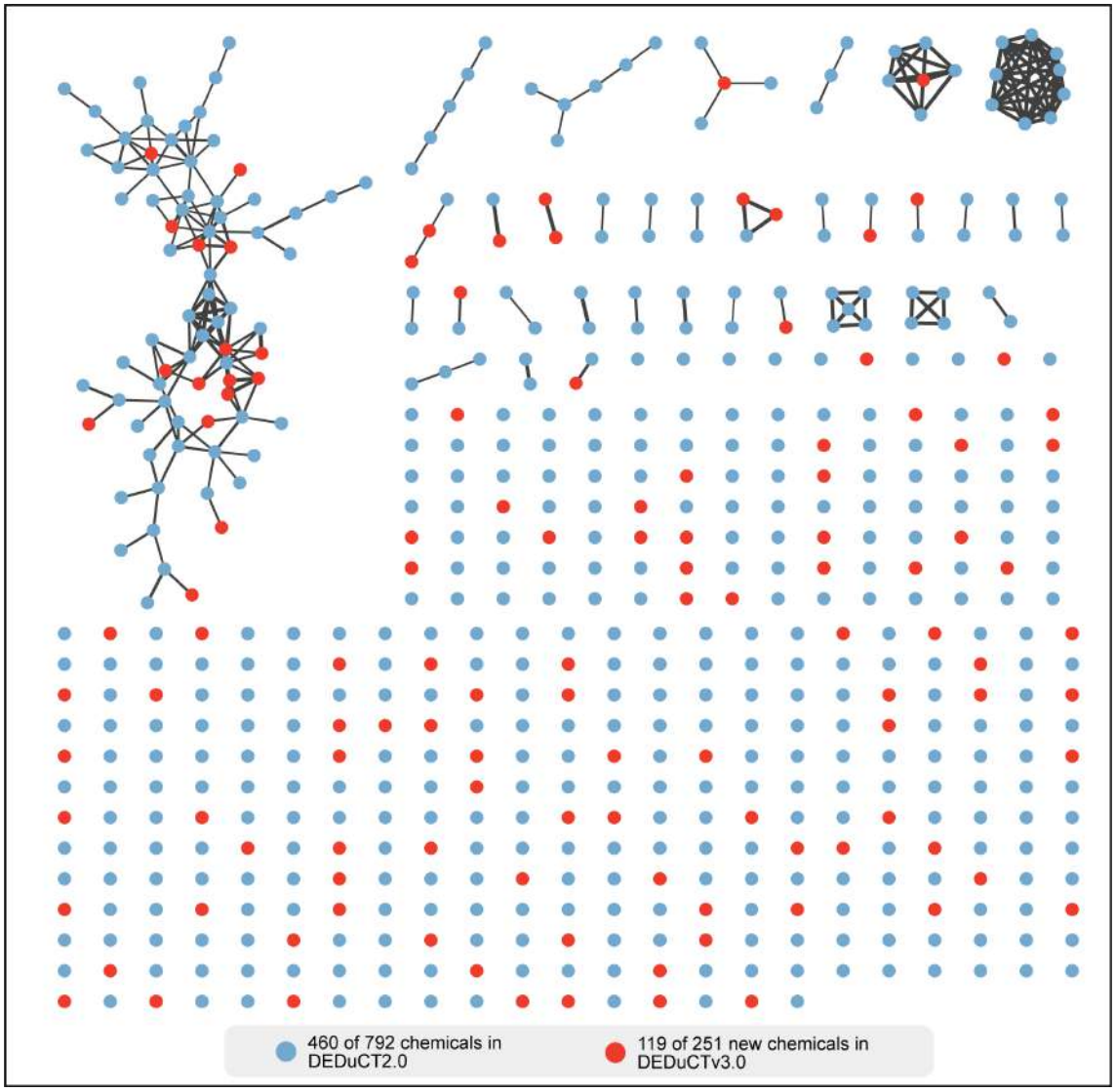


Figure 3.7: Network visualization of the target similarity network (TSN) constructed based on the Jaccard indices computed for each chemical pair based on shared gene targets. In this figure, edges are filtered to represent only those with Jaccard index > 0.5, with edge thickness scaled proportionally to the corresponding Jaccard index. The 460 of 792 chemicals compiled in DEDuCT 2.0 are represented in blue, while the 119 of 251 new chemicals curated in DEDuCTv3.0 are shown in red.

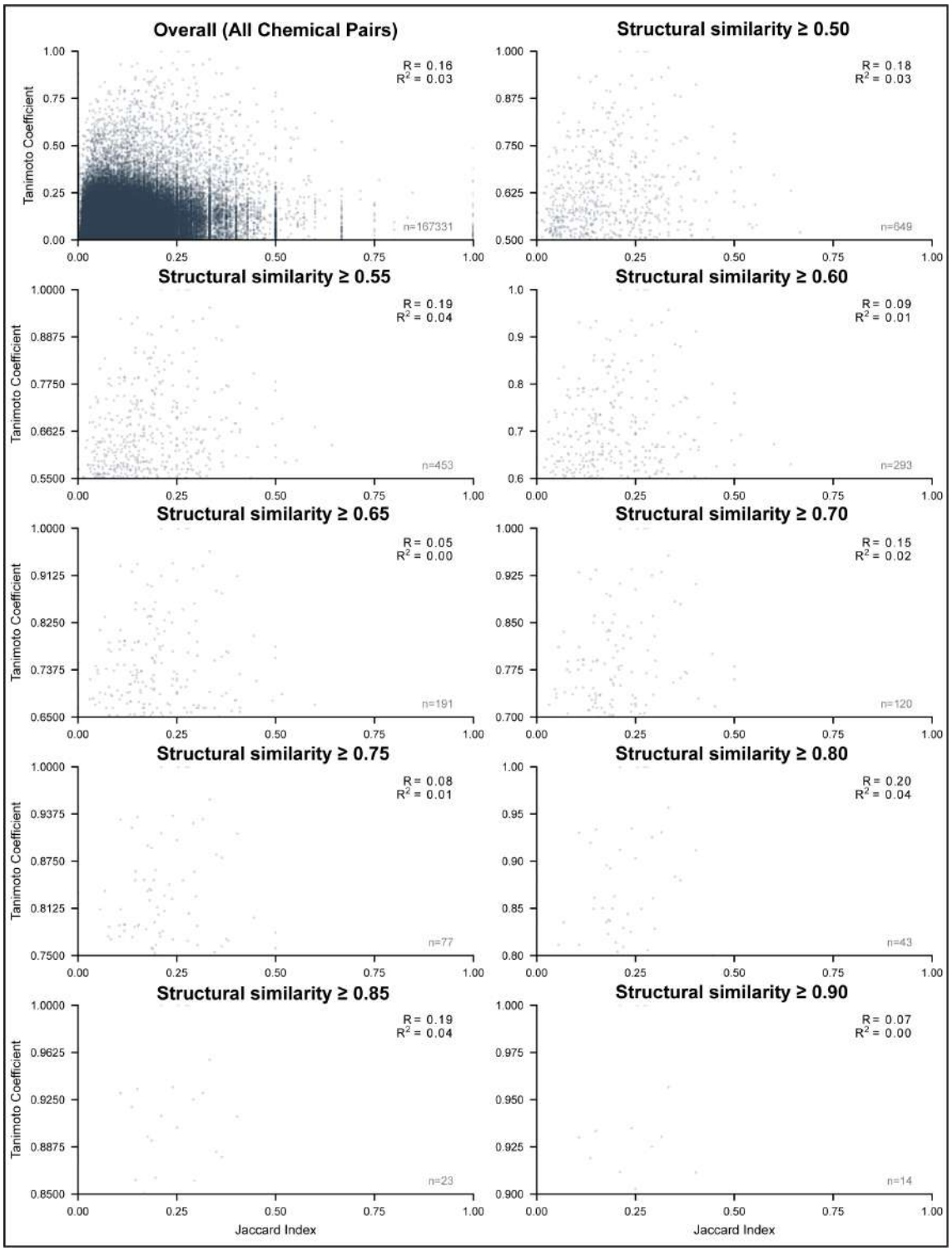


Figure 3.8: The scatter plot comparing structural and target similarity between chemical pairs, where structural similarity is represented by Tanimoto coefficient, and target similarity by the Jaccard index. The plots are further divided based on different structural similarity cut-off values. The Pearson correlation coefficient (R) and the coefficient of determination (R^2) values are denoted at the top-right of each graph. Further, the number of chemical pairs in each graph is denoted at the bottom right.

DEDuCTv3.0, we constructed the stressor-AOP network by connecting 949 of 1043 potential EDCs and 381 curated AOPs from AOP-Wiki [50], based on 759 associated Key Events (KEs) (Supplementary Tables S3.5-S3.5).

Supplementary Table S3.6 provides further details on these KEs and their associated chemical-endpoint associations, categorized by the source from which each association was derived. The constructed stressor-AOP network enabled the exploration of EDC-induced toxicity mechanisms through the AOP framework. We observed that bisphenol A (CAS:80-05-7) had maximum associated KEs (390 KEs), which is expected given its extensive toxicological characterization within DEDuCTv3.0 (Supplementary Table S3.6). It is immediately followed by cadmium (CAS:7440-43-9), which is linked to 289 KEs, and bis(2-ethylhexyl) phthalate (CAS:117-81-7), which is linked to 260 KEs (Supplementary Table S3.6). Moreover, we observed that the KE 'Increased, Cancer' (KE:885), which is also considered an adverse outcome across various curated AOPs, was associated with the highest number of chemicals, i.e., 348 of 949 chemicals, signifying that several chemicals have carcinogenic potential. Among the KEs annotated as MIE in at least one of the 381 curated AOPs, we observed that the MIE 'Activation, Estrogen Receptor Alpha' (KE:1065) was associated with the highest number of EDCs, i.e., 236 of 949 chemicals, suggesting that it represents one of the most frequently triggered initiating events underlying EDC-mediated toxicity pathways.

In the stressor-AOP network, we observed that 949 chemicals were linked to 381 curated AOPs through one or more key events (KEs) spanning multiple levels of biological organization. To further characterize these chemical-AOP associations, we classified them based on levels of relevance, wherein Level 5 denotes that there is evidence of both MIE and AO within an AOP associated with the chemical, and that the MIE and AO are connected through causal linkages [56]. We observed that 250 chemicals in DEDuCTv3.0 were associated with 215 curated AOPs from AOP-Wiki (Supplementary Table S3.7) at 'Level 5' level of relevance. Notably, we observed that 96 of these 250 chemicals were associated with 31 endorsed AOPs, signifying the utility of AOPs in characterizing EDC-

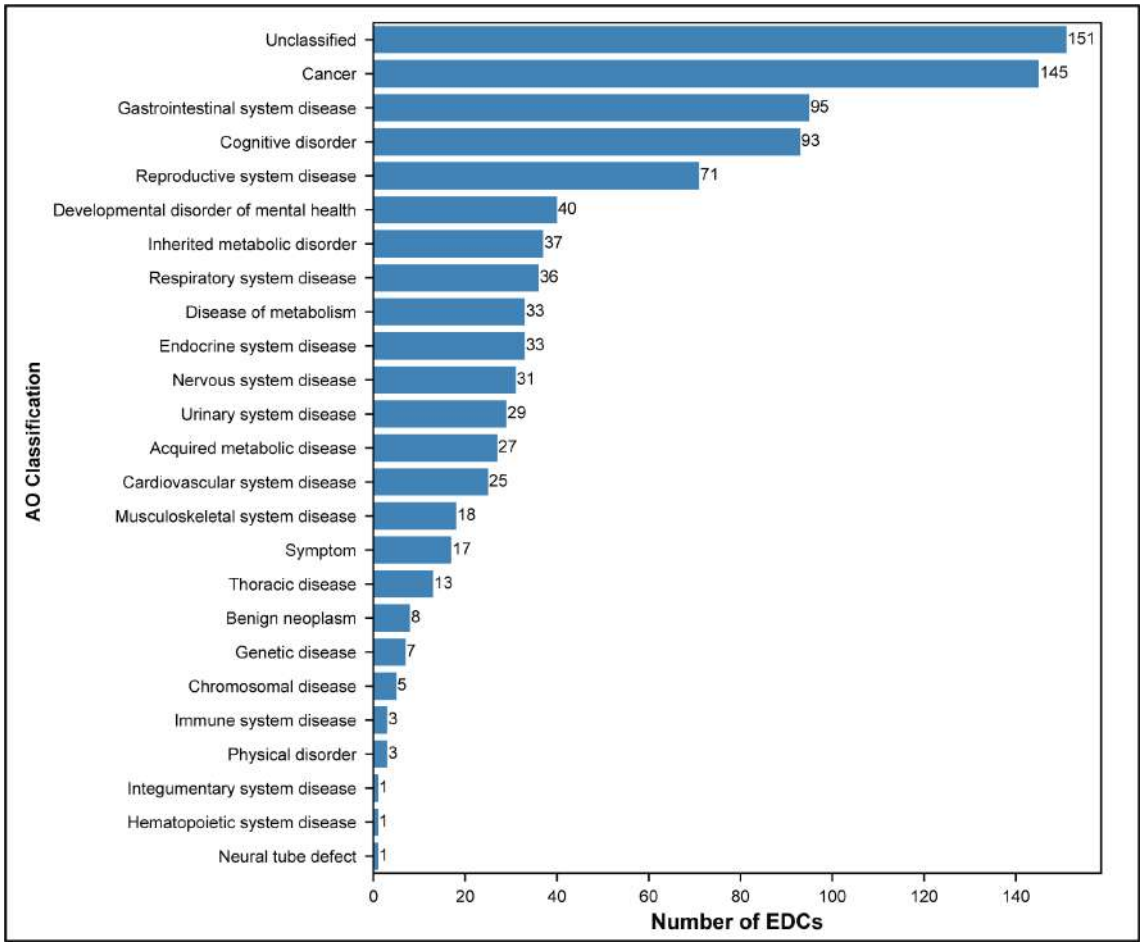


Figure 3.9: Distribution of 250 chemicals in DEDuCTv3.0 across 28 disease categories based on the AO classification of their associated AOPs at Level 5 relevance.

specific mechanisms (Supplementary Table S3.3). Among the 250 chemicals, we observed that 113 chemicals were linked to the KE 'Increase, Reactive oxygen species' (KE:1115), indicating that oxidative stress is a common mechanism through which EDCs exert their downstream adverse effects. Furthermore, we observed that 145 of the 250 chemicals were linked to 29 of the 215 AOPs that were categorized as 'Cancer' based on their AOs (Supplementary Table S3.7; Figure 3.9). This was followed by 95 chemicals linked to 17 AOPs that were categorized as 'Gastrointestinal system disease', and 93 chemicals linked to 7 AOPs that were categorized as 'Cognitive disorder' (Supplementary Table S3.7; Figure 3.9). This suggests that mechanisms underlying cancer, gastrointestinal disease, and cognitive disorders were most represented for chemicals in DEDuCTv3.0. Note that the current bias towards carcinogenic mechanisms may stem from the over representation

of cancer-associated AOPs within AOP-Wiki [56]. It should further be noted that the identified mechanisms reflect the broader toxic effects of EDCs and are not exclusive to endocrine-mediated pathways, as we derived the associations by integrating chemical-specific biological endpoints from multiple sources beyond the curated endocrine-mediated endpoints.

3.2.3 DEDuCT-KG: A toxicology knowledge graph on EDCs

A toxicology knowledge graph organizes chemical-relevant toxicological data from diverse sources into a network comprising chemical-associated toxicological data connected through semantic relationships [39]. Such a construction enables the querying of subgraphs that can highlight previously unexplored associations between entities, thereby facilitating novel understanding of chemical-induced toxicities [39]. Here, we systematically curated the 1043 potential EDCs and their 796 associated endocrine-mediated endpoints, and used them as the foundation to construct a large-scale toxicology knowledge graph, referred to as DEDuCT-KG (Figure 3.2). We further expanded DEDuCT-KG by integrating semantic relationships from several external sources, including ToxCast, CTD, and AOP-Wiki, to connect chemicals, endocrine-mediated endpoints, genes, phenotypes, diseases, KEs and AOPs into a unified network comprising 75106 nodes and 1065949 edges (Figure 3.2). Among these nodes, 48999 are genes, 3098 are diseases, 19238 are phenotypes, 1547 are KEs within AOP-Wiki, and 385 are the curated AOPs from AOP-Wiki. Based on the association information from different sources, we classified the edges into 31 distinct edge types connecting these nodes. Moreover, the LCC comprised 75048 nodes and 1065913 edges, which connected 1043 EDCs, 796 endpoints, 3096 diseases, 19219 phenotypes, 48979 genes, 1532 KEs, and 383 AOPs. This LCC signifies a vast corpus of information associated with the chemicals curated in DEDuCTv3.0. Moreover, we hosted DEDuCT-KG on a Neo4j graph database [180] and made it available on the webserver through an interactive user interface [181] to enable further research (Figure 3.10).

In a previous study on reproductive toxicants, it was shown that knowledge graph

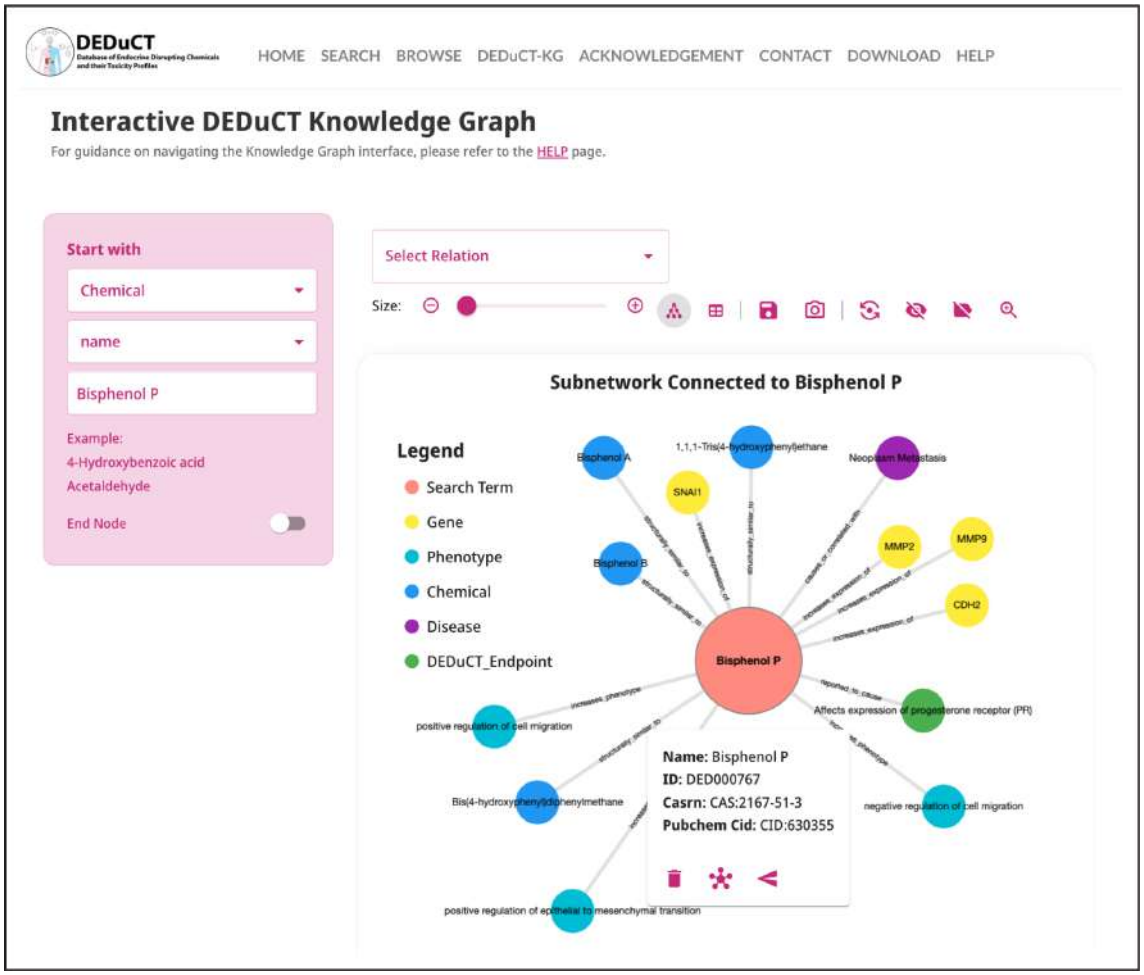


Figure 3.10: Screenshot of the interactive user interface of DEDuCT-KG in the webserver.

construction enabled the exploration of underexplored mechanisms of action (MOAs) associated with those chemicals [39]. Similarly, we utilized the chemicals and associated endocrine-mediated endpoint data from DEDuCTv3.0 to gain more information on EDC-associated MOAs. We observed that the endocrine-mediated endpoints were not directly associated with the genes to help gauge their endocrine-specific MOAs (Figure 3.2). Instead, the endocrine-mediated endpoints were linked to phenotypes and diseases, which were further linked to genes (Figure 3.2). Previously, the concept of a CGPD-tetramer was proposed as a bioinformatics-driven framework to bridge knowledge gaps between chemical exposures and outcomes by identifying intermediate molecular mechanisms [184]. Here, we adapted this framework by integrating endocrine-mediated endpoint-phenotype and endocrine-mediated endpoint-disease associations derived from the constructed knowl-

edge graph (Supplementary Table S3.10). This allowed for the construction of subgraphs such as chemical-gene-phenotype-endpoint and chemical-gene-disease-endpoint, which revealed indirect gene-to-endpoint links. These genes can then be used to explore the potential MoA associated with the chemical-induced endocrine-mediated effects [39, 184]. We demonstrate the utility of this knowledge graph using the following case studies.

Potential mechanism of action associated with obesogens in DEDuCTv3.0

Obesogens are a class of EDCs that promote obesity in exposed individuals by altering metabolic homeostasis and increasing susceptibility to weight gain [185]. We observed that DEDuCTv3.0 contained an endocrine-mediated endpoint titled 'Lead to obesity', which was categorized as 'Metabolic endocrine-mediated perturbations' (Supplementary Table S3.2). Therefore, we selected this endpoint to gain more insight into the potential obesogens within DEDuCTv3.0. In the constructed knowledge graph, we observed that this endpoint was associated with 52 chemicals, and diseases such as 'Obesity, Abdominal' (MESH:D056128), 'Weight Gain' (MESH:D015430), 'Obesity' (MESH:D009765), 'Overweight' (MESH:D050177), and 'Obesity, Morbid' (MESH:D009767), that were compiled from CTD. Subsequently, we selected the chemical-gene, chemical-disease, and gene-disease links from the knowledge graph to construct a chemical-gene-disease-endpoint subgraph comprising 18 chemicals, 260 genes, 4 diseases, and 1 endpoint (Supplementary Table S3.10). Importantly, a prior study by Heindel *et al.* [185] classified the chemical groups to which these 18 chemicals belong as obesogenic, thereby supporting the construction of our knowledge graph and highlighting its utility in extracting toxicity relevant information.

Among the 18 chemicals, we observed that bisphenol A (CAS:80-05-7) was linked to 256 out of 260 genes, followed by tetrachlorodibenzodioxin (CAS:1746-01-6) linked to 244 genes (Supplementary Table S3.11). Among the 260 genes, we observed that TNF (NCBI Gene ID: 7124) and AHR (NCBI Gene ID: 196) were linked to all 18 chemicals, followed by NFE2L2 (NCBI Gene ID: 4780), CD36 (NCBI Gene ID: 948) and ADIPOQ

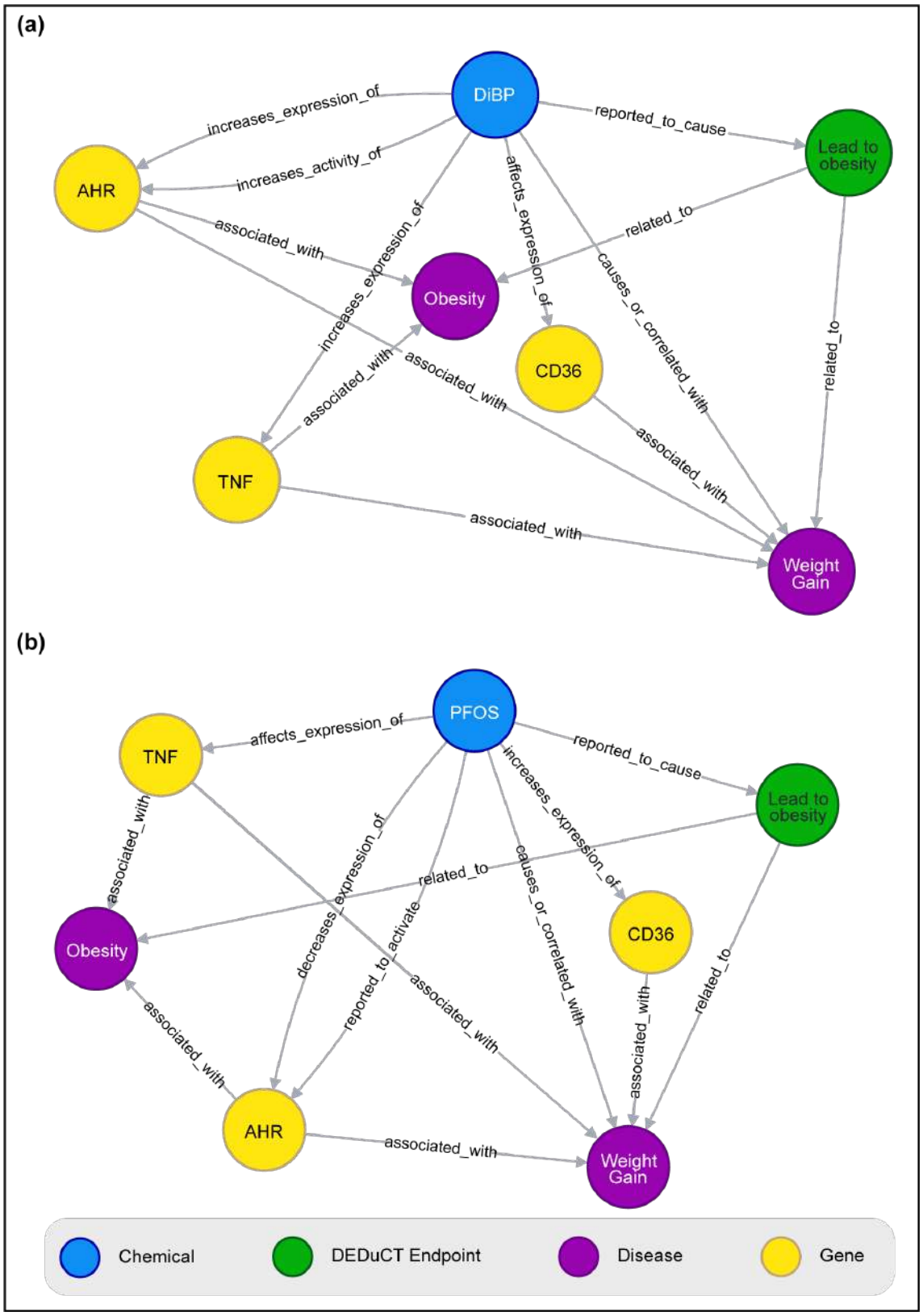


Figure 3.11: Chemical-gene-disease-endpoint subgraphs depicting the genes, diseases, and endocrine-mediated endpoint obtained from DEDuCT-KG, associated with obesogens: (a) dibutyl phthalate (DiBP; CAS:84-74-2); (b) perfluorooctane sulfonic acid (PFOS; CAS:1763-23-1). These subgraphs were generated using Neo4j.

(NCBI Gene ID: 9370) which were linked to at least 15 of the 18 chemicals (Supplementary Table S3.11). Alternatively, PHAROS [186] is an NIH-based repository that integrates data from the Target Central Resource Database (TCRD) to enable gene-disease association studies [187]. From this database, 370 genes were found to be associated with 'obesity disorder', including the highly connected TNF, AHR, CD36, and ADIPOQ genes. The overlap of genes between the knowledge graph and PHAROS underscores the biological relevance of these targets. For instance, the overexpression of the TNF gene and its role in inducing inflammation have been well-studied as central factors in the pathogenesis of obesity [188–190]. AHR gene is found to have a dual role in the pathogenesis of obesity [191], with its activation in adipocytes promoting their differentiation and eventually leading to the promotion of obesity [192]. The CD36 gene, which functions as a fatty acid translocase, showed increased expression in preadipocytes of obese mice [193], and positively correlated with abdominal adipose tissue distribution in humans [194]. Finally, the dysregulation of the ADIPOQ gene has been extensively studied in the pathogenesis of metabolic syndromes, including obesity [195, 196].

Next, we selected chemical-gene-disease-endpoint subgraphs for two obesogens, namely, dibutyl phthalate (DiBP; CAS:84-74-2) and perfluorooctane sulfonic acid (PFOS; CAS:1763-23-1), to further explore their MOAs (Figure 3.11). From these subgraphs, we observed that these two chemicals are associated with TNF, AHR, and CD36 genes (Figure 3.11). In particular, we observed that these chemicals increase expression of TNF and CD36 genes based on evidence curated from CTD (Figure 3.11). Additionally, we observed that DiBP and PFOS activate AHR, based on evidence from CTD and ToxCast invitrodb v4.3, respectively (Figure 3.11). These observations highlight that the two chemicals can potentially induce obesity through interaction with TNF, AHR, and CD36.

Potential mechanism of action underlying EDC-associated neurodegenerative disorders

Exposure to EDCs has been associated with a wide range of neurological manifestations, with growing evidence implicating their role in the development and progression of neurodegenerative diseases [197]. We observed that DEDuCTv3.0 contained an endpoint titled 'Neuronal degeneration', which was categorized as 'Neurological endocrine-mediated perturbations' (Supplementary Table S3.2). Therefore, DEDuCTv3.0 was utilized to investigate the potential mechanisms through which EDCs may contribute to the development of neurodegenerative disorders. From the constructed knowledge graph, we observed that the endpoint was associated with 13 chemicals and two diseases, namely, 'Nerve Degeneration' (MESH:D009410) and 'Neurodegenerative diseases' (MESH:D019636), that were compiled from CTD. Similar to the case study on obesogens, we constructed a chemical-gene-disease-endpoint subgraph, which comprised 6 chemicals, 83 genes, 2 diseases, and 1 endpoint (Supplementary Table S3.12). Given that Alzheimer's disease and Parkinson's disease are the two well-known neurodegenerative disorders for which environmental pollutants have been identified as risk factors [198], we selected them as the focus of this case study.

Among the six chemicals in the subgraph, we observed that bisphenol A (CAS:80-05-7) was linked to all 83 genes (Supplementary Table S3.13). Among the 83 genes, we observed that TH (NCBI Gene ID:7054), CASP3 (NCBI Gene ID: 836), MAPK1 (NCBI Gene ID: 5594), PINK1 (NCBI Gene ID: 65018), SLC18A2 (NCBI Gene ID: 6571), and PRKN (NCBI Gene ID: 5071) were associated with all six chemicals (Supplementary Table S3.13). Among the other chemicals present in the subgraph, chlorpyrifos (CPF; CAS:2921-88-2) is an organophosphate insecticide, and atrazine (ATZ; CAS:1912-24-9) is a triazine herbicide. Based on the data curated within the DEDuCT webserver, we observed that these chemicals are actively used, produced in high volumes, and have been flagged as chemicals of concern across different jurisdictions. Moreover, it was observed

that these chemicals have been detected in different human biospecimens. Therefore, to identify the potential MoA associated with the EDCs in these diseases, we obtained the corresponding gene sets from PHAROS [186], and constructed subgraphs for both CPF and ATZ.

Alzheimer's disease was associated with 482 genes, among which the highly connected CASP3 and PINK1 genes were identified. The CASP3 gene, a critical regulator of neuronal apoptosis, has been observed at elevated levels in Alzheimer's disease, where it is associated with neuronal loss and the conversion of Tau and amyloid-β proteins into more toxic derivatives [199]. The PINK1 gene, a regulator of mitochondrial quality and synaptic integrity, has been observed at reduced levels in Alzheimer's disease, where it is associated with amyloid-β accumulation, mitochondrial dysfunction, and cognitive decline [200]. From the constructed subgraphs, it was observed that both CPF and ATZ increase the expression of or activate the CASP3 gene with evidence from CTD and ToxCast invitrodb v4.3 (Figure 3.12). While CPF was observed to decrease the expression of the PINK1 gene, ATZ was observed to increase its expression (Figure 3.12). This suggests that CPF can potentially induce Alzheimer's disease through both CASP3 activation and PINK1 downregulation, while ATZ can potentially induce Alzheimer's disease through CASP3 activation alone.

From the knowledge graph, we observed that Parkinson's disease was associated with 248 genes, among which the highly connected TH, SLC18A2, PINK1, and PRKN genes were present. The TH gene, which encodes the tyrosine hydroxylase enzyme involved in brain catecholamine biosynthesis, has been observed at reduced expression levels in Parkinson's disease, where it is associated with diminished dopamine synthesis [201]. The SLC18A2 gene, responsible for packaging dopamine into synaptic vesicles, has been observed at reduced levels in Parkinson's disease, where it is associated with impaired dopamine storage and loss of dopaminergic neurons [202]. The PRKN and PINK1 genes, both of which regulate mitochondrial quality and mitophagy in neurons, have been observed at reduced levels in Parkinson's disease, where their loss is associated with the accumulation

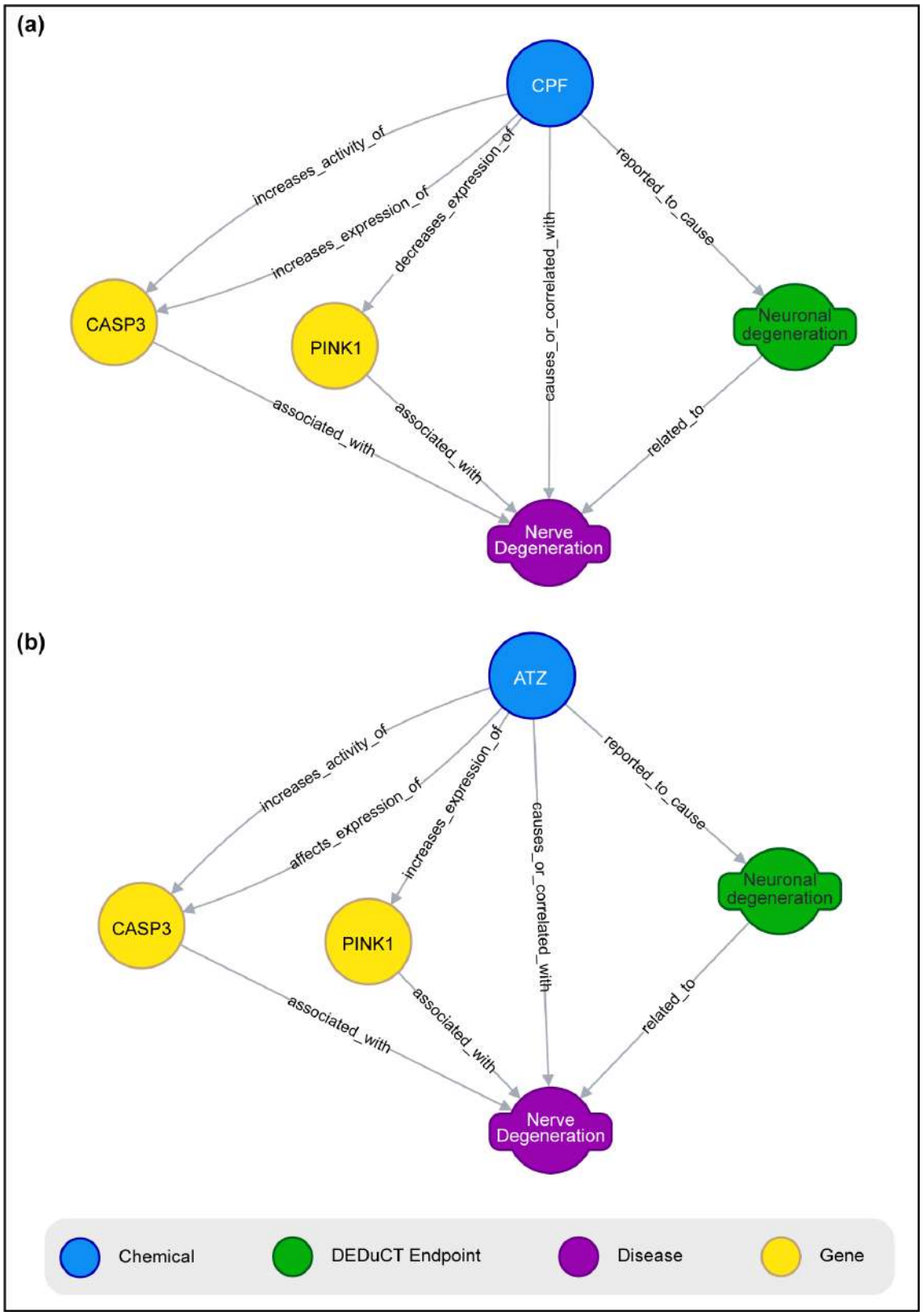


Figure 3.12: Chemical-gene-disease-endpoint subgraphs depicting the genes, diseases, and endocrine-mediated endpoint obtained from DEDuCT-KG, associated with Alzheimer's disease: (a) chlorpyrifos (CPF; CAS:2921-88-2); (b) atrazine (ATZ; CAS:1912-24-9). These subgraphs were generated using Neo4j.

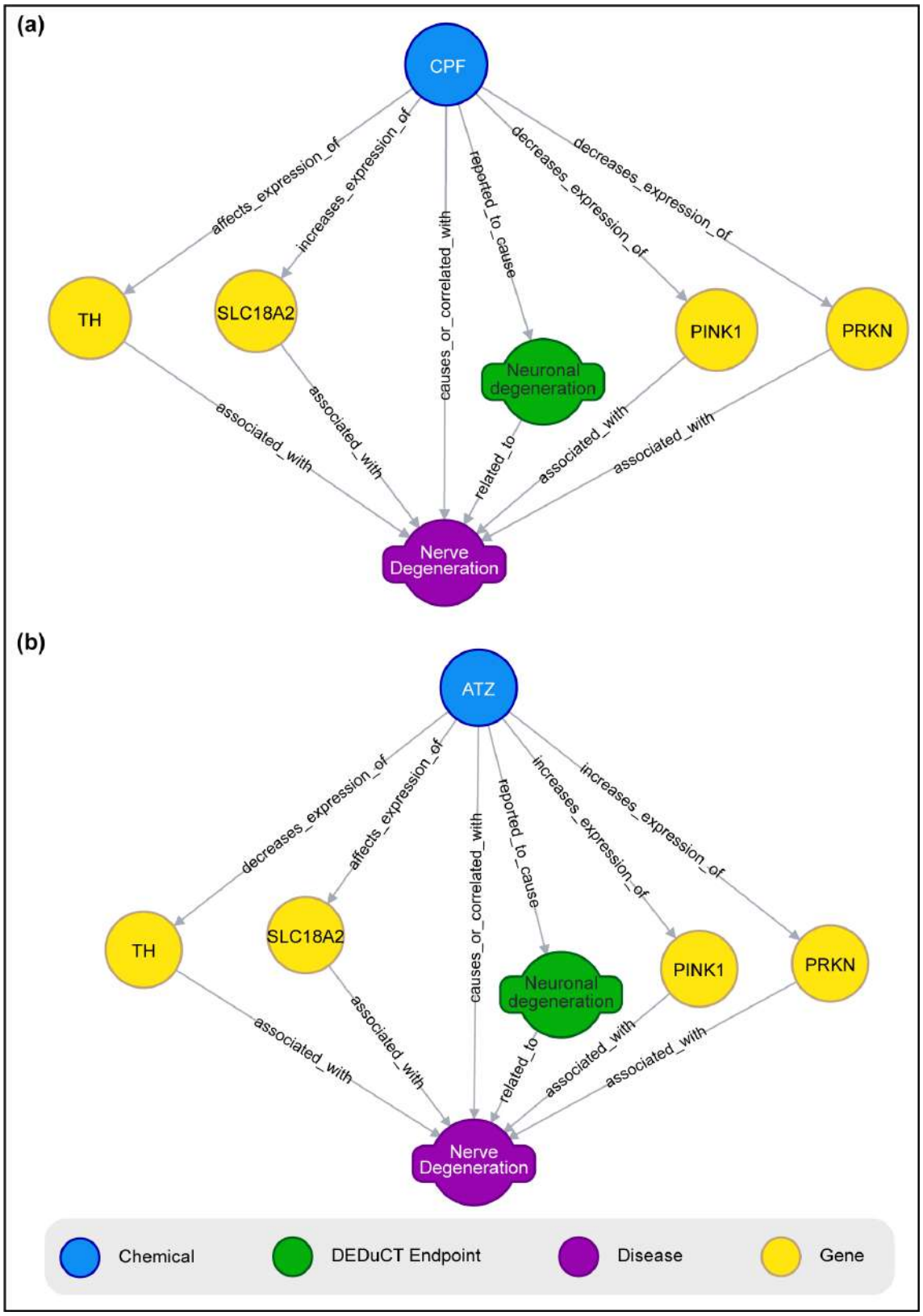


Figure 3.13: Chemical-gene-disease-endpoint subgraphs depicting the genes, diseases, and endocrine-mediated endpoint obtained from DEDuCT-KG, associated with Parkinson's disease: (a) chlorpyrifos (CPF; CAS:2921-88-2); (b) atrazine (ATZ; CAS:1912-24-9). These subgraphs were generated using Neo4j.

of damaged mitochondria and loss of dopaminergic neurons [203]. Notably, mutations in the PRKN gene have been strongly associated with hereditary forms of Parkinson's disease, making it one of the key genes implicated in familial Parkinsonism [204]. From the constructed subgraph for CPF, we observed that it decreases the expression of PRKN and PINK1, and affects the expression of TH (Figure 3.13a). From the subgraph for ATZ, we observed that it decreases the expression of TH and affects SLC18A2 expression (Figure 3.13b). This suggests that CPF potentially triggers Parkinson's disease through mitochondrial dysfunction in the neurons, while ATZ potentially triggers Parkinson's disease through a reduction in dopaminergic neurons.

Overall, through these case studies, we were able to utilize the knowledge graph to gain biological insights into the potential MoA underlying EDC-associated obesity and neurodegeneration, which were otherwise not apparent within DEDuCT. This demonstrates the utility of the knowledge graph in organizing the knowledge within the database and gaining biological insights into the potential MoA underlying endocrine disruption.

3.2.4 Potential exposure to EDCs through consumer products

Daily, humans encounter a wide variety of products, some of which may contain potential EDCs that pose a risk to human health. Therefore, we used CPDatv4 [163] to explore the exposure sources of potential EDCs in DEDuCTv3.0. First, we observed that CPDatv4 cataloged 542 of 1043 potential EDCs, with corresponding products grouped into product use categories (PUCs) (Supplementary Table S3.1). We observed that CPDatv4 utilizes a hierarchical classification of PUCs, where 'Kind' represents the highest level describing the physical nature of the product [163], spanning three types: (i) 'Article' denoting products designed for long term use in the environment; (ii) 'Formulation' denoting products kept in containers; (iii) 'Occupation' denoting products meant for use in work environments [163].

Among the 542 chemicals present in CPDatv4, we observed that 171 were present in products categorized as 'Articles'. Among these 171 chemicals, we observed that 133

chemicals were present in products categorized as 'Construction and building materials', followed by 81 chemicals present in products categorized as 'Batteries' (Supplementary Table S3.1). This potentially suggests that researchers must check such chemicals for chronic or low dose exposures. Similarly, we identified 334 of 542 chemicals as present in products categorized as 'Formulation' (Supplementary Table S3.1). Among these, 168 chemicals were present in products categorized as 'Personal care', followed by 138 in 'Home maintenance' (Supplementary Table S3.1). This highlights a significant pathway of exposure to potential EDCs through common consumer products. Finally, we observed that 418 of 542 chemicals were present in products categorized as 'Occupational' (Supplementary Table S3.1). Among these, 344 were present in products categorized as 'Laboratory supplies', followed by 198 chemicals as 'Raw materials' (Supplementary Table S3.1).

In addition to PUC data, CPDatv4 provides the functional uses associated with the chemicals in consumer products [163]. This information is valuable for identifying less toxic chemical alternatives to replace the chemical of potential concern. To refine this dataset, we filtered out chemicals with a reported function of 'unknown' or 'not reported'. Through this process, we identified 473 of 1043 potential EDCs with known functional use data (Supplementary Table S3.1). Among these, we linked 417 chemicals to 73 standardized OECD functional terms (Supplementary Table S3.1) [163]. Among these, we observed 179 chemicals function as 'Biocides', followed by 92 chemicals that function as 'Fragrances', in their corresponding products.

3.2.5 Distribution of potential EDCs in chemical lists

EDCs are of prime concern and regulatory agencies worldwide increasingly regulate them. To check the regulatory coverage of the potential EDCs compiled in this chapter, we utilized several chemical lists. First, we compiled a unique list of 2049 potential EDCs from DEDuCTv3.0, the WHO report [160], the TEDx [161], and EDCs Databank version 2015 [162]. Next, we compiled 54 different chemical lists and categorized them as substances in use and substances of concern (Supplementary Table S3.8). Based on the

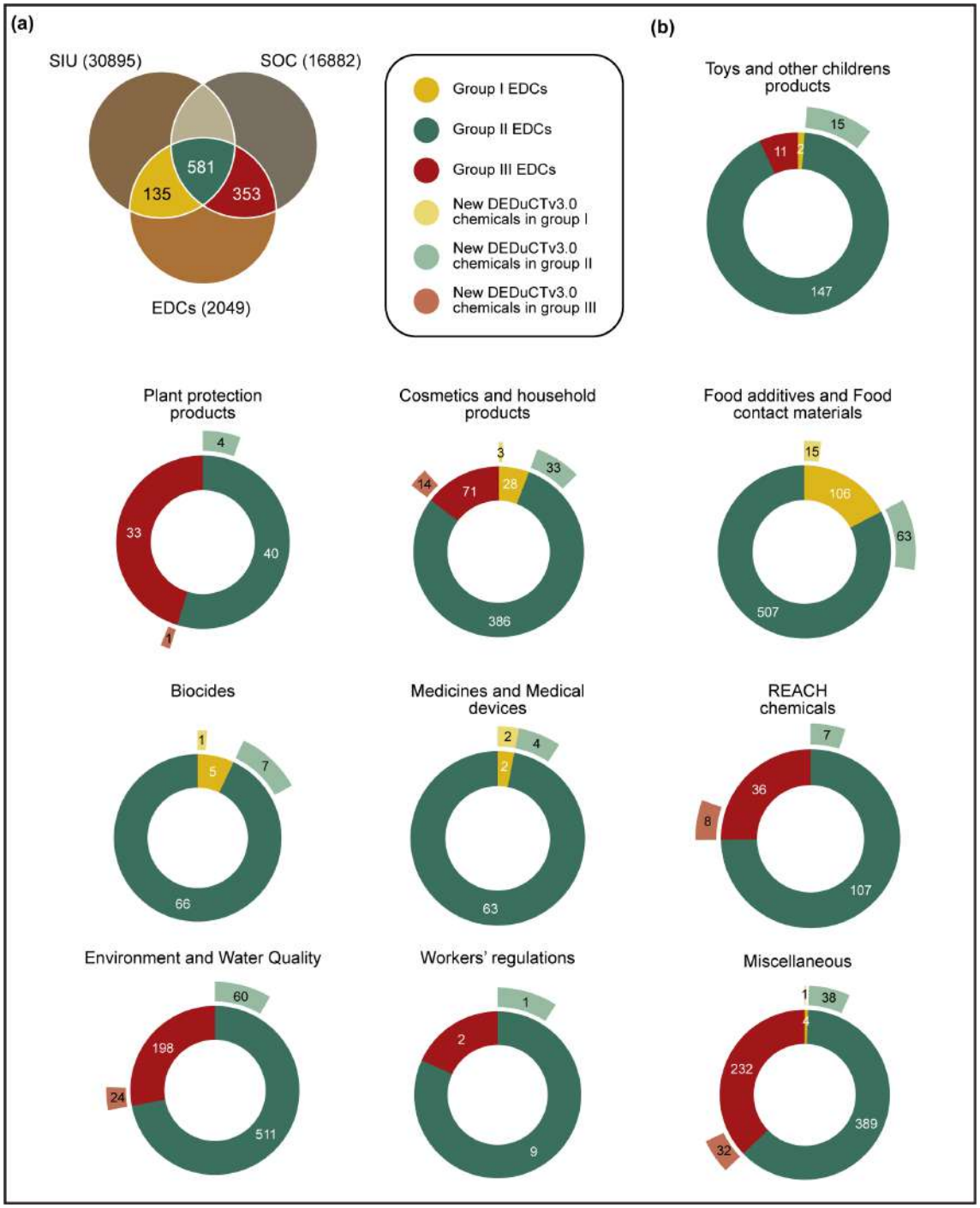


Figure 3.14: Distribution of 2049 unique potential EDCs from DEDuCTv3.0, TEDx, EDCs Databank, and WHO report across 54 chemical lists. (a) Venn diagram depicting chemicals categorized into groups I, II, and III. (b) Sunburst plots depicting the distribution of potential EDCs across 10 chemical list categories. The inner ring represents the number of potential EDCs categorized into groups I, II, or III within each category, while the outer ring highlights the number of newly compiled chemicals in DEDuCTv3.0 across each group.

procedure detailed in section 3.1.6, we categorized the chemicals within each list into three groups, with 19501 chemicals in group I, 11394 in group II, and 5488 in group III. We overlapped this data with 2049 potential EDCs and identified that 135 were in group I, 581 were in group II, and 353 were in group III (Figure 3.14a). Among these, we observed that 20 are newly compiled within DEDuCTv3.0 in group I, 67 are newly compiled in group II, and 50 are newly compiled in group III (Figure 3.14a).

Next, we examined the distribution of these three chemical groups across the different categories of chemical lists. We observed that the newly compiled chemicals were present in all groups, and predominantly belonged to group II (Figure 3.14b). Notably, we observed that 15 of the new group I chemicals were currently in use as food additives or food contact materials, highlighting potential concern about dietary exposure. Among these 15 chemicals, 11 contained information from *in vitro* human cell line studies on evidence of endocrine-mediated effects (Supplementary Table S3.1). Furthermore, we observed that three of the new chemicals were present in cosmetics, with two chemicals having *in vitro* human evidence of endocrine-mediated effects (Supplementary Table S3.1). We also observed that none of these new chemicals were cataloged as high production volume chemicals by either the USEPA or OECD. Overall, these findings highlight that some of the newly identified chemicals in DEDuCTv3.0 are currently in active use but remain unflagged by major regulatory agencies, demonstrating the utility of this compilation in uncovering previously overlooked chemicals of concern.

3.2.6 Enhancements to DEDuCT webserver

We have substantially updated the DEDuCT webserver, accessible at `https://cb.imsc.res.in/deduct/`, to align more closely with the FAIR principles [16]. Notably, we have provided unique identifiers for the chemicals in DEDuCT, along with links to their respective chemical pages in databases such as PubChem [11] and CAS Common Chemistry [93], and the USEPA's CompTox Chemicals Dashboard [21, 22] (Figure 3.15; Supplementary Table S3.1). In this update, we have also enriched the chemical pages with more

toxicology-relevant data curated from multiple relevant databases. For instance, we curated information regarding chemical-associated genes, phenotypes, and diseases from the CTD [151] and made them available as separate sections within the chemical pages (Figure 3.16). Additionally, we curated the latest data from ToxCast invitrodb v4.3 [20] and made them available in respective chemical pages for improved access (Figure 3.17). Here, we further categorized the data based on the type of target, such as RNA, protein, molecular messenger, pathway, cellular, and hormone. Moreover, we curated the PUC data and the functional uses of chemicals in consumer products from CPDatv4 [163] and made them accessible through a dedicated section in the chemical page (Figure 3.18). We categorized the functional uses under specific vocabularies from the USEPA and OECD (Figure 3.18), while the PUCs were categorized based on the physical nature of the corresponding product (Figure 3.18). We updated the regulatory lists to include the latest chemical inventories. Chemicals are searchable based on their identifiers, names, synonyms, physicochemical properties, or structures (Figure 3.19).

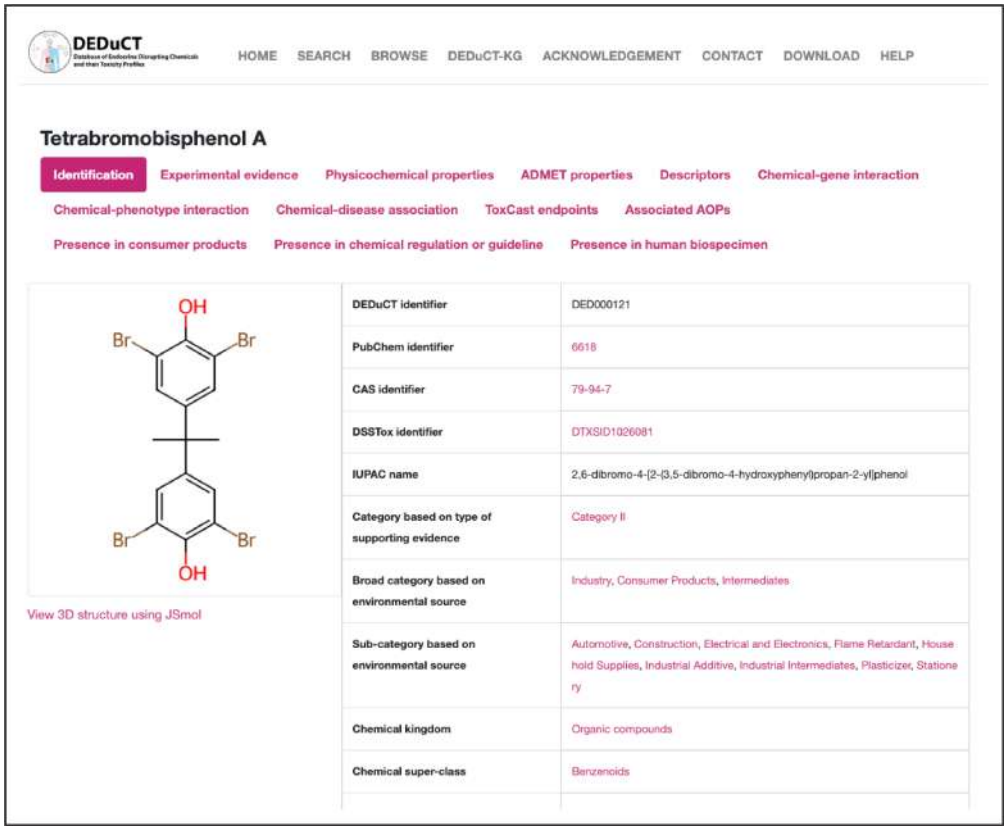


Figure 3.15: Screenshot of the updated chemical information page in DEDuCT webserver.

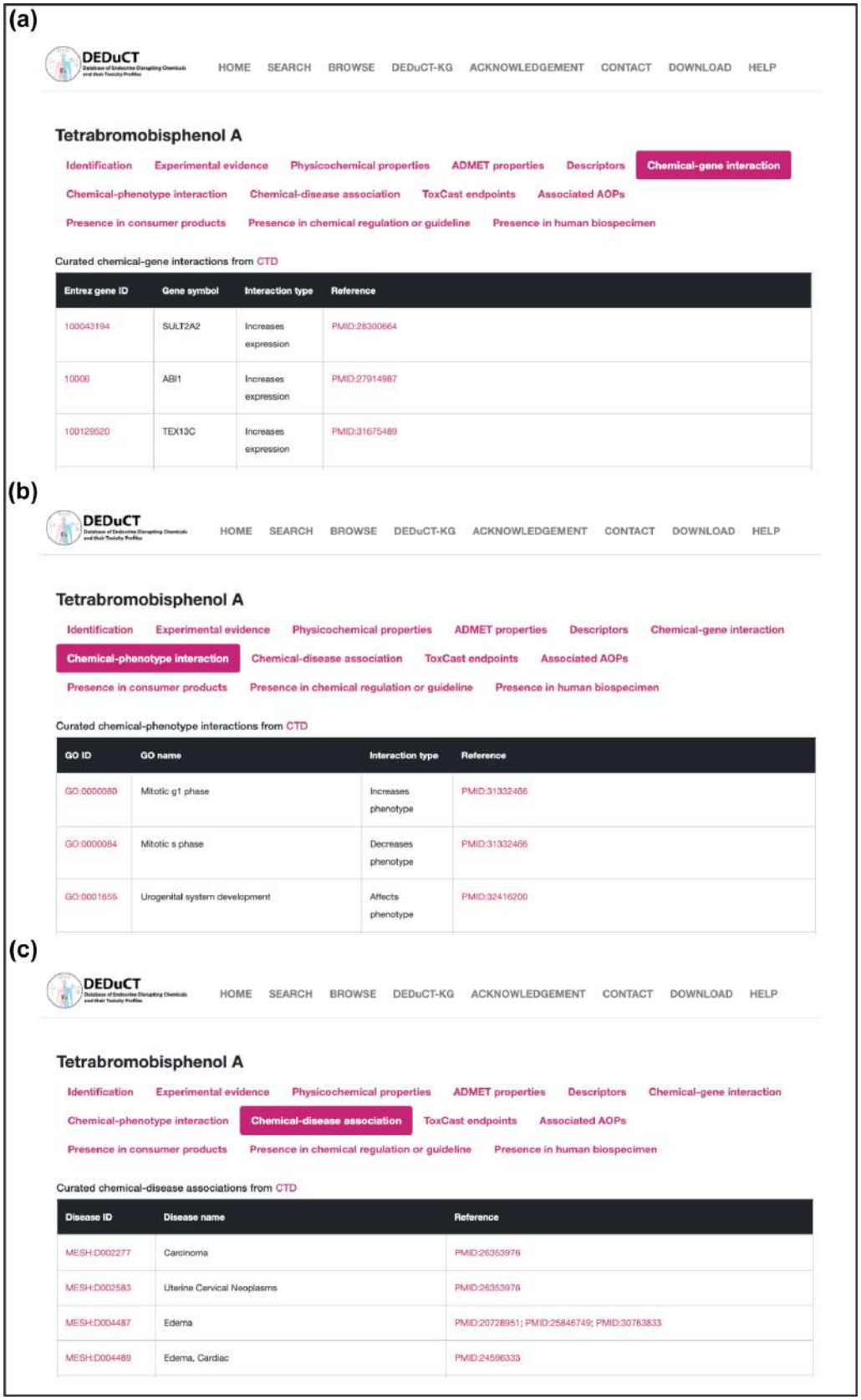


Figure 3.16: Screenshots from DEDuCT webserver of the updated chemical information pages on: (a) chemical-gene interactions; (b) chemical-phenotype interactions; (c) chemical-disease associations.

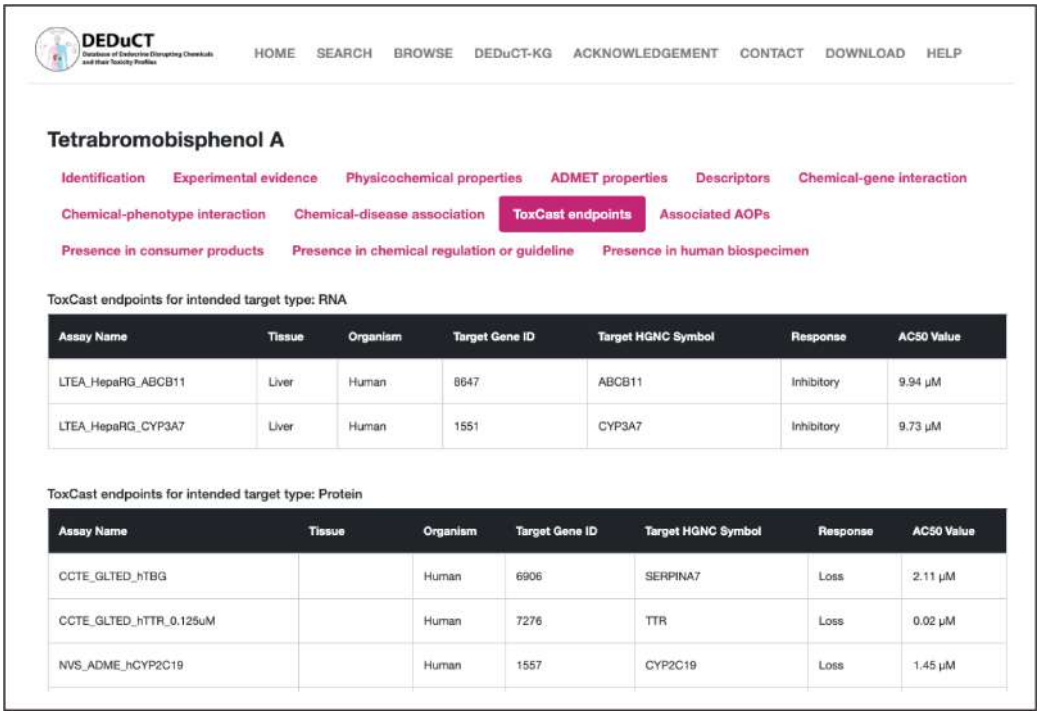


Figure 3.17: Screenshot from DEDuCT webserver of the updated chemical information page on associated ToxCast endpoints.

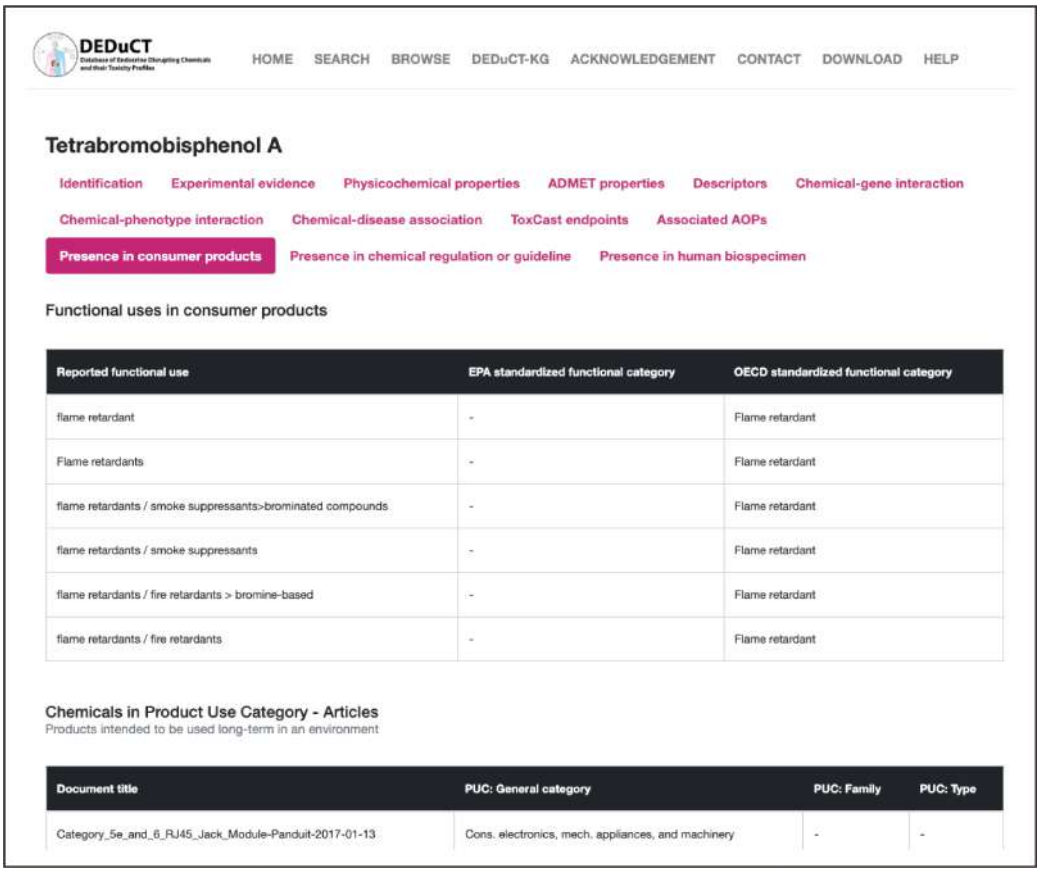


Figure 3.18: Screenshot from DEDuCT webserver of the updated chemical information page on associated data from the Chemical and Products Database (CPDat).

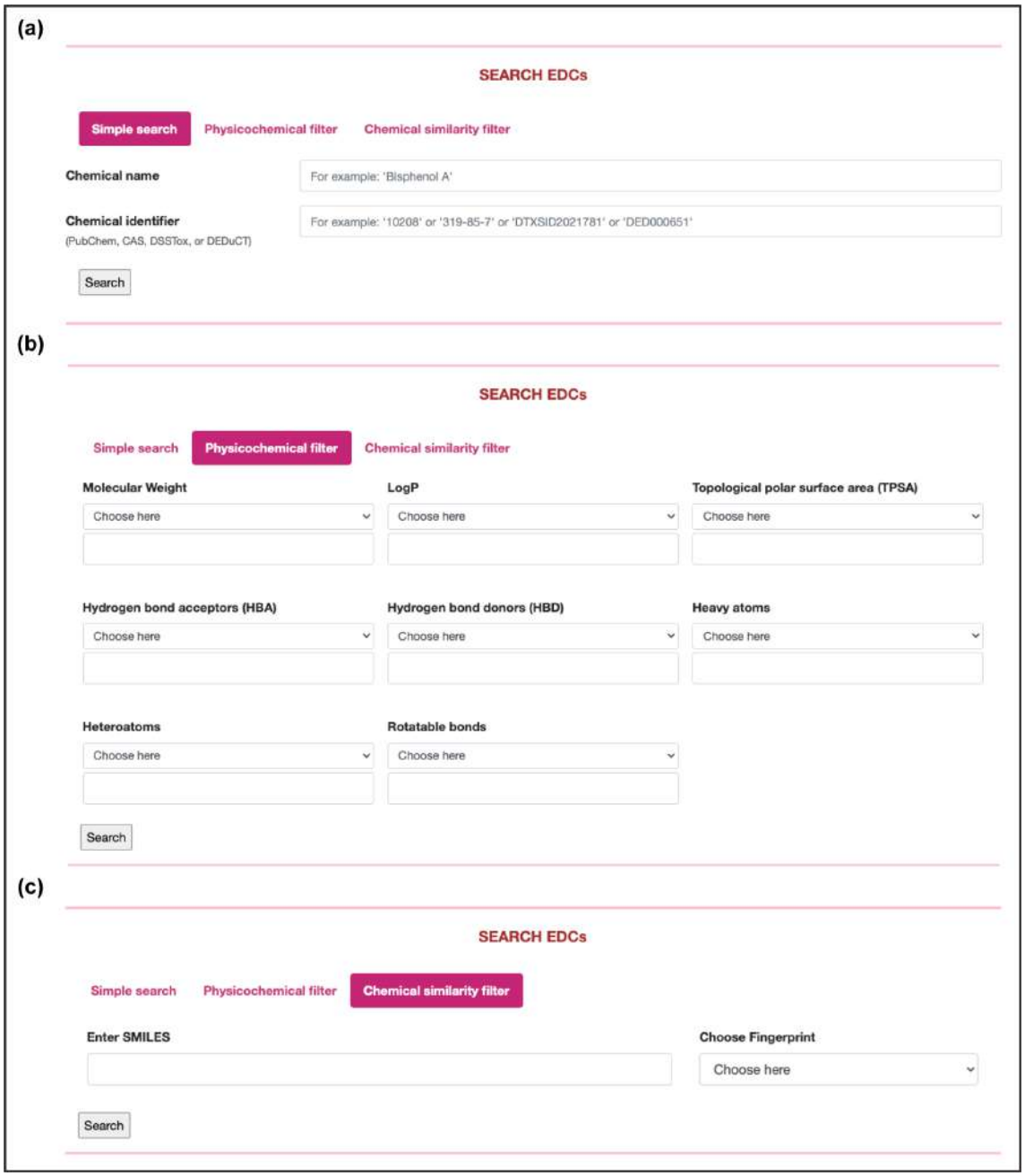


Figure 3.19: Screenshots from DEDuCT webserver on the updated search options. (a) Search for chemicals using chemical information such as names, synonyms, PubChem identifier, CASRN, DSSTox identifier, and DEDuCT identifier. (b) Search for chemicals using their physicochemical properties. (c) Search for chemicals based on structural similarity.

Additionally, we constructed a large-scale toxicology knowledge graph for EDCs, termed DEDuCT-KG, and made it accessible through an interactive user interface [181] on the webserver. Overall, DEDuCTv3.0 is a culmination of significant updates to the resource, making it more FAIR-compliant and enhancing the toxicological knowledge associated with the compiled EDCs. We have made the data compiled in DEDuCTv3.0 accessible through machine-readable file formats in the Downloads section of the webserver, enabling further research on EDCs.

3.3 Discussion

The Database of Endocrine Disrupting Chemicals and their Toxicity Profiles (DEDuCT), accessible at https://cb.imsc.res.in/deduct/, is a manually curated and an actively maintained database on potential EDCs that has organized and structured data from scientific literature. In this update, we systematically collected EDC-specific toxicity data from toxicology-relevant resources such as CTD, ToxCast, and AOP-Wiki, and made it available on individual chemical pages, providing easy single-window access to all EDC-related toxicity information. Additionally, we organized the compiled data to build a large-scale EDC-specific toxicology knowledge graph, called DEDuCT-KG, and made it accessible through an interactive interface. The construction of DEDuCT-KG enabled us to explore possible mechanisms of action underlying the curated endocrine-mediated endpoints. Furthermore, we identified potential exposure sources using the USEPA's CPDatv4, showing exposure through everyday products such as cosmetics and home maintenance products, among others. Comparison with existing regulations also revealed chemicals that are currently overlooked. Overall, these analyses show that integrating diverse toxicological information has improved our understanding of the health effects associated with these potential EDCs and supports further research in this field.

While DEDuCT is a comprehensive, manually curated resource on EDCs, it only focuses on compiling data from published scientific literature and does not include studies

conducted by regulatory bodies or industries, which are typically published as chemical reports. Moreover, it primarily focuses on mammalian toxicities and does not address the ecotoxicity of such chemicals. It also does not cover the mixture effects of potential EDCs. Although the latest chemical lists were used for the regulatory analysis in this chapter, they might not fully capture all possible regulated chemicals. Moreover, the knowledge graph was inherently limited by the scope of its underlying data sources. CTD predominantly captured literature-curated associations, which may over-represent well-studied chemicals and diseases, while ToxCast data derived from *in vitro* assays may not fully capture *in vivo* biological complexity. Although AOP-Wiki is a continuously updated resource, a snapshot of the resource was utilized during construction of the updated webserver, and may not capture all recent additions. As we relied on curated information from these limited sources, the absence of a link may reflect limited data within those sources.

Nonetheless, to the best of our knowledge, we present the first toxicology knowledge graph specifically developed for EDCs, termed DEDuCT-KG, enabling large-scale exploration of their adverse effects. DEDuCT-KG can facilitate the identification of mechanisms of action, and when supported with external validation, it can reveal underexplored toxic effects associated with these chemicals. The new interactive user interface of DEDuCT-KG supports EDC research through specialized knowledge graph techniques. Moreover, the curated EDC-AOP associations will improve the understanding of chemical toxicities and help in designing targeted *in vitro* testing strategies for EDCs. Consumer product information from CPDat has allowed for the identification of exposure routes for several potential EDCs and highlighted primary sources of concern. In sum, this update of DEDuCT organizes the large amount of information on EDCs generated since the last version. It brings together diverse toxicology data into a central resource that can accelerate the research and regulation of EDCs.

Supplementary Information

Supplementary Tables S3.1-S3.13 associated with this chapter are available for download from the GitHub repository: https://github.com/asamallab/PhDThesis-Nikhil_C/blob/main/SI/ST_Chapter3.xlsx.

Chapter 4

Network toxicology focused investigation on the impacts of inorganic arsenic and cadmium on human and ecosystem health in India

Heavy metals like arsenic and cadmium are categorized as priority pollutants by the United States Agency for Toxic Substances and Disease Registry (ATSDR), based on their frequency of occurrence, associated toxicity, and potential for human exposure [205]. In particular, inorganic forms of arsenic and cadmium are hazardous due to their higher rates of absorption, toxic valence states, and enhanced ability to cause cellular and organ-level damage [206–209]. Moreover, they are categorized as endocrine disrupting chemicals in DEDuCT with both *in vitro* human and *in vivo* rodent evidence. Therefore, these heavy metals need to be regulated and controlled to mitigate their adverse effects on human and ecosystem health.

Across India, these metals are widely reported in groundwater, agricultural soils, and food crops. Arsenic contamination is most evident in the Ganga–Brahmaputra–Meghna

plains [210–213], while cadmium is frequently associated with industrial and agricultural regions across India [214–216]. Despite their widespread occurrence and multi-species impacts, regulatory standards in India are largely limited to drinking water guidelines, with little attention to ecological toxicity. Thus, organizing disparate data within network toxicology frameworks can provide a basis to support risk assessment, and aid in policy decision-making in the Indian context.

In Chapter 3, we utilized a network toxicology framework called Adverse Outcome Pathways (AOPs) to investigate the mammalian-centric toxicity mechanisms induced by endocrine disrupting chemicals (EDCs). Specifically, we curated data relevant to humans and rodents from five exposome-relevant sources to focus on the adverse effects associated with EDCs. In this chapter, we expand the scope by including ecotoxicologically-relevant data from ECOTOX Knowledgebase [63, 217], and integrate other network toxicology-based frameworks to gain a comprehensive understanding of the impacts of inorganic arsenic and cadmium from a One Health perspective. Notably, we construct the first large-scale Aggregate Exposure Pathway (AEP) networks for inorganic arsenic and cadmium in India. Next, we integrate these with the AOPs to explore the complete source-to-outcome pathways. We further construct stressor-species networks, and integrate these with the derived aquatic hazard concentrations to provide a novel approach for visualizing particularly vulnerable species. Finally, we conduct environmental risk assessments for these chemicals across Indian rivers. **The work reported in this chapter is contained in the published manuscript [75].**

4.1 Methods

4.1.1 Compilation and curation of inorganic compounds of arsenic and cadmium

In this chapter, we demonstrate the utility of network toxicology-based analysis to gain a holistic understanding of inorganic arsenic- and cadmium-induced toxicities from a One Health perspective. We first extracted toxicologically-relevant biological endpoint data associated with arsenic, cadmium and their compounds, from six exposome-relevant databases, namely ToxCast [18], Comparative Toxicogenomics Database (CTD) [25, 151], DEDuCT [105, 120, 121], NeurotoxKb [126, 127], AOP-Wiki [50], and ECOTOX [63, 217]. Next, we curated chemical information and mapped the compounds to standardized PubChem [11] and Chemical Abstracts Service (CAS) [93] identifiers, to ensure a non-redundant chemical list. Subsequently, we retrieved two-dimensional (2D) structural data from PubChem and used Open Babel [106] to generate key structural representations, including SMILES, InChI, and InChIKey. We used ClassyFire [96, 97] to classify these chemicals as either inorganic or organic. For the compounds that could not be classified by ClassyFire, we manually verified and classified them based on their structural definitions [96]. Through this extensive systematic workflow, we identified 62 inorganic arsenic compounds and 18 inorganic cadmium compounds (Supplementary Table S4.1). We subsequently used these compounds for all further analysis in this chapter.

4.1.2 Construction of aggregate exposure pathways (AEPs) for inorganic arsenic and cadmium in India

The Aggregate Exposure Pathway (AEP) framework traces how a chemical moves from its source to the biological site of action [42–45]. It is represented as a directed graph, where nodes are key exposure states (KESs) and links are key transitional relationships (KTRs) [42–45]. In this chapter, we present the first AEP networks constructed for

inorganic arsenic and cadmium in the Indian context, to characterize their distribution and exposure risks.

First, we conducted a PubMed search to retrieve relevant articles detailing the sources of arsenic and cadmium, and their environmental dispersion. We ran the following queries on 6 August 2025, on PubMed [91]: '"arsenic*"[TIAB] AND "India*"[TIAB]' and '"cadmium*"[TIAB] AND "India*"[TIAB]'. These queries resulted in 1095 arsenic-relevant articles and 594 cadmium-relevant articles. Next, we manually screened these articles based on their titles and abstracts, and excluded the article if: (i) it is a review article; (ii) it is written in a language other than English; (iii) it is not specific to India; (iv) it did not contain information on presence of arsenic or cadmium in environmental media or biological samples (Figure 4.1). Consequently, we screened the full-text of the articles that passed the previous stage and excluded an article if: (i) the full text was not available; (ii) the arsenic or cadmium concentrations were not experimentally detected (Figure 4.1). Through this extensive manual process, we identified 177 arsenic-relevant and 182 cadmium-relevant articles (Figure 4.1). From these articles, we extracted data on the sampling location, year, and detection method. We also recorded the chemical type and its concentrations in environmental media or biological samples (Supplementary Table S4.2). Additionally, we curated the information from these articles regarding the flow of detected chemicals across different media (Supplementary Table S4.3). We then identified the environmental and biological states, and organized the corresponding links to construct a conceptual AEP network for arsenic and cadmium. This network provides a holistic view of their distribution and exposure based on India-specific data.

4.1.3 Compilation and curation of AOPs within AOP-Wiki

The AOP framework organizes the biological events (key events (KEs)) underlying a stressor-induced perturbation into a directional network, beginning with a molecular interaction (molecular initiating event (MIE)) and culminating in an AO relevant for regulation [47–49, 167]. In this directed network, KEs function as vertices, with the MIE

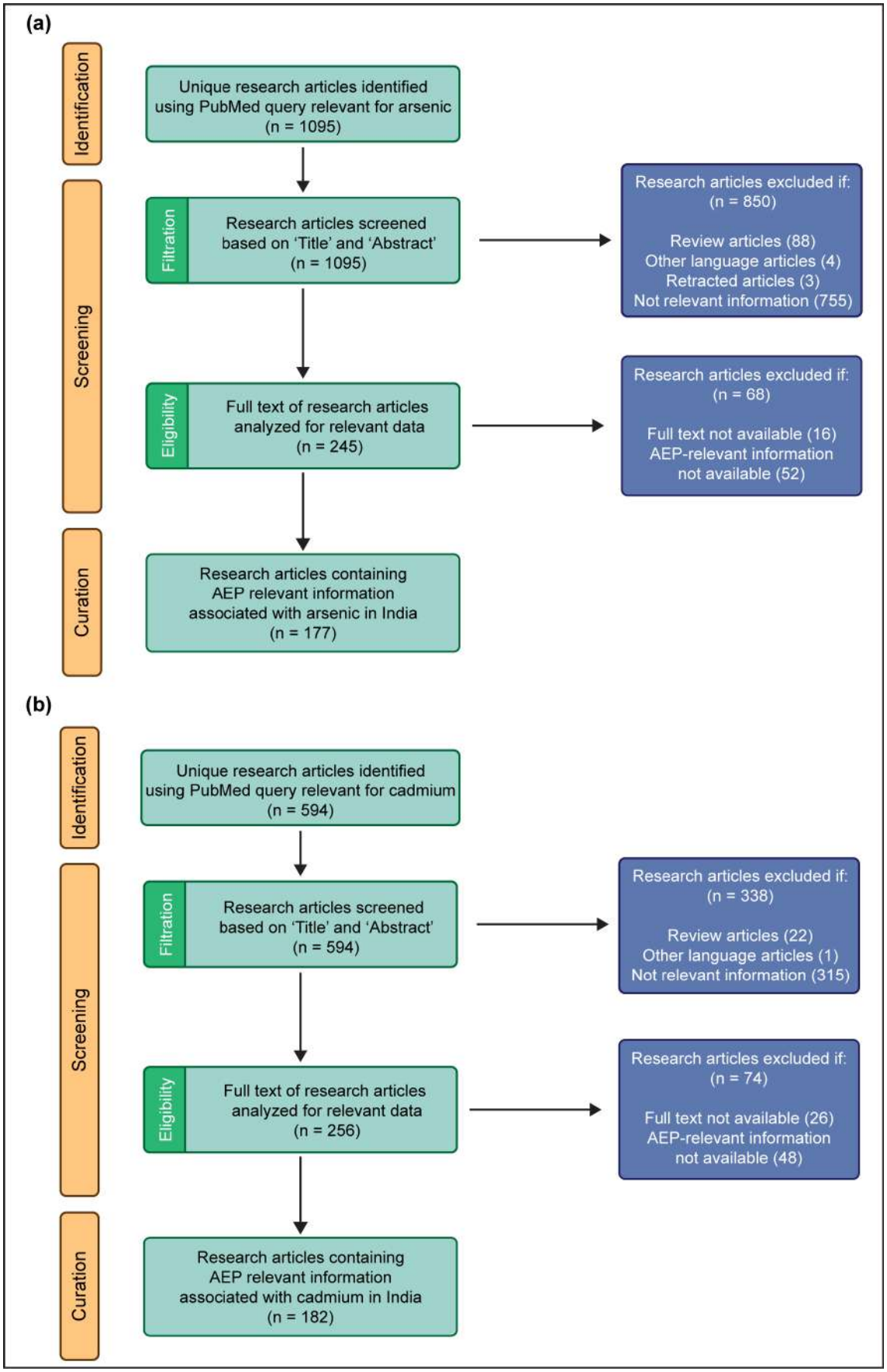

(a)
Identification
Unique research articles identified using PubMed query relevant for arsenic (n = 1095)
Screening
Filtration
Research articles screened based on 'Title' and 'Abstract' (n = 1095)
Research articles excluded if: (n = 850)
Review articles (88)
Other language articles (4)
Retracted articles (3)
Not relevant information (755)
Eligibility
Full text of research articles analyzed for relevant data (n = 245)
Research articles excluded if: (n = 68)
Full text not available (16)
AEP-relevant information not available (52)
Curation
Research articles containing AEP relevant information associated with arsenic in India (n = 177)
(b)
Identification
Unique research articles identified using PubMed query relevant for cadmium (n = 594)
Screening
Filtration
Research articles screened based on 'Title' and 'Abstract' (n = 594)
Research articles excluded if: (n = 338)
Review articles (22)
Other language articles (1)
Not relevant information (315)
Eligibility
Full text of research articles analyzed for relevant data (n = 256)
Research articles excluded if: (n = 74)
Full text not available (26)
AEP-relevant information not available (48)
Curation
Research articles containing AEP relevant information associated with cadmium in India (n = 182)

Figure 4.1 ***(previous page)*:** Workflow to curate AEP-relevant information associated with arsenic and cadmium in India. (a) PRISMA diagram for selecting relevant articles associated with arsenic. (b) PRISMA diagram for selecting relevant articles associated with cadmium.

as the source vertex and the AO as the terminal vertex, while the connections between them, known as key event relationships (KERs), serve as the edges, collectively forming AOPs [47–49, 167].

AOPs are stressor-agnostic and context-specific, meaning each AOP is tailored to a particular biological scenario [47–49, 167]. They are developed globally, targeting diverse adverse outcomes, and are deposited into AOP-Wiki [50], the largest global collaborative platform for AOP development and associated scientific evidence, hosted by the Society for the Advancement of Adverse Outcome Pathways (SAAOP). Therefore, in this chapter, we relied on AOP-Wiki to obtain the latest available AOPs.

Compilation of AOPs within AOP-Wiki

AOP-Wiki makes its complete repository available as a downloadable XML file. In order to obtain information on AOPs within AOP-Wiki, we first downloaded this XML file (released 1 April 2024) from the 'Project Downloads' page in AOP-Wiki [218]. We then parsed and extracted information associated with AOPs, like AOP identifier, AOP title, associated KEs (including MIEs and AOs) and KERs, linked stressors, handbook version followed for development of AOP, status according to OECD, and biological applicability information such as taxonomy, sex and life-stage of the organism, and their corresponding weight of evidence. Additionally, for each KE, we extracted information like KE title, KE identifier, level of biological organization, action name, object name, object identifiers and process name, and information associated with KERs like upstream/downstream KEs, evidence for biological plausibility of KER, adjacency, and the extent of quantitative understanding of KER.

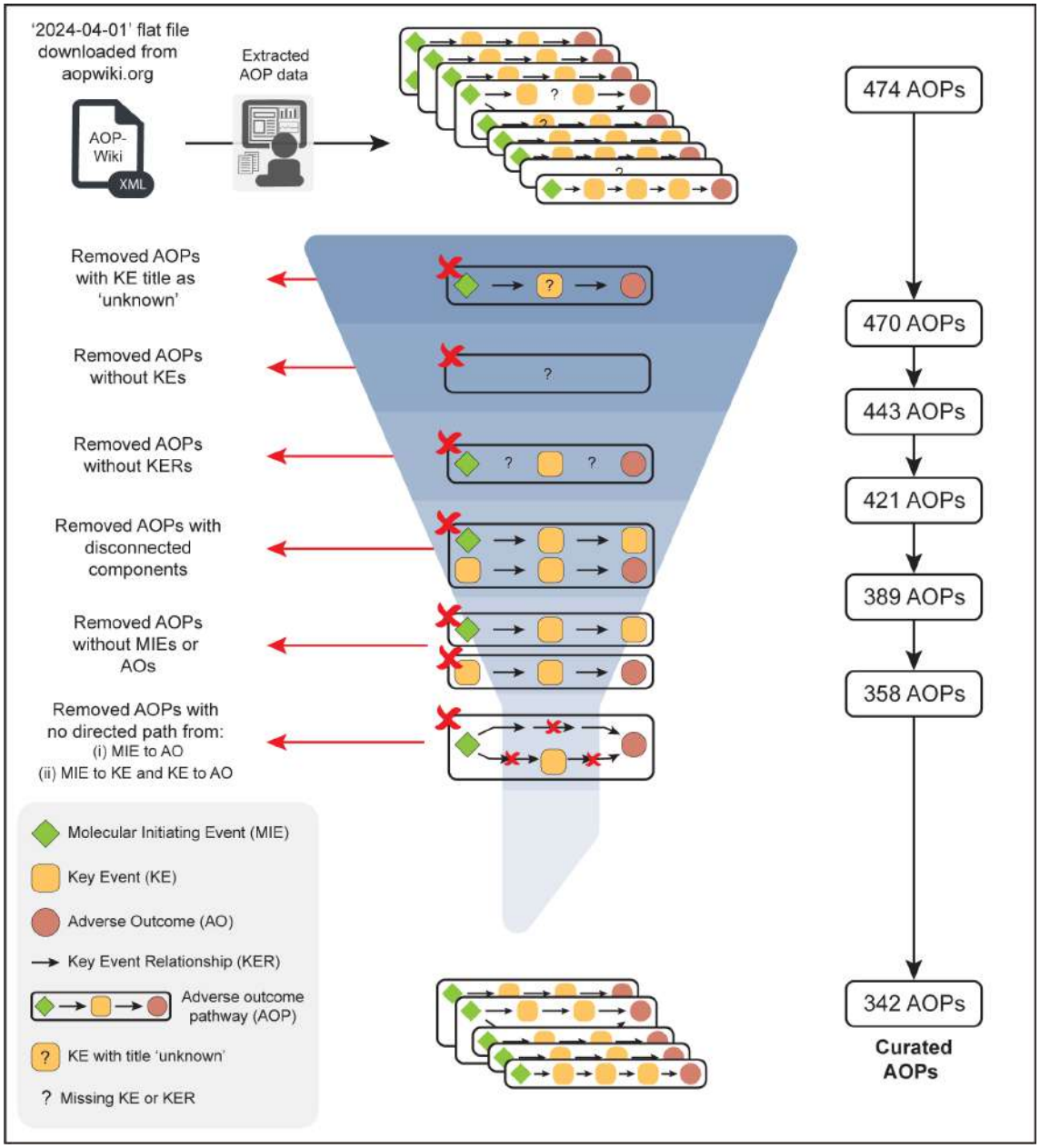


Figure 4.2: Workflow to curate complete and connected Adverse Outcome Pathways (AOPs) from AOP-Wiki by employing computation and manual curation in conjunction.

Identification of complete and connected AOPs within AOP-Wiki

AOPs are considered living documents as they are collaboratively developed globally, and are continuously updated based on new scientific evidence [51]. Consequently, many AOPs may remain incomplete, or lack sufficient information for further analysis [51]. Therefore, based on our previous works [55, 56], we systematically checked each AOP to filter complete and connected AOPs (Figure 4.2). First, we manually checked and removed AOPs that comprised KEs with title as 'unknown', or lacked any KEs or KERs. Next, we employed NetworkX library [168] in Python to check for disconnected AOPs, and manually updated and filtered out AOPs that contained disconnected components. Finally, we checked for the presence of MIE, AO, and a directed path between them, and filtered out AOPs that lacked any. Through these combined manual and computational efforts, we identified 342 complete, and connected AOPs from AOP-Wiki (last accessed on 9 July 2024), which we designated as 'curated AOPs' (Supplementary Table S4.4). These 342 curated AOPs comprised 1163 unique KEs (Supplementary Table S4.5) and 1827 unique KERs (Supplementary Table S4.6).

Computation of cumulative weight of evidence (WoE) of AOPs

AOPs carry an endorsement status within AOP-Wiki [51], with 35 of 342 curated AOPs being fully endorsed, 13 currently under review, and the remainder still under development (Supplementary Table S4.4). Since endorsement is often a time-consuming process, the complete AOPs can be regarded as plausible hypotheses for toxicity mechanisms, even if not yet formally endorsed. To further assess the confidence associated with these curated AOPs, we evaluated the cumulative weight of evidence based on the support for individual KERs within each of the curated AOPs (Supplementary Table S4.6).

AOP-Wiki provides a weight of evidence (WoE) information for each KER in an AOP based on its biological plausibility (Supplementary Table S4.6). The WoE is a qualitative score represented in terms of 'High', 'Moderate', 'Low' and 'Not Specified'. Following Ravichandran *et al.* [54], we computed the fraction of KERs within an AOP with

'High' WoE [represented as F('High')], 'Moderate' WoE [represented as F('Moderate')], 'Low' WoE [represented as F('Low')] and 'Not Specified' WoE [represented as F('Not Specified')], and subsequently assigned a cumulative WoE to the AOPs based on the following criteria [54]:

i. High: If F('High') ≥ 0.5

ii. Moderate: If F('High') < 0.5, but (F('High') + F('Moderate')) ≥ 0.5

iii. Low: If (F('High') + F('Moderate')) < 0.5, but (F('High') + F('Moderate') + F('Low')) ≥ 0.5

iv. Not Specified: If none of the above criteria are satisfied

4.1.4 Construction of stressor-AOP network for inorganic arsenic and cadmium

Stressor-AOP network provides a broader perspective on impacts of stressors across diverse biological processes by linking the stressors to AOPs [56, 166]. To better understand the perturbances caused by inorganic arsenic and cadmium, we constructed a stressor-AOP network which is a bipartite graph linking arsenic and cadmium to different AOPs within AOP-Wiki. Building on our previous works [55–57], we first curated all biological endpoint data in humans and ecologically relevant species associated with inorganic arsenic and cadmium from ToxCast [18], CTD [25, 151], DEDuCT [105, 120, 121], NeurotoxKb [126, 127], and ECOTOX [63, 217]. We then mapped these endpoints to KEs within AOP-Wiki to identify chemical-relevant KEs. The following sections elaborate the process of endpoint extraction, followed by their mapping to KEs within AOPs.

Identification of KEs using ToxCast

ToxCast is a program by the United States Environmental Protection Agency (USEPA) designed to enhance chemical toxicity predictions through *in vitro* high-throughput screening of various environmental chemicals [18]. These *in vitro* approaches provide crucial infor-

mation on the associated biological processes and gene alterations triggered by stressor interactions, which can aid in identification of potential MIEs related to active chemicals [52, 55–57, 166, 219]. Therefore, ToxCast invitrodb v4.1, available on the USEPA repository [220], was used to identify KEs/MIEs associated with inorganic arsenic and cadmium.

First, we extracted the chemicals and their corresponding assay information from the '`mc5-6_winning_model_fits-flags_invitrodb_v4_1_SEPT2023.csv`' file, and identified active assay endpoints for each chemical (defined as 'hitc' ≥ 0.9) [19]. Next, we determined whether these active chemicals exhibited an 'activatory' or 'inhibitory' response based on the 'top' value of the corresponding winning model from the '`mc4_all_model_fits_invitrodb_v4_1_SEPT2023.csv`' file [19].

To ensure that these active endpoints are not due to non-specific activation of reported genes (referred to as 'cytotoxicity-associated burst' phenomenon) [164], we applied the following Z-score statistic proposed by Judson *et al.* [164]:

$$\text{Z(chemical, assay)} = \frac{-\text{logAC}_{50}\text{(chemical, assay)} - \text{median}[-\text{logAC}_{50}\text{(chemical, cytotox)}]}{\text{global cytotoxicity MAD}}$$

Here, 'logAC_{50}(chemical, assay)' is the logarithm of the AC_{50} value of the chemical in the assay, 'logAC_{50}(chemical, cytotox)' is the logarithm of the AC_{50} value of the chemical in the corresponding cytotoxicity assay, and the 'global cytotoxicity MAD' is the median of the MAD (median absolute deviation) of the 'logAC_{50}(chemical, cytotox)' distributions across all chemicals [164]. Based on this definition, Z-scores ranging between +3 and -3 are considered indicative of cytotoxicity-associated bursts [164].

We retrieved the global cytotoxicity MAD and logAC_{50}(chemical, cytotox) (given by the column titled 'cytotox_median_log') from the '`cytotox_invitrodb_v4_1_SEPT2023.xlsx`' file to identify the cytotoxicity-associated bursts for inorganic arsenic and cadmium compounds [19]. We then discarded assays with Z-scores between +3 and -3, resulting in 281 assay endpoints linked to one inorganic arsenic compound and four inorganic cadmium

compounds, which we subsequently mapped to KEs within AOP-Wiki.

We first extracted the genes corresponding to these 281 assay endpoints from the 'assay_annotations_invitrodb_v4_1_SEPT2023.xlsx' file, along with assay endpoint-gene mappings from the 'assay_gene_mappings_invitrodb_v4_1_SEPT2023' file. Next, using the KE-gene annotations provided by Saarimäki *et al.* [178], we identified gene sets linked to KEs at the molecular or cellular levels of biological organization. We then mapped these KEs to ToxCast assay endpoints based on gene overlaps and manually filtered the results using the assay endpoint descriptions. Through this comprehensive toxicogenomics-based manual curation, we identified 52 KEs associated with 40 assay endpoints across one inorganic arsenic compound (CAS:1327-53-3) and two inorganic cadmium compounds (CAS:10108-64-2, CAS:10325-94-7) (Supplementary Table S4.7).

Identification of KEs using CTD

CTD is a comprehensive public resource that links environmental chemicals (C), genes (G), phenotypes (P), and diseases (D) to advance understanding of their effects on health [25, 151]. This resource facilitates the construction of CGPD-tetramers, which help identify confident associations between chemicals, phenotypes, and diseases, enabling their mapping to KEs within AOP-Wiki [221, 222]. Based on our previous works [55–57], we retrieved high confidence CGPD-tetramers associated with inorganic arsenic and cadmium compounds, and leveraged them to identify KEs within AOP-Wiki.

We accessed the CTD June 2024 release and constructed the CGPD-tetramers where we considered: (i) chemical-gene and chemical-phenotype associations with literature evidence; (ii) chemical-disease and gene-disease associations with 'marker/mechanism' or 'marker/mechanism|therapeutic' evidence; (iii) gene-phenotype associations with GO annotations based on only the experimental results [171]. This process resulted in a list of 17749 CGPD-tetramers comprising three inorganic arsenic compounds and three inorganic cadmium compounds, 1162 genes, 483 phenotypes and 213 diseases (Supplementary Table S4.8). Next, we generated the immediate neighboring GO terms for the CGPD-tetramer

phenotype GO terms using the GOSim package [172] in R. We then overlapped these GO terms with the process identifiers of KEs in AOP-Wiki, manually examined them, and identified 197 KEs linked to 105 phenotypes across three inorganic arsenic compounds (CAS:1327-53-3, CAS:56320-22-0, CAS:7784-34-1) and three inorganic cadmium compounds (CAS:10108-64-2, CAS:7440-43-9, CAS:10124-36-4) (Supplementary Table S4.7). Additionally, we manually examined the disease terms and identified 92 KEs associated with 106 diseases across three inorganic arsenic compounds (CAS:1327-53-3, CAS:56320-22-0, CAS:7784-34-1) and two inorganic cadmium compounds (CAS:7440-43-9, CAS:10108-64-2) (Supplementary Table S4.7).

Identification of KEs using DEDuCT and NeurotoxKb

DEDuCT is one of the largest databases compiling curated information on endocrine disrupting chemicals (EDCs) and their corresponding endocrine-mediated endpoints from published literature [105, 120, 121]. Therefore, we extracted the endocrine-mediated endpoints corresponding to inorganic arsenic and cadmium compounds within DEDuCT and considered them to identify relevant KEs within AOP-Wiki. We manually inspected both the endocrine-mediated endpoints and KE titles in AOP-Wiki, and identified 76 KEs associated with 46 endocrine-mediated endpoints across two inorganic arsenic compounds (CAS:7440-38-2, CAS:1327-53-3) and three inorganic cadmium compounds (CAS:7440-43-9, CAS:10108-64-2, CAS:10325-94-7) (Supplementary Table S4.7).

NeurotoxKb is a manually curated resource focusing on mammalian neurotoxicity endpoints associated with environmental chemicals, compiled from published literature [126, 127]. Therefore, we extracted the neurotoxic endpoints corresponding to inorganic arsenic and cadmium compounds within NeurotoxKb, and considered them to identify relevant KEs within AOP-Wiki. We manually inspected both the neurotoxic endpoints and KE titles in AOP-Wiki, and identified 20 KEs associated with 13 neurotoxic endpoints across one inorganic arsenic compound (CAS:7440-38-2) and two inorganic cadmium compounds (CAS:7440-43-9, CAS:10108-64-2) (Supplementary Table S4.7).

Identification of KEs using AOP-Wiki

AOP-Wiki catalogs information on prototypical stressors for each AOP, based on well-documented associations with these stressors [51]. We first extracted information on prototypical stressors associated with AOPs from the downloaded XML file. We then observed that four AOPs (AOP:263, AOP:413, AOP:499, AOP:500) were associated with two inorganic arsenic compounds (CAS:7440-38-2, CAS:15502-74-6), and four AOPs (AOP:257, AOP:296, AOP:499, AOP:500) were associated with two inorganic cadmium compounds (CAS:7440-43-9, CAS:10108-64-2). Therefore, we considered the KEs within these AOPs to be associated with inorganic arsenic and cadmium, respectively.

Identification of KEs using ECOTOX

ECOTOX is a comprehensive USEPA resource that consolidates toxicological information for over 12000 chemicals across more than 13000 terrestrial and aquatic species from published literature, offering valuable insights into ecological impacts of these chemicals [63]. Therefore, we leveraged ECOTOX to identify ecotoxicologically-relevant KEs within AOP-Wiki associated with inorganic arsenic and cadmium compounds.

We first downloaded the ECOTOX dataset files (released June 2024) by selecting the 'Download ASCII Data' option on the ECOTOX website [217]. Then, using an in-house Python script, we parsed 'tests.txt', 'results.txt', 'chemicals.txt' and 'species.txt' files to extract data associated with inorganic arsenic and cadmium compounds. ECOTOX curates the biological effects associated with the chemicals as 'Effect', the corresponding measure as 'Measurement', and the trend of the effect as 'Trend'. We removed the data points where the Measurement value was either 'Not Reported' or empty, as they lack sufficient information on adverse effects, and proceeded to map the remaining data to KEs within AOP-Wiki. We manually inspected the 'Effect', 'Measurement' and 'Trend' information from ECOTOX, and KE titles within AOP-Wiki, and identified 273 KEs to be associated with 499 unique ECOTOX endpoints across 25 inorganic arsenic compounds (CAS:7784-46-5, CAS:7778-43-0, CAS:1327-53-3, CAS:13464-37-4, CAS:1303-11-3,

CAS:7440-38-2, CAS:7631-89-2, CAS:7778-39-4, CAS:10103-60-3, CAS:13466-06-3, CAS:7784-41-0, CAS:7778-44-1, CAS:15584-04-0, CAS:15502-74-6, CAS:69162-99-8, CAS:13464-38-5, CAS:22541-54-4, CAS:17428-41-0, CAS:1303-33-9, CAS:1303-28-2, CAS:7784-34-1, CAS:7778-41-8, CAS:17029-22-0, CAS:15120-17-9, CAS:12777-38-7) and 8 inorganic cadmium compounds (CAS:10108-64-2, CAS:7440-43-9, CAS:7790-80-9, CAS:10325-94-7, CAS:10124-36-4, CAS:10022-68-1, CAS:1306-19-0, CAS:1306-23-6) (Supplementary Table S4.7).

S. No.	Resource	Type of endpoint obtained	Number of chemicals	Number of endpoints	Number of mapped KEs within AOP-Wiki
1	AOP-Wiki	AOPs with prototypical stressor information	4	6	27
2	CTD	Diseases from CGPD-tetramers constructed using experimental data	5	106	92
		Phenotypes from CGPD-tetramers constructed using experimental data	6	105	197
3	DEDuCT	Endocrine-mediated endpoints	5	46	76
4	ECOTOX	Measurement and Trend (excluding 'No Effect' trend) information	33	499	273
5	NeurotoxKb	Neurotoxicant endpoints	3	13	20
6	ToxCast	Active endpoints (hitc $\geq$ 0.9) post-cytotoxicity burst analysis	3	40	52

Table 4.1: Summary of resources and criteria for endpoint identification with chemical counts and mapped key events (KEs). [AOPs: Adverse Outcome Pathways; CTD: Comparative Toxicogenomics Database; CGPD-tetramers: chemical-gene-phenotype-disease tetramers; hitc: hit call value in ToxCast invitrodb]

Overall, we identified 579 KEs associated with either inorganic arsenic or cadmium compounds by integrating heterogeneous toxicogenomics and biological endpoints data from six exposome-relevant resources, namely, ToxCast, CTD, DEDuCT, NeurotoxKb, AOP-Wiki, and ECOTOX (Supplementary Table S4.7). Table 4.1 provides an overview of

the resources utilized, the criteria applied to extract endpoints from them, and the number of chemicals and endpoints obtained from each.

Construction of stressor-AOP network

Stressor-AOP network provides a broader perspective on impacts of stressors across diverse biological processes by linking the stressors to AOPs [56, 166]. To construct the stressor-AOP network, we first grouped KEs associated with various inorganic arsenic compounds to identify common KEs related to inorganic arsenic. Similarly, we grouped KEs for different inorganic cadmium compounds to identify those associated with inorganic cadmium. To obtain high confidence associations between the chemicals and AOPs, we relied solely on the curated list of 342 AOPs (Supplementary Table S4.4). Finally, leveraging these arsenic-KE and cadmium-KE associations, we constructed a bipartite stressor-AOP network, linking arsenic and cadmium to 310 and 318 curated AOPs, respectively (Supplementary Table S4.9).

Thereafter, we characterized these links based on two criteria, namely, the coverage score and the level of relevance. The coverage score of a stressor-AOP link is defined as the ratio of the number of KEs within that AOP associated with the stressor to the total number of KEs within that AOP [174]. This score is a real valued number between 0 and 1, and we denote it as the edge weight of the linkage between a stressor and an AOP in our stressor-AOP network. The level of relevance is a qualitative score used to assess the relevance of the stressor-AOP association within the stressor-AOP network, and we denote this score as an edge attribute in our stressor-AOP network [56]. Level of relevance is a five-level criterion defined as follows:

- Level 1: The stressor is associated with at least one KE within an AOP, where the KE is neither MIE nor AO within that AOP
- Level 2: The stressor is associated with at least one AO within an AOP, but not associated with any MIE within that AOP
- Level 3: The stressor is associated with at least one MIE within an AOP, but not

associated with any AO within that AOP

- Level 4: The stressor is associated with at least one MIE and one AO within an AOP
- Level 5: The stressor is associated with at least one MIE and one AO within an AOP, and there exists a directed path between the associated MIE and AO

Supplementary Table S4.9 contains all the data on the stressor-AOP network constructed for inorganic arsenic and cadmium, including the coverage score and level of relevance for each of the stressor-AOP links. We visualized the resultant networks using Cytoscape [132].

4.1.5 Construction of stressor-species networks by utilizing ECOTOX data

The stressor-species network provides a framework to understand how multiple environmental stressors interact with various species, offering insights into their effects on ecosystems and biodiversity [57, 60, 62]. Therefore, we constructed stressor-species networks for the inorganic compounds arsenic and cadmium by leveraging ECOTOX data on chemical concentrations associated with acute and chronic toxicity, as well as bioconcentration factors (BCFs), to gain a comprehensive understanding of their toxicities and accumulation potential across different species. Acute toxicity is measured in terms of median lethal concentration (LC_{50}) or effect concentration (EC_{50}), that refers to the concentration which causes lethality or induces desired biological response in 50% of the tested organisms, respectively [63]. Chronic toxicity is measured in terms of no observed effect concentration (NOEC), which is the highest tested concentration that causes no observable adverse effect on tested organisms [63]. BCF is the degree to which different species take up chemicals from their environment via respiration and skin contact, excluding dietary uptake [223].

We first normalized the acute toxicity concentrations to parts per million (ppm) equivalents (mg/L or mg/kg) by relying on corresponding conversion factors (Supplementary

Table S4.10). Then, for each stressor-species pair, we considered studies with durations between 24 and 96 hours, and for pairs with multiple concentration values, we computed their geometric means. Thereafter, we constructed a network linking 15 inorganic arsenic and 8 inorganic cadmium compounds to 167 and 644 species, respectively (Supplementary Table S4.11). Similarly, for chronic toxicity values, we first normalized the concentrations to ppm equivalents, and computed their geometric mean in case of more than one concentration for a stressor-species pair, and constructed a network linking 16 arsenic and 4 cadmium compounds to 130 and 264 species, respectively (Supplementary Table S4.11). Similarly, we normalized the BCF values, and the highest BCF value per stressor-species pair was selected, and a network was constructed linking 5 arsenic and 4 cadmium compounds to 32 and 147 species, respectively (Supplementary Table S4.12). We visualized the resultant networks using Cytoscape [132].

4.1.6 Construction of species sensitivity distributions (SSDs) for inorganic arsenic and cadmium compounds by utilizing ECOTOX data

Species sensitivity distribution (SSD) is a widely used statistical method in ecological risk assessment to estimate species' sensitivities to environmental stressors, and derive hazard concentration affecting 5% of the species (HC05), which informs regulatory guidelines for biodiversity protection [64–67]. In this chapter, we curated relevant data from ECOTOX to construct SSDs for inorganic arsenic and cadmium compounds.

First, we filtered the data to obtain species with habitat as 'Water', and standardized the concentrations to ppm equivalents for the inorganic compounds of arsenic and cadmium. To obtain relevant acute toxicity data, we considered only LC_{50} or EC_{50} values obtained from studies ranging between 24 and 96 hours [224], and calculated the geometric mean if multiple values were available per stressor-species pair [225, 226]. To obtain relevant chronic toxicity data, we considered only NOEC values, and calculated the geometric mean

if multiple values were available per stressor-species pair. Finally, to ensure robust SSDs, we included only chemicals with data across at least five ECOTOX species groups [227]. Through this extensive process, we identified acute toxicity data for four inorganic arsenic compounds (CAS:1327-53-3, CAS:7778-43-0, CAS:7784-46-5, CAS:13464-38-5), and four inorganic cadmium compounds (CAS:7440-43-9, CAS:10108-64-2, CAS:10124-36-4, CAS:10325-94-7) (Supplementary Table S4.13). Furthermore, we identified that chronic toxicity data is available for three inorganic arsenic compounds (CAS:1327-53-3, CAS:7778-43-0, CAS:7784-46-5) and four inorganic cadmium compounds (CAS:7440-43-9, CAS:10108-64-2, CAS:10124-36-4, CAS:10325-94-7) (Supplementary Table S4.14). We utilized this data to construct SSDs based on acute and chronic toxicities associated with inorganic arsenic and cadmium.

To construct the SSDs for inorganic compounds of arsenic and cadmium, we relied on two tools namely, SSD Toolbox developed by the USEPA [228] and R-based ssdtools package developed by the Ministry of Environment and Climate Change Strategy of British Columbia [225]. In both these tools, we fit five statistical distributions, namely, Log-Normal, Log-Logistic, Log-Gumbel, Weibull and Burr Type III to the curated toxicity data through maximum-likelihood method, and obtained the HC05 value based on the best-fit model determined by the minimum corrected Akaike Information Criterion (AICc) value [225,229,230]. Furthermore, we set the number of bootstrap resampling iterations as 10000 to quantify uncertainty in fitted parameters and estimate confidence intervals [225, 230]. We observed that the Burr Type III model did not adequately fit the chronic toxicity data for the inorganic cadmium compound CAS:10325-94-7 in both tools. Therefore, only the remaining four statistical distributions were considered to compute the SSD for this chemical.

Recently, model averaging, which assigns weights to individual distributions to derive a weighted average, has been proposed as a more reliable method for computing SSDs, as it has been observed to produce more stable HC05 values compared to single-model estimates [225,231]. Accordingly, we additionally computed model-averaged HC05 values

for each inorganic arsenic and cadmium compound using both tools.

4.1.7 Construction of toxicity-normalized SSD (SSDn) for inorganic arsenic and cadmium

The toxicity-normalized SSD (SSDn) approach refines ecological risk assessment by normalizing chemical toxicity data based on a sensitive reference species, thereby providing a framework to pool data of chemicals that share toxicological similarities [68, 69]. This SSDn approach further allows the determination of chemical-specific HC05 values by back-calculating HC05n values (i.e., HC05 value derived from SSDn) using the toxicity data from the corresponding reference species [68, 69]. Therefore, the SSDn approach was utilized to assess the overall toxicity of inorganic arsenic and cadmium, and allow comparison with the chemical-specific SSDs.

Lambert *et al.* [69] suggested the following criterion to choose an appropriate reference species (referred to as 'nSpecies') for the construction of SSDn:

- All chemicals of interest must have toxicity data for the nSpecies
- When there are multiple potential nSpecies, the most sensitive species should be used for normalization
- When possible, normalize toxicity values to the nSpecies with the most similar sensitivity to all compounds (i.e., smallest range of percentiles in single-chemical SSDs)

In the case of arsenic, we observed that *Daphnia magna* was the only species with acute toxicity data across the four inorganic arsenic compounds (Supplementary Table S4.13). Therefore, we chose *Daphnia magna* as the nSpecies for acute toxicity data-based SSDn computation for arsenic. We observed that six species, namely, *Cyprinus carpio*, *Danio rerio*, *Daphnia magna*, *Oncorhynchus mykiss*, *Rhinella arenarum*, and *Terapon jarbua* had chronic toxicity data for all three inorganic arsenic compounds (Supplementary Table S4.14). Among these, the chronic toxicity data of *Terapon jarbua* satisfied the other

two criterion outlined by Lambert *et al.* [69], and therefore we chose *Terapon jarbua* as nSpecies for chronic toxicity data-based SSDn computation for arsenic.

In the case of cadmium, we observed that 10 species, namely, *Artemia salina*, *Asellus aquaticus*, *Brachionus calyciflorus*, *Danio rerio*, *Daphnia magna*, *Daphnia pulex*, *Gambusia affinis*, *Hyalella azteca*, *Oncorhynchus mykiss*, and *Pimephales promelas* had acute toxicity data for all the four inorganic cadmium compounds (Supplementary Table S4.13). Among these, the toxicity data of *Hyalella azteca* satisfied the other two criterion outlined by Lambert *et al.* [69], and therefore we chose *Hyalella azteca* as nSpecies for acute toxicity data-based SSDn computation for cadmium. We observed that two species, namely, *Cyprinus carpio*, and *Oncorhynchus mykiss* had chronic toxicity data for all four inorganic cadmium compounds (Supplementary Table S4.14). Among these, the chronic toxicity data of *Oncorhynchus mykiss* satisfied the other two criterion outlined by Lambert *et al.* [69], and therefore we chose *Oncorhynchus mykiss* as nSpecies for chronic toxicity data-based SSDn computation for cadmium.

Next, we normalized the toxicity data with respect to that of nSpecies for the identified inorganic compounds of arsenic and cadmium, and we designated them as nSpecies equivalents. Then, for each species, we computed the geometric mean of the nSpecies equivalents across different chemicals to obtain the corresponding species mean toxicity values (SMTVs) (Supplementary Tables S4.15-S4.16). Subsequently, we followed procedure used for SSD construction, and computed the toxicity-normalized hazard concentration (HC05n) values for inorganic arsenic and cadmium by relying on their corresponding SMTV values. We estimated the chemical specific HC05 values from the corresponding HC05n using the following formula suggested by Lambert *et al.* [69]:

$$\mathrm{HC05}_{\mathrm{chemical,SSDn}} = \mathrm{HC05n} * \mathrm{SMTV}_{\mathrm{chemical,nSpecies}}$$

Here, $\mathrm{SMTV}_{\mathrm{chemical,nSpecies}}$ refers to the SMTV for the nSpecies identified for a chemical, and $\mathrm{HC05}_{\mathrm{chemical,SSDn}}$ represents the derived HC05 value for that chemical from the corre-

sponding HC05n value. Additionally, Lambert *et al.* [69] calculated the fold change as a ratio of the HC05 value derived from SSDn to that of single-chemical SSD, and used it as a metric to assess the accuracy of the values computed using the SSDn approach. Here, we used the following formula to compute the fold change:

$$\text{Fold change} = \frac{\text{HC05}_{\text{chemical,SSDn}}}{\text{HC05}_{\text{chemical}}}$$

where, $\text{HC05}_{\text{chemical}}$ represents the HC05 value calculated from the single-chemical SSD for the corresponding chemical. As per Lambert *et al.* [69], values falling within a 5-fold range are considered 'Good'.

4.1.8 Calculation of risk quotient (RQ) based on the environmental concentrations of arsenic and cadmium

The risk quotient (RQ) of a chemical is defined as the ratio of its measured environmental concentration (MEC) to its predicted no-effect concentration (PNEC) [71, 72]. In 2019, the Central Water Commission (CWC) under Department of Water Resources, RD & GR, Ministry of Jal Shakti, Government of India, published a report titled '*Status of Trace and Toxic Metals in Indian Rivers*' [232] providing the analysis and data on various metal elements in Indian rivers, including average seasonal concentrations categorized as 'monsoon' and 'non-monsoon' across sampling stations. Importantly, this report provided total arsenic and total cadmium related concentration data across 424 sampling stations, 247 rivers, and 24 states. Therefore, to compute the RQ of the total arsenic and cadmium reported at these sites, we first computed the PNEC values of the inorganic arsenic and cadmium compounds.

We first standardized the reported concentrations to parts per billion (ppb) by converting their units to ppb equivalents (μg/L) and considered them as measured environmental concentrations (MECs). Thereafter, the predicted no-effect concentration (PNEC) values for inorganic arsenic and cadmium compounds were computed using the following

formula [70]:

$$\text{PNEC} = \frac{\text{HC05}}{\text{AF}}$$

where, HC05 is obtained from the SSD constructed using chronic toxicity data in the USEPA's SSD Toolbox, and AF is the assessment factor. Given the diversity of the data in ECOTOX, we selected the most conservative AF of 5 to minimize the risk of underestimation, aligning with the EU Commission's 'Technical Guidance Document on Risk Assessment' [70] and published literature [57, 233–237].

Thereafter, we computed the geometric mean of the computed PNECs for inorganic arsenic and cadmium compounds to obtain the values for arsenic and cadmium, respectively. Finally, the RQ was computed using the following formula [71, 72]:

$$\text{RQ} = \frac{\text{MEC}}{\text{PNEC}}$$

where, MEC is the measured environmental concentration obtained from the report and PNEC is the computed geometric mean of arsenic or cadmium compounds.

We further utilized these computed RQs to perform three comparative analyses: non-monsoon and monsoon RQs for arsenic, non-monsoon and monsoon RQs for cadmium, and arsenic and cadmium RQs across sampling stations. We used the Shapiro-Wilks test [238] to check for normality of the distribution of differences. If the distribution was found to be normal, we used the paired t-test [239], else we used the Wilcoxon signed-rank test [240]. We performed this analysis using the rstatix package in R version 4.3.1.

4.2 Results

4.2.1 Inorganic arsenic and cadmium compounds curated from six exposome-relevant datasets

The inorganic forms of arsenic and cadmium are particularly hazardous due to their higher absorption rates, multiple valence states, and their potential to cause damage at both cellular and organ levels [206–209]. In this chapter, we relied on network toxicology-based approaches to assess the adverse effects of these contaminants on human and ecosystem health. First, we systematically curated 62 inorganic arsenic and 18 inorganic cadmium compounds from six exposome-relevant databases (Supplementary Table S4.1). Among these, we observed that many compounds are prohibited from usage in the European Union (EU) [241], with some listed as high concern chemicals under EU's Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) regulations [242] (Supplementary Table S4.1). Notably, we also observed that some of these compounds have been detected in human biospecimens, including the placenta, urinary bladder, brain, skin, bone, kidney, and pancreas [243, 244] (Supplementary Table S4.1). Furthermore, we utilized the USEPA's Chemical and Products Database (CPDat) [116] and observed that some of these compounds are present in consumer products such as batteries, electrical appliances, and construction materials (Supplementary Table S4.1).

Chemical speciation strongly influences the toxicity, mobility and bioavailability of arsenic and cadmium [245, 246]. For instance, As(III) (arsenite) is generally far more toxic and more easily absorbed in biological systems than As(V) (arsenate), owing to its greater reactivity with thiol groups and its higher capacity to cross cell membranes [247]. Similarly, for cadmium, both its free ionic form and its soluble complexes tend to be more bioavailable and thus more toxic, while insoluble compounds exhibit lower immediate uptake [248, 249]. Importantly, chemical speciation is governed by several environmental conditions, including pH and the presence of other salts, and requires specialized techniques

to identify the different fractions [246]. Since the chemical data was curated from existing repositories, we manually checked for the parameters such as oxidation state to gauge the different species of heavy metals among the curated compounds. For arsenic, we identified 12 arsenites, 36 arsenates, 11 arsines (-3 oxidation state), 1 elemental arsenic, and 2 compounds with a +2 oxidation state, while for cadmium we identified 1 elemental cadmium, 3 isotopes, and 14 different compounds (Supplementary Table S4.1).

We used this set of chemicals for further network toxicology-based analyses (Figure 4.3). We also made all results reported in this chapter readily accessible for further research through the associated website: https://cb.imsc.res.in/heavymetaltox/.

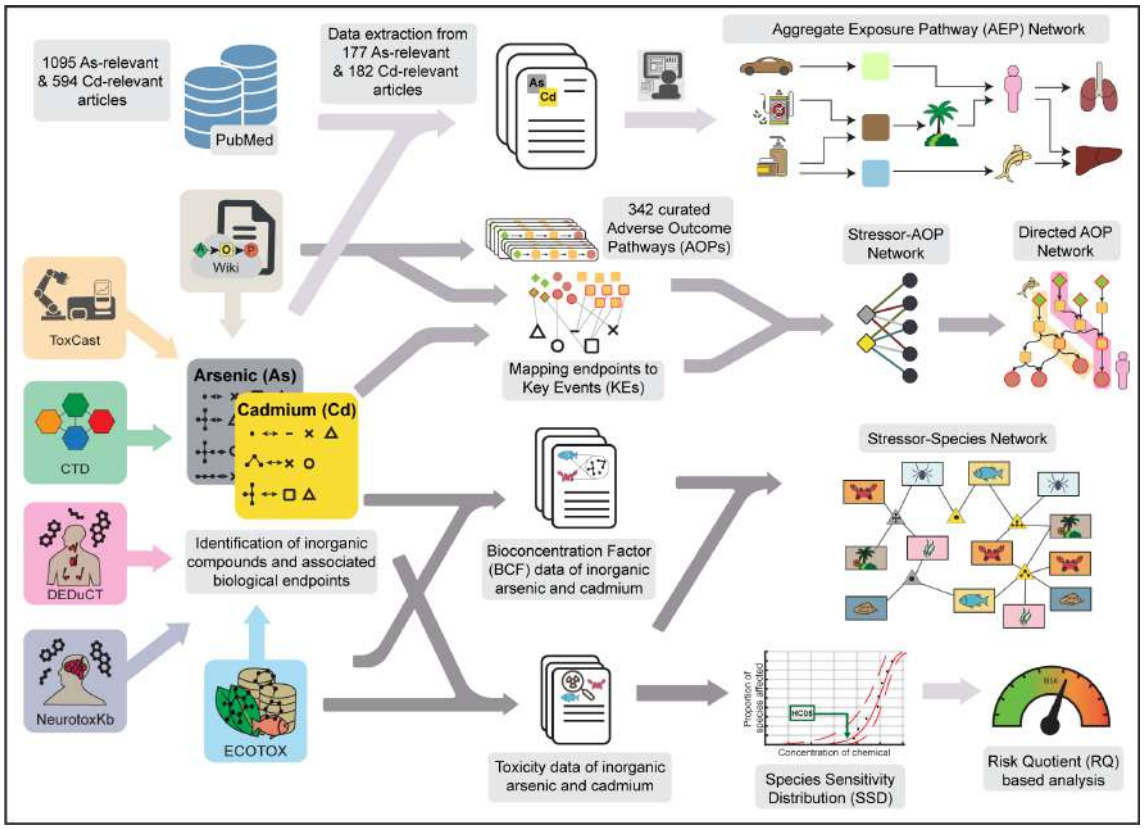


Figure 4.3: Summary of the workflow followed to identify inorganic arsenic and cadmium compounds from six exposome-relevant resources, followed by the exploration of their impacts on human and ecosystem health through Aggregate Exposure Pathway (AEP) networks, stressor-Adverse Outcome Pathway (AOP) networks, stressor-species networks, and Species sensitivity distributions (SSDs) based on their toxicity values, followed by their risk quotient (RQ)-based analysis to assess their risks.

4.2.2 AEP networks for inorganic arsenic and cadmium in India

Arsenic and cadmium are highly toxic contaminants that are widely present in the environment [206, 209]. AEPs are structured frameworks that organize exposure information by linking contaminant sources, environmental transport and transformation, and internal concentrations at target sites [42–45]. By connecting these processes through empirically measurable KESs, AEPs provide a mechanistic view of how chemicals move from the environment to biological receptors [44, 45].

Here, we constructed AEP networks for both inorganic arsenic and cadmium by systematically curating India-specific data from published literature available in PubMed [91]. First, we systematically screened through arsenic-relevant articles and curated information on environmental concentrations and flow of chemical through different states. We then standardized the information into KESs and KTRs to construct an AEP network comprising 64 unique nodes, or KESs, and 142 edges, or KTRs (Figure 4.4; Supplementary Tables S4.2-S4.3). We observed that the KES 'Human' had the highest in-degree in this AEP network, indicating it serves as a major sink or receptor for multiple exposure paths. Conversely, we observed that KES 'Industry' had the highest out-degree, indicating its role as a critical source for downstream exposure pathways. Furthermore, we observed that multiple studies across various geographic regions reported the KES 'Geological' (Supplementary Table S4.2), emphasizing its role as a recurring source for arsenic in India. Moreover, we observed that multiple studies frequently reported groundwater contamination, particularly in the Ganga-Brahmaputra-Meghna plains. In this region, research has linked arsenic exposure to various health effects through consumption of contaminated groundwater [210–213]. Additionally, evidence indicates that arsenic bioaccumulates in rice grains, fruits and vegetables cultivated in these plains due to contaminated irrigation water, highlighting a significant dietary exposure pathway for both humans and other organisms [250–254].

Similarly, for cadmium, we systematically compiled data from 182 cadmium-relevant

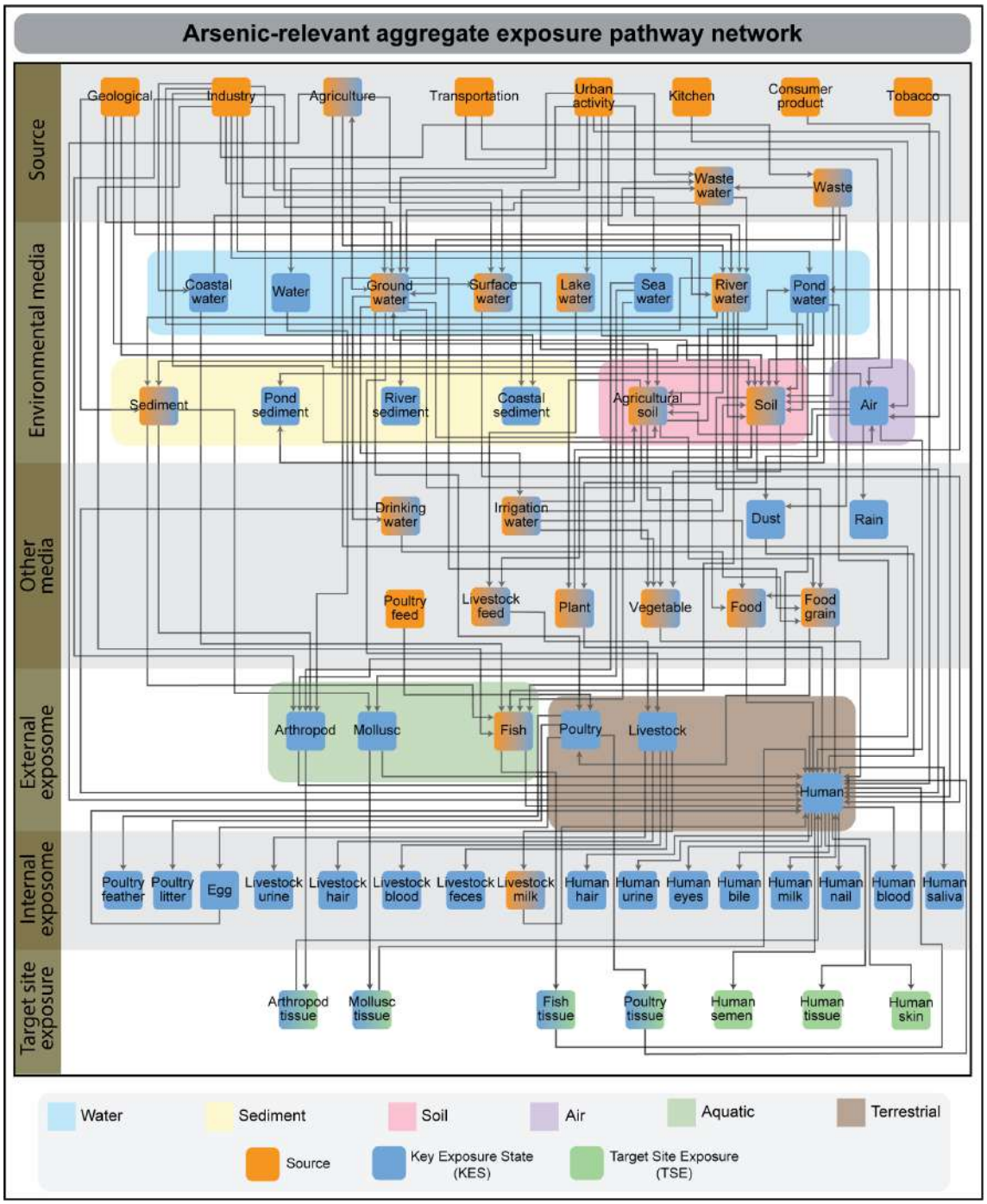


Figure 4.4: Aggregate Exposure Pathway (AEP) network corresponding to arsenic in the Indian context, comprising 64 nodes, or key exposure states (KESs), and 142 edges, or key transitional relationships (KTRs). Among the 64 KESs, 8 are categorized as Source (denoted by orange squares), 19 are categorized as both Source and KES (denoted by orange/blue squares), 30 are categorized as KES (denoted by blue squares), 4 are categorized as KES and target site exposure (TSE) (denoted by blue/green squares), and 3 are categorized as TSE (denoted by green squares). In this figure, the 64 KESs are arranged vertically according to their level of organization.

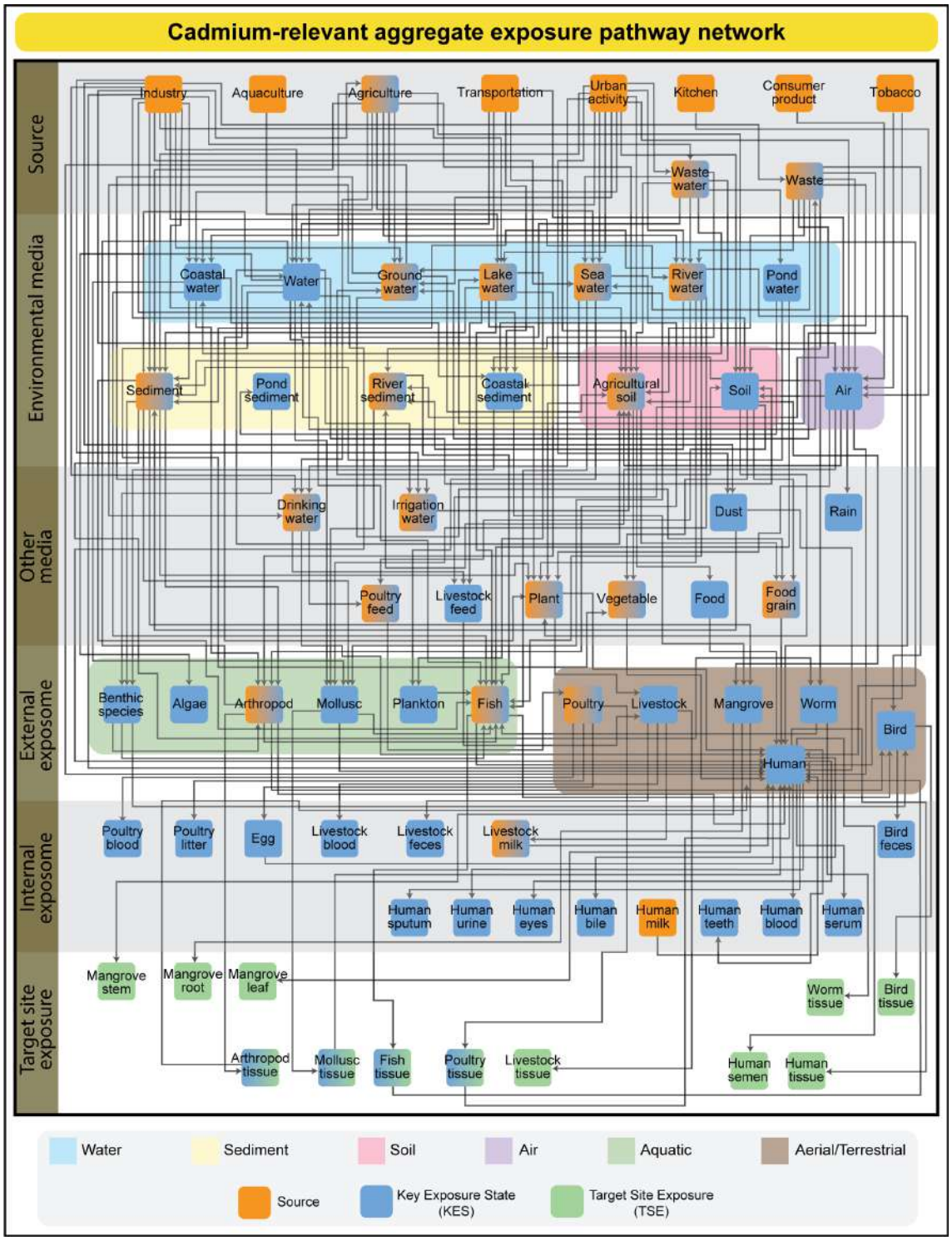


Figure 4.5: Aggregate Exposure Pathway (AEP) network corresponding to cadmium in the Indian context, comprising 73 nodes, or key exposure states (KESs), and 204 edges, or key transitional relationships (KTRs). Among the 73 KESs, 8 are categorized as Source (denoted by orange squares), 20 are categorized as both Source and KES (denoted by orange/blue squares), 33 are categorized as KES (denoted by blue squares), 4 are categorized as KES and target site exposure (TSE) (denoted by blue/green squares), and 8 are categorized as TSE (denoted by green squares). In this figure, the 73 KESs are arranged vertically according to their level of organization.

articles, and constructed an AEP network comprising 73 KESs and 204 KTRs (Figure 4.5; Supplementary Tables S4.2-S4.3). In this AEP network, we observed that the KES 'Industry' emerged as a major anthropogenic source of cadmium contamination across India, with the majority of articles reporting it across various regions (Supplementary Table S4.2). Furthermore, we observed that the KES 'Human' has been documented as the primary receptor of cadmium across multiple sources. Several studies highlighted cadmium accumulation in agricultural soils and subsequent uptake in staple crops, particularly rice, which represents a significant dietary exposure pathway in India [215, 255–257]. Additionally, multiple studies reported cadmium presence in surface waters downstream of industrial clusters, pointing to a geographically widespread contamination route that can impact both human populations and aquatic ecosystems (Supplementary Table S4.2).

4.2.3 Stressor-AOP network for inorganic arsenic and cadmium

To explore the mechanisms of toxicity underlying such chemical exposures, we constructed a stressor-AOP network linking inorganic arsenic and cadmium to their associated curated AOPs through a systematic data integration approach (Methods; Supplementary Table S4.9). We observed that only a limited number of studies provided chemical speciation information during AEP construction, while most studies focused on total arsenic and cadmium concentrations (Supplementary Table S4.2). Therefore, we unified all endpoints associated with different arsenic and cadmium compounds to obtain inorganic arsenic- and cadmium-specific information, resulting in 51 AOPs linked to arsenic (denoted as arsenic-AOPs), and 78 AOPs to cadmium (denoted as cadmium-AOPs), with Level 5 relevance and coverage score ≥ 0.4 (Supplementary Table S4.9). Of the 35 endorsed AOPs, we observed that only 5 were linked to arsenic (Supplementary Table S4.4), while the remaining 46 AOPs were under development. Similarly, we observed that cadmium was linked to only 6 endorsed AOPs, while 5 AOPs were under review and 67 were under development (Supplementary Table S4.4). This indicates that a sole reliance on endorsed AOPs risks excluding other plausible and potentially informative pathways. In addition to the 4 AOPs

already linked to arsenic and the 4 linked to cadmium in AOP-Wiki, we observed that our data-centric approach identified 47 and 74 additional AOPs, respectively, that could also be associated with these stressors. Notably, this is the first extensive identification of arsenic associated AOPs, and especially ecotoxicologically-relevant AOPs.

Finally, we utilized these associated AOPs to construct undirected AOP networks for arsenic and cadmium, revealing 48 arsenic-AOPs (Figure 4.6; Supplementary Table S4.17) and 72 cadmium-AOPs (Figure 4.7; Supplementary Table S4.18) as comprising the largest connected component (LCC), respectively.

Chai *et al.* [174] had previously constructed AOP networks for male reproductive disorders induced by inorganic arsenic. Although this study identified 16 AOPs associated with arsenic, not all were high quality and complete AOPs, and the study relied on toxicity endpoints from a single database. In comparison, our current workflow utilized integrative data-centric approach to identify human- and ecotoxicologically-relevant endpoints from six databases, expanding the associations of arsenic to 310 curated AOPs with varying levels of relevance and coverage scores. In our earlier work on inorganic cadmium [55], we relied solely on mammalian-centric databases to identify 30 curated AOPs, within which the LCC comprised 18 cadmium-AOPs. In comparison, we observed that our current workflow retained these associations and expanded to identify 78 curated AOPs, with the LCC comprising 72 cadmium-AOPs. This significant expansion underscores the importance of integrating ecotoxicological data alongside mammalian data, thereby providing a more comprehensive understanding of the underlying toxicity mechanisms.

Thereafter, we constructed directed AOP networks corresponding to the LCCs in the arsenic-AOPs (comprising 152 unique KEs and 243 unique KERs) (Figures 4.8-4.10; Supplementary Table S4.19) and cadmium-AOPs undirected networks (comprising 263 unique KEs and 422 unique KERs) (Figures 4.11-4.13; Supplementary Table S4.20) to gain deeper insights into their respective toxicity mechanisms. Further, we computed different node-centric network measures for the resulting directed arsenic-AOP and cadmium-AOP networks.

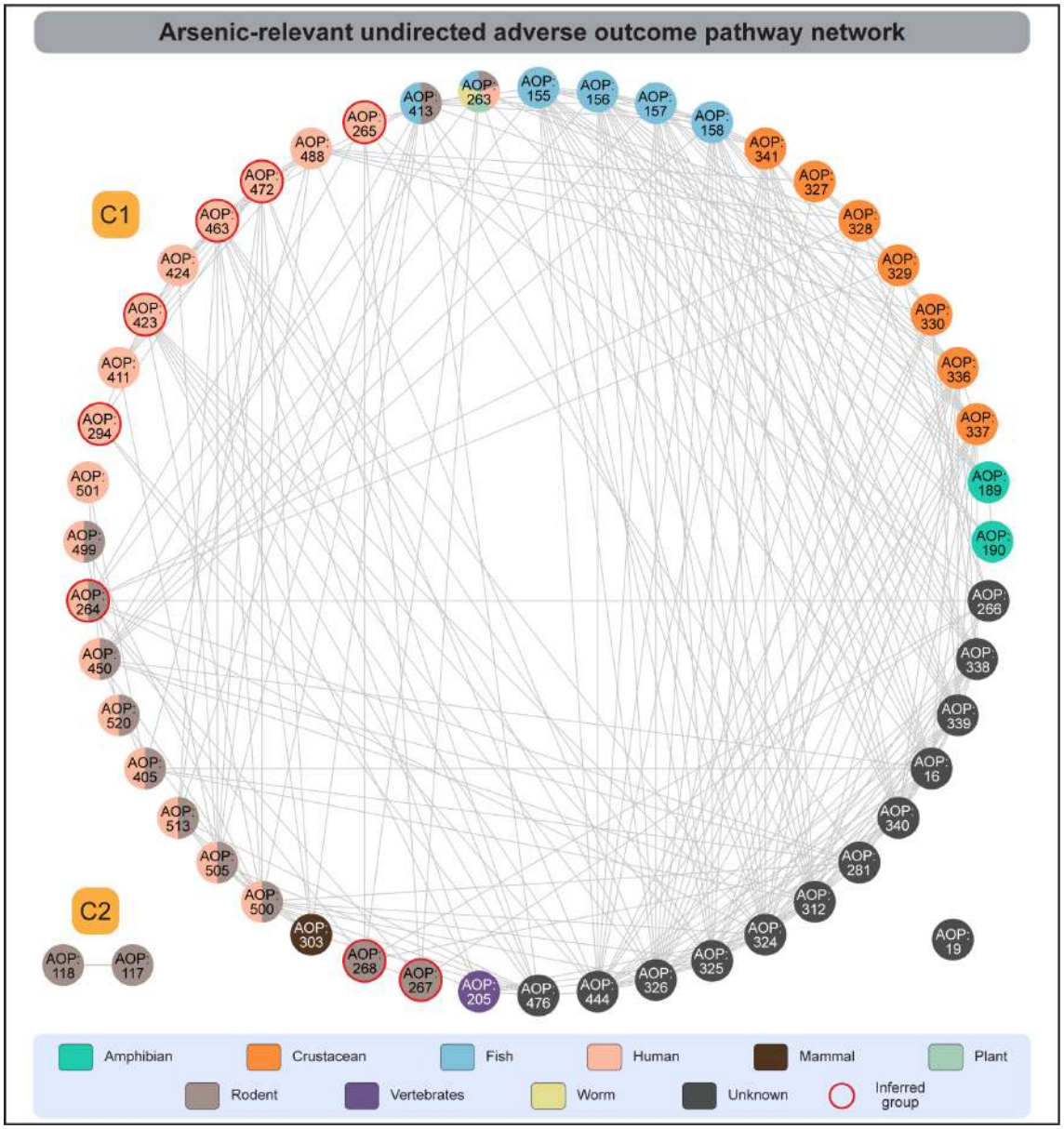


Figure 4.6: Undirected network of 51 Adverse Outcome Pathways (AOPs) linked to inorganic arsenic (denoted as arsenic-AOPs). Each node corresponds to an arsenic-AOP, and an edge between two nodes denotes that the two arsenic-AOPs share at least one Key Event (KE). This undirected network has 2 connected components (with two or more nodes) which are labeled as C1 and C2, and 1 isolated node. Each arsenic-AOP is colored based on the taxonomic group of applicability, and circled in red border if the applicability was inferred in this chapter.

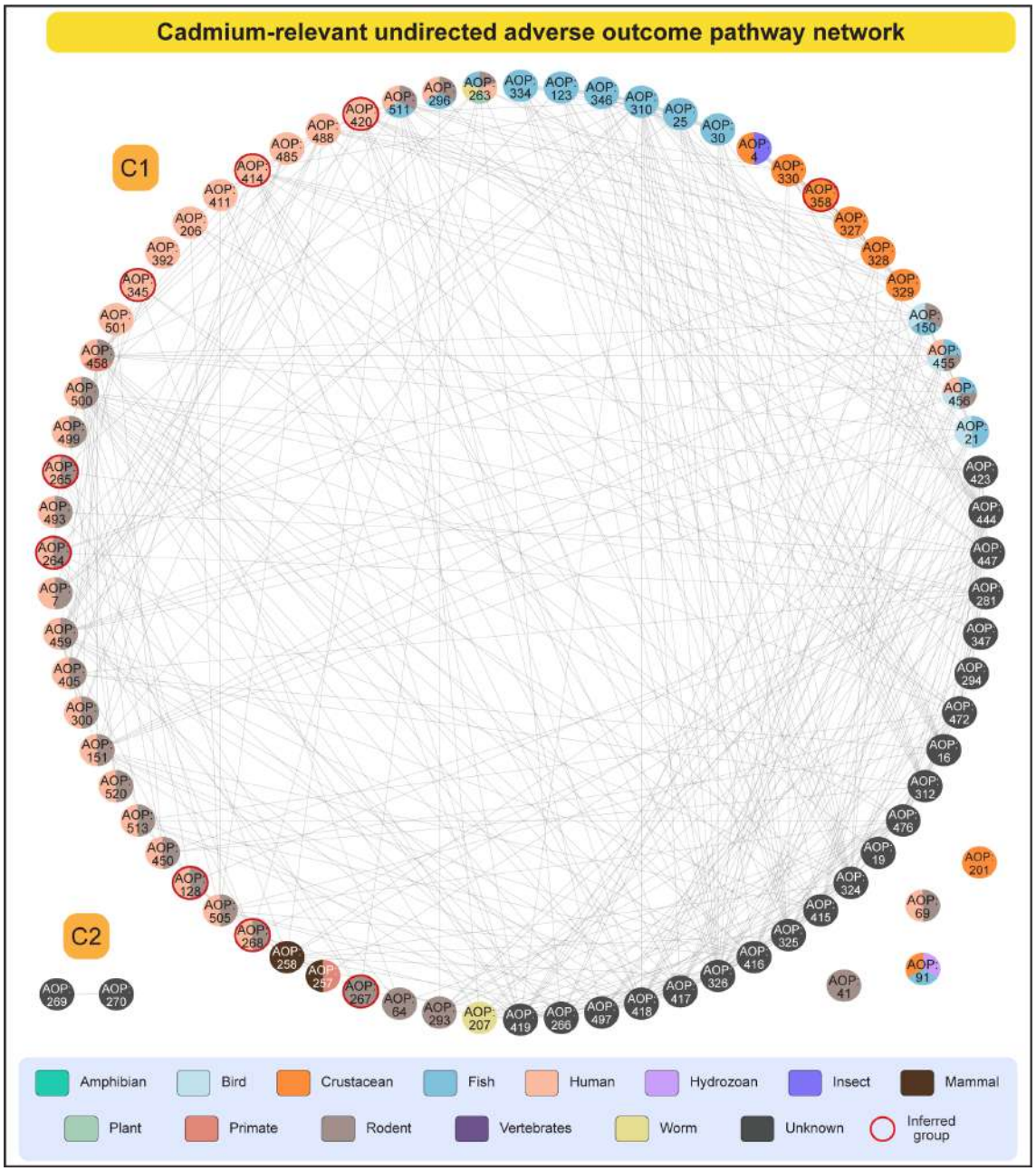


Figure 4.7: Undirected network of 78 Adverse Outcome Pathways (AOPs) linked to inorganic cadmium (denoted as cadmium-AOPs). Each node corresponds to a cadmium-AOP, and an edge between two nodes denotes that the two cadmium-AOPs share at least one Key Event (KE). This undirected network has 2 connected components (with two or more nodes) which are labeled as C1 and C2, and 4 isolated nodes. Each cadmium-AOP is colored based on the taxonomic group of applicability, and circled in red border if the applicability was inferred in this chapter.

In the directed arsenic-AOP network, we observed that the MIE 'N/A, mitochondrial dysfunction 1' (KE:177) had the highest out-degree of 9, while the AOs 'Increased mortality' (KE:351) and 'Decrease, population growth rate' (KE:360) had the highest in-degree of 9 (Supplementary Table S4.19). Further, we observed that the MIE 'Increased, reactive oxygen species' (KE:1115) had the highest betweenness centrality value, signifying that several toxicity pathways are passing through it in this network (Figure 4.9) [53], while the KE 'Tissue resident cell activation' (KE:1492) had the highest eccentricity, denoting that it is the most remotely placed KE in this network (Figure 4.10) [258].

In the directed cadmium-AOP network, we observed that the MIE 'N/A, mitochondrial dysfunction 1' (KE:177) had the highest out-degree of 13, while the KE 'Increased,vascular endothelial dysfunction' (KE:1928) and AO 'Decrease, population growth rate' (KE:360) had the highest in-degree of 9 (Supplementary Table S4.20). Further, we observed that the MIE 'N/A, mitochondrial dysfunction 1' (KE:177) had the highest betweenness centrality value, signifying that several toxicity pathways are passing through it in this network (Figure 4.12) [53], while the MIE 'DNA adduct formation' (KE:10058) and KE 'Releasing, apoptosis-inducing factor (AIF)' (KE:2031) had the highest eccentricity, denoting that they are the most remotely placed KEs in this network (Figure 4.13) [258].

We also observed that 83 of 152 KEs in the directed arsenic-AOP network were associated with inorganic arsenic-induced toxicity through this systematic data-centric approach (Supplementary Table S4.19). Similarly, we observed that 156 of 263 KEs were associated with inorganic cadmium-induced toxicity within the directed cadmium-AOP network (Supplementary Table S4.20). To further support KEs within directed arsenic-AOP and cadmium-AOP networks that are not linked to the stressors through our systematic approach, we used natural language processing (NLP)-based tools like Abstract Sifter [259] and AOP-HelpFinder [260, 261] and identified associated published evidence (Supplementary Tables S4.19-S4.20). We further utilized such evidence to infer novel taxonomic applicability of the associated AOPs.

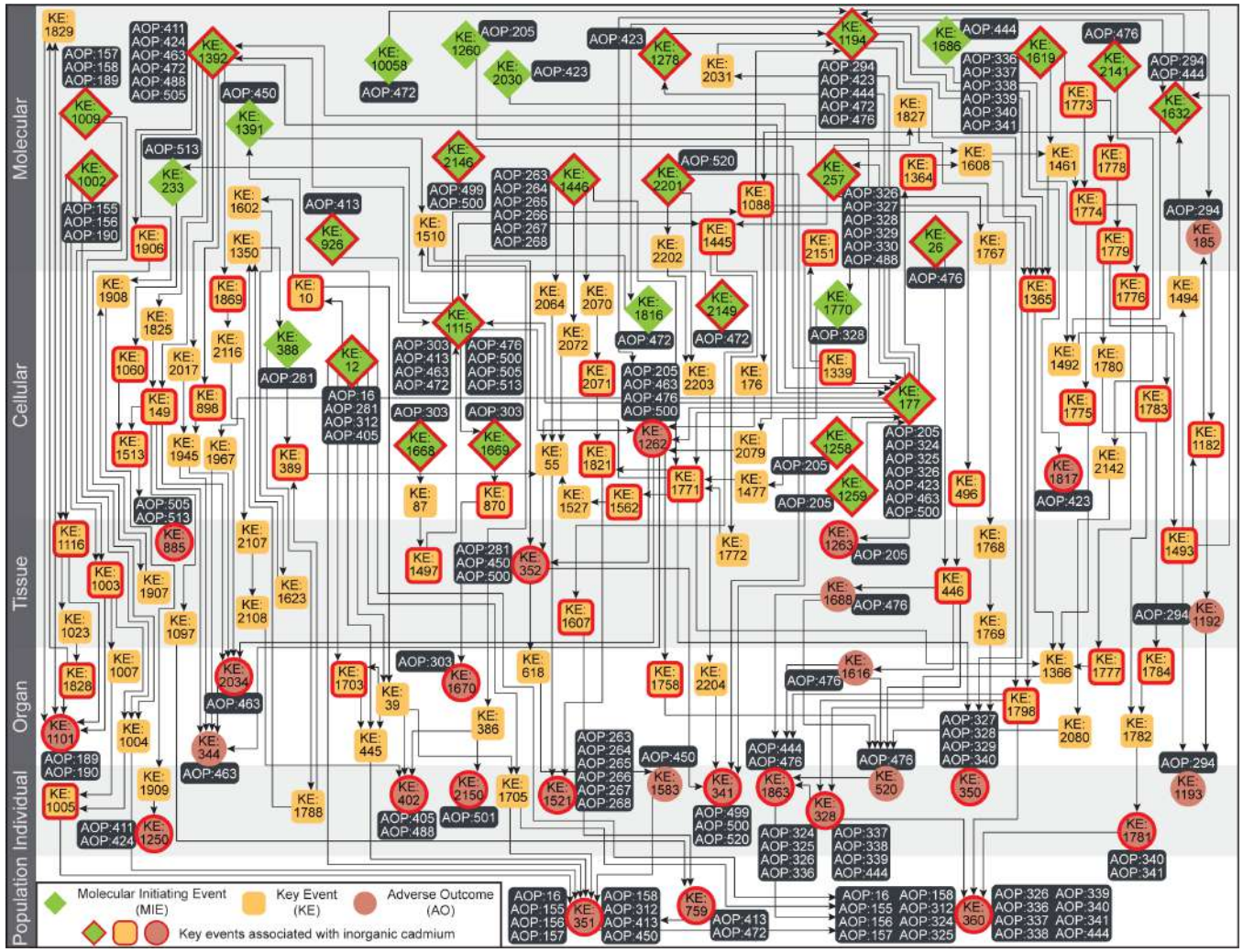


Figure 4.8: Directed network corresponding to the largest connected component (C1) in the directed arsenic-AOP network comprising 152 KEs and 243 KERs. Among the 152 KEs, 31 are categorized as MIEs (denoted as diamond), 28 are categorized as AOs (denoted as circle), and the remaining 93 are categorized as KEs (denoted as rounded square). The 83 KEs (including MIEs and AOs) associated with inorganic arsenic are marked in 'red'. In this figure, the 152 KEs are arranged vertically according to their level of biological organization.

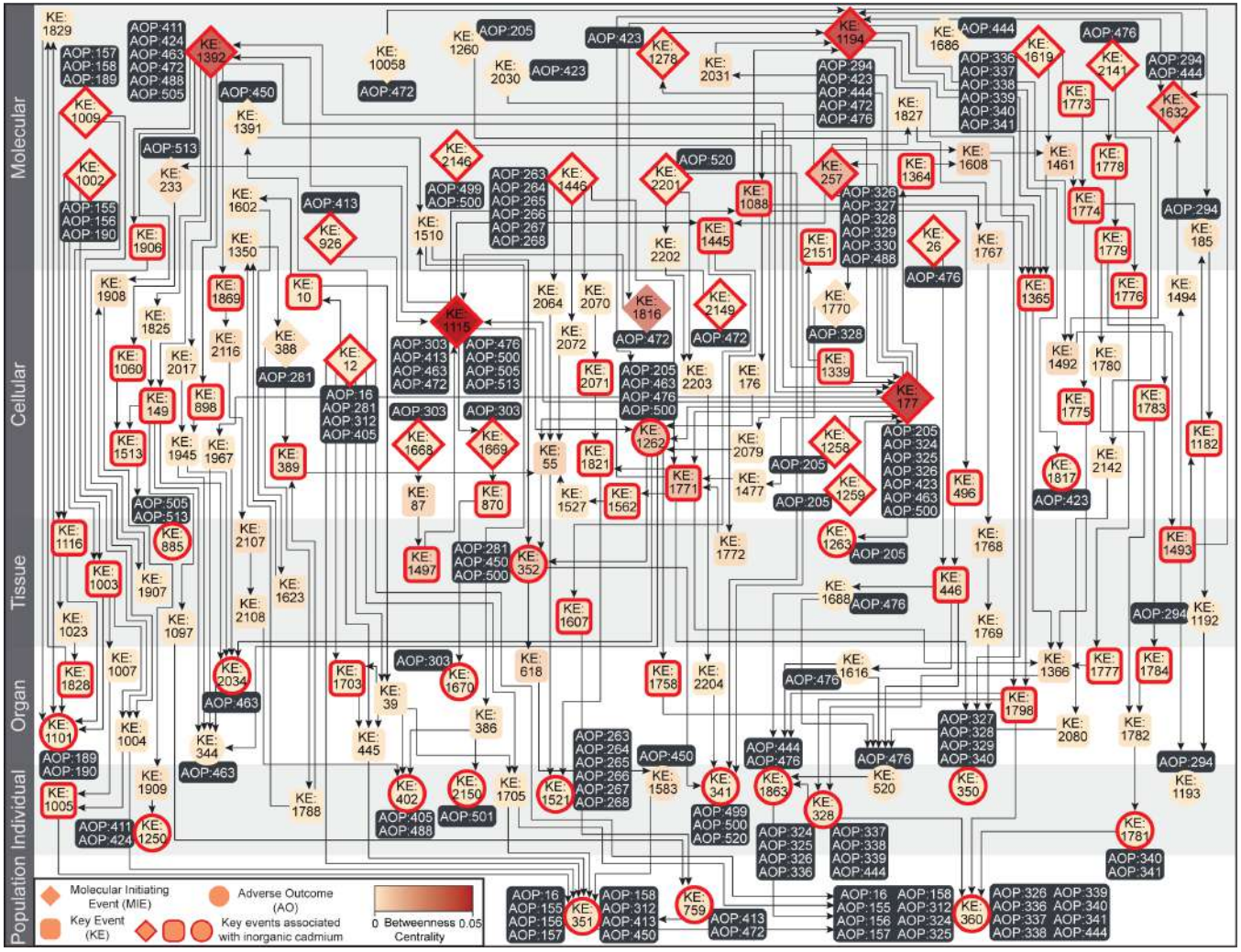


Figure 4.9: Directed network corresponding to the largest connected component (C1) in the arsenic-AOP network, where the KEs (including MIEs and AOs) are colored based on their betweenness centrality values. The 83 KEs (including MIEs and AOs) associated with inorganic arsenic are marked in 'red'. In this figure, the 152 KEs are arranged vertically according to their level of biological organization.

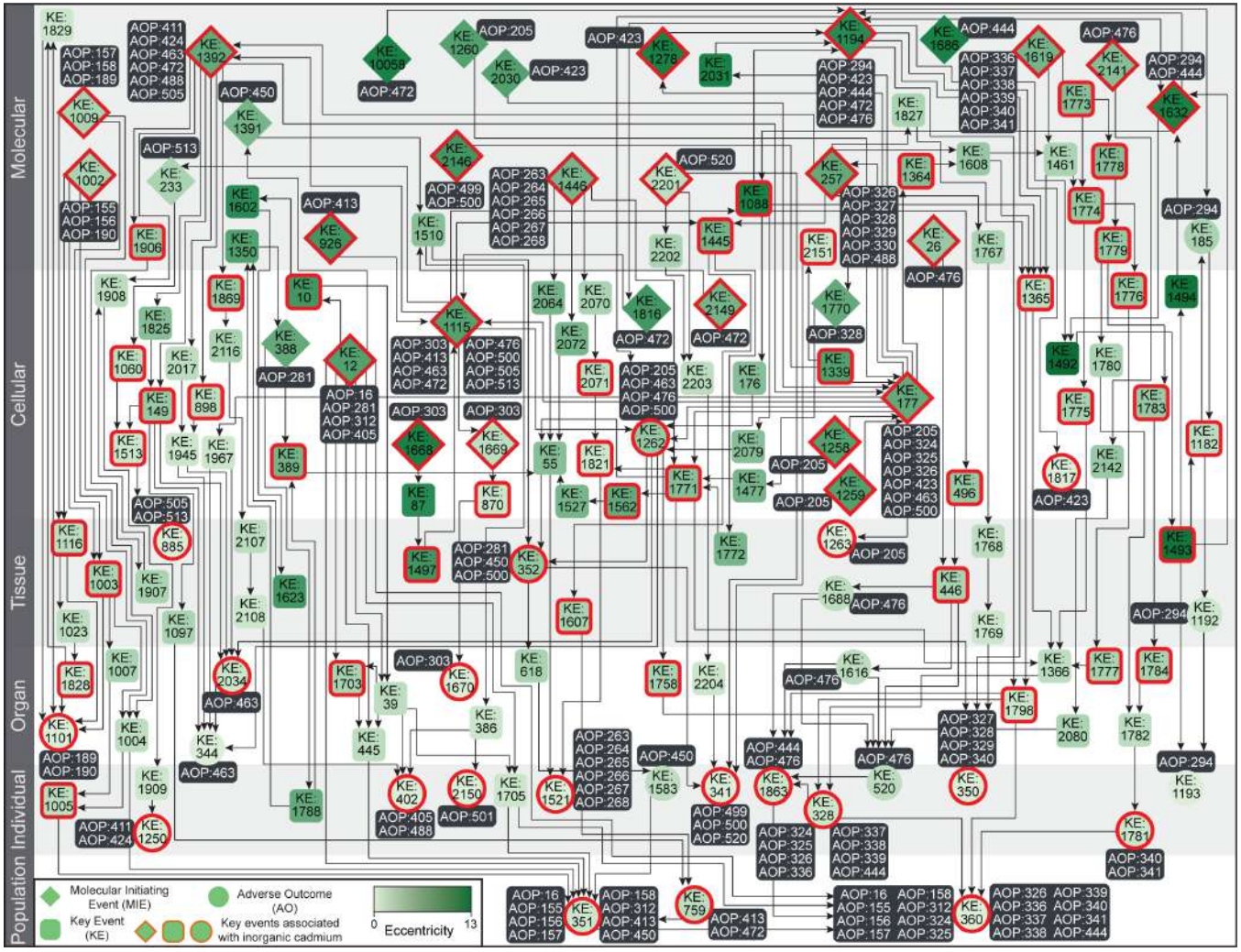


Figure 4.10: Directed network corresponding to the largest connected component (C1) in the arsenic-AOP network, where the KEs (including MIEs and AOs) are colored based on their eccentricity values. The 83 KEs (including MIEs and AOs) associated with inorganic arsenic are marked in 'red'. In this figure, the 152 KEs are arranged vertically according to their level of biological organization.

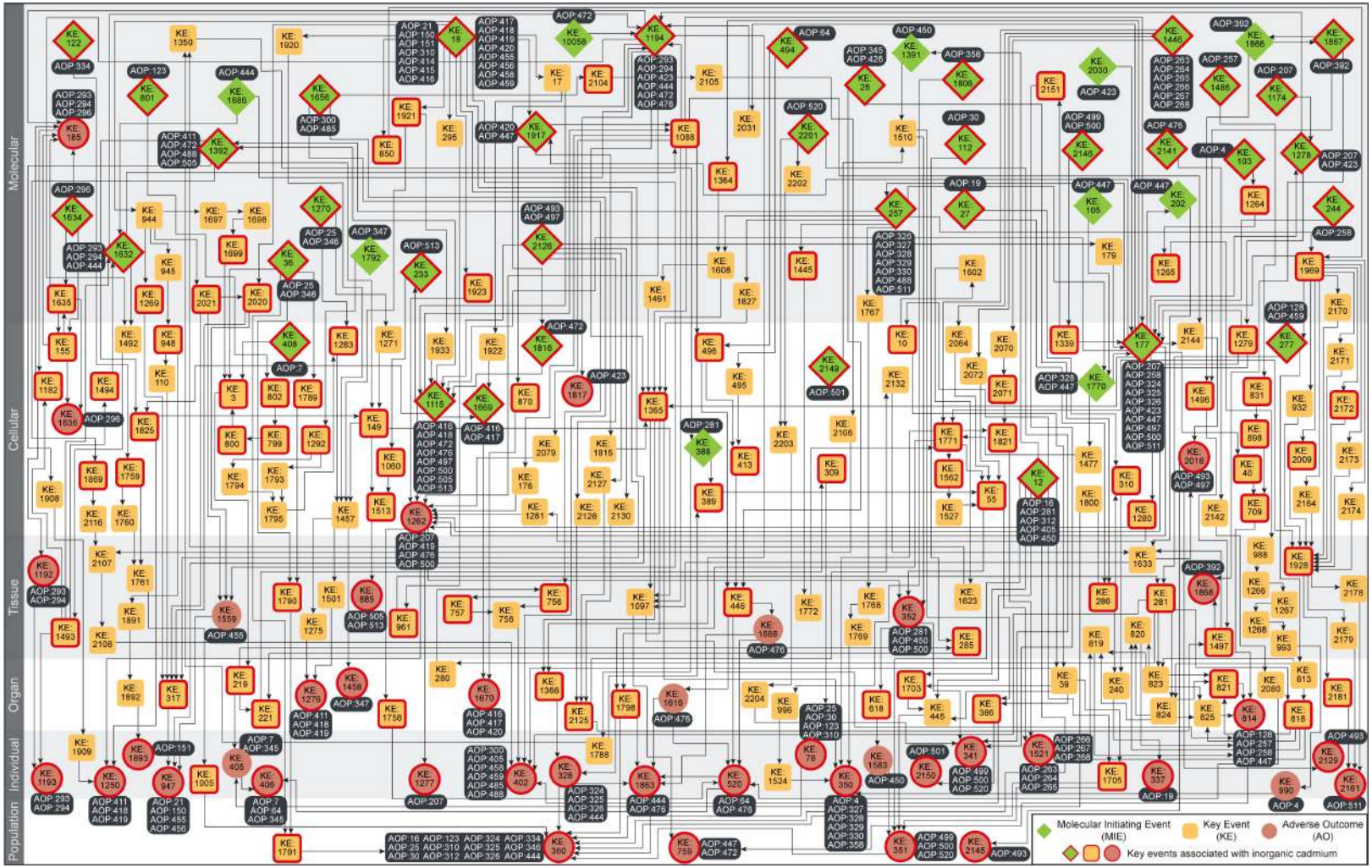


Figure 4.11: Directed network corresponding to the connected component (C1) in the directed cadmium-AOP network comprising 263 KEs and 422 KERs. Among the 263 KEs, 47 are categorized as MIEs (denoted as diamond), 41 are categorized as AOs (denoted as circle), and the remaining 175 are categorized as KEs (denoted as rounded square). The 156 KEs (including MIEs and AOs) associated with inorganic cadmium are marked in 'red'. In this figure, the 263 KEs are arranged vertically according to their level of biological organization.

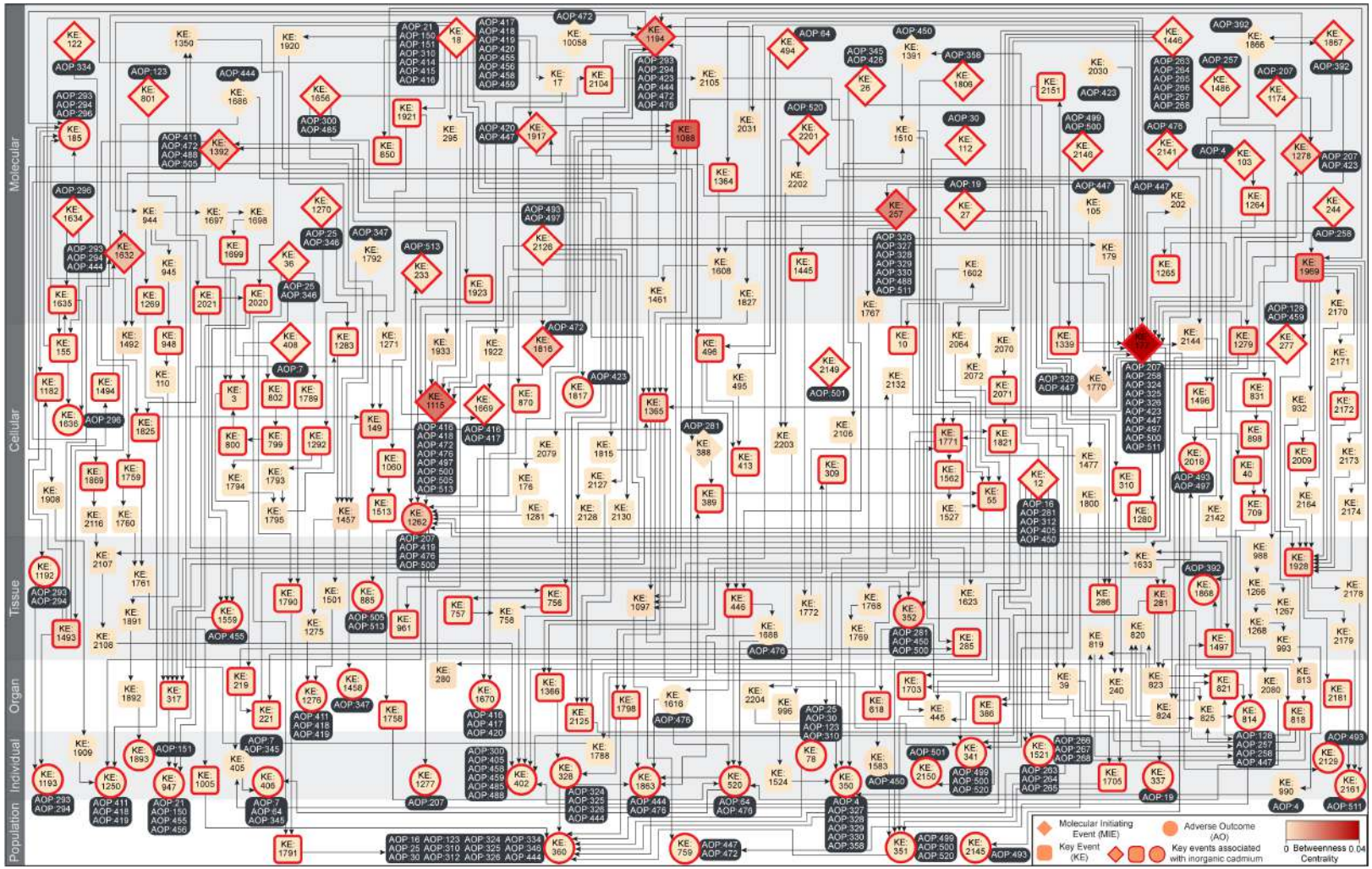


Figure 4.12: Directed network corresponding to the largest connected component (C1) in the cadmium-AOP network, where the KEs (including MIEs and AOs) are colored based on their betweenness centrality values. The 156 KEs (including MIEs and AOs) associated with inorganic cadmium are marked in 'red'. In this figure, the 263 KEs are arranged vertically according to their level of biological organization.

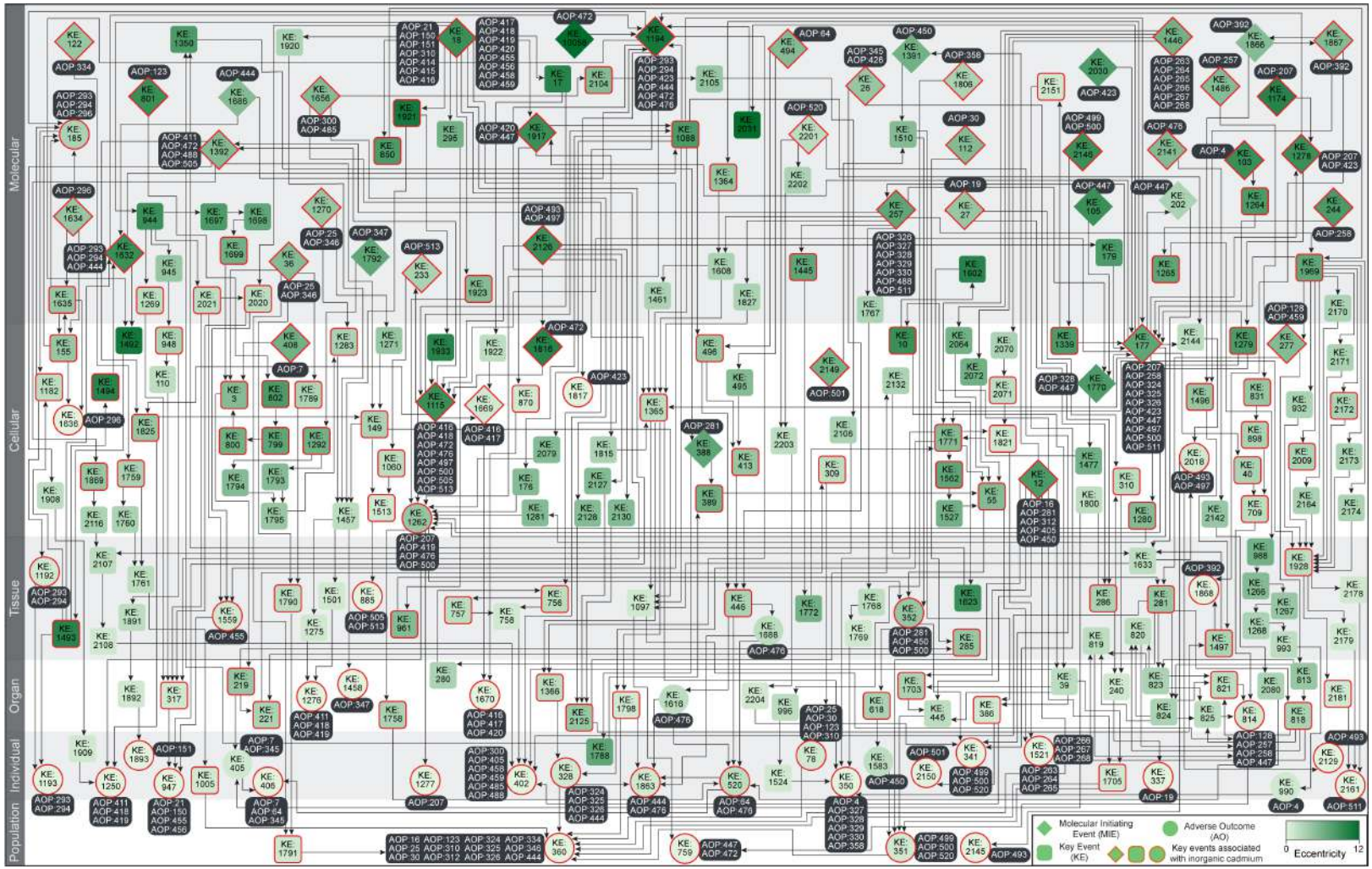


Figure 4.13: Directed network corresponding to the largest connected component (C1) in the cadmium-AOP network, where the KEs (including MIEs and AOs) are colored based on their eccentricity values. The 156 KEs (including MIEs and AOs) associated with inorganic cadmium are marked in 'red'. In this figure, the 263 KEs are arranged vertically according to their level of biological organization.

4.2.4 Taxonomic applicability of arsenic- and cadmium-AOPs

The taxonomic domain of applicability of an AOP is essential for regulatory decision-making, particularly when assessing the protection of untested species [262]. We observed that AOP-Wiki cataloged taxonomic applicability information for 28 of 48 arsenic-AOPs in the LCC of the arsenic-AOP network (Supplementary Table S4.4), and 42 of 72 cadmium-AOPs in the LCC of the cadmium-AOP network (Supplementary Table S4.4). To better understand the applicability of the remaining AOPs, we obtained the taxonomic information for each KE within the arsenic- and cadmium-AOPs from both AOP-Wiki (Supplementary Table S4.5) and the corresponding auxiliary evidence related to arsenic- and cadmium-induced toxicity (Supplementary Tables S4.19-S4.20). We then used the taxonomic information to compute the taxonomic coverage or AOPs lacking direct taxonomic annotation. Here, we defined taxonomic coverage as the ratio of KEs annotated with a given taxonomic group to the total number of KEs within that AOP.

We observed that eight arsenic-AOPs (Supplementary Table S4.21) and eight cadmium-AOPs (Supplementary Table S4.22) had a taxonomic coverage of 1 for at least one taxonomic group. This indicates that all KEs within these AOPs belong to that group, thereby inferring potential taxonomic applicability for such AOPs. Figure 4.6 shows the taxonomic applicability of arsenic-AOPs, while Figure 4.7 shows that of cadmium-AOPs. In both the figures, the AOPs with inferred taxonomic applicability are highlighted in 'red'.

4.2.5 Exploring AEP-AOP constructs associated with inorganic arsenic and cadmium

Integrating AEPs with AOPs provides a powerful framework to link sources of environmental contaminants through exposure, internal dose, and mechanistic toxicity, thereby improving the mechanistic basis for risk assessment and highlighting key data gaps [44,45]. Therefore, we constructed AEP–AOP networks for specific pathways involving arsenic and cadmium, and conducted case studies to illustrate how exposure, transport, and effect

measurements combine to yield insights into real-world risks (Figure 4.14).

AEP-AOP construct linking arsenic exposure to decreased lung function in humans

From the AEP network, we observed that industrial sources such as brick factory, thermal power plants, and lead factory release arsenic into the air (Supplementary Table S4.2), which can further reach humans either through direct inhalation or dermal contact. Interestingly, Zhang *et al.* [263] noted that atmospheric arsenic concentrations showed the maximum increase during 2005-2015 in India. In the Indian context, only limited studies have explored the inhalation/exposure pathway of arsenic through air (Supplementary Table S4.2). However, other studies suggest inhaled arsenic can be absorbed into the lungs [264, 265]. While no studies have demonstrated the bioaccumulation of arsenic in human lungs in the Indian context, human blood has often been used as a surrogate marker to assess arsenic accumulation [266]. Therefore, we selected the AOP leading to impaired lung function as a natural extension of this AEP, to gain mechanistic insights into the molecular and cellular processes underlying arsenic-induced pulmonary toxicity.

Arsenic exposure has been shown to induce oxidative stress in both human *in vitro* studies and *in vivo* rat models [267, 268]. Bomberger *et al.* [269] demonstrated that arsenic promotes degradation of the cystic fibrosis transmembrane conductance regulator (CFTR) function, leading to reduced airway surface liquid volume, decreased cilia beat frequency, and impaired mucociliary clearance in human airway epithelial cells. This impairment in clearance can cause pathogen accumulation, ultimately reducing lung function [270]. Additionally, a population-wide study found an association between arsenic exposure and impaired lung function [271]. Thus, we were able to explore a potential source-to-outcome pathway underlying arsenic-induced decrease in lung function by leveraging various published evidence (Figure 4.14a).

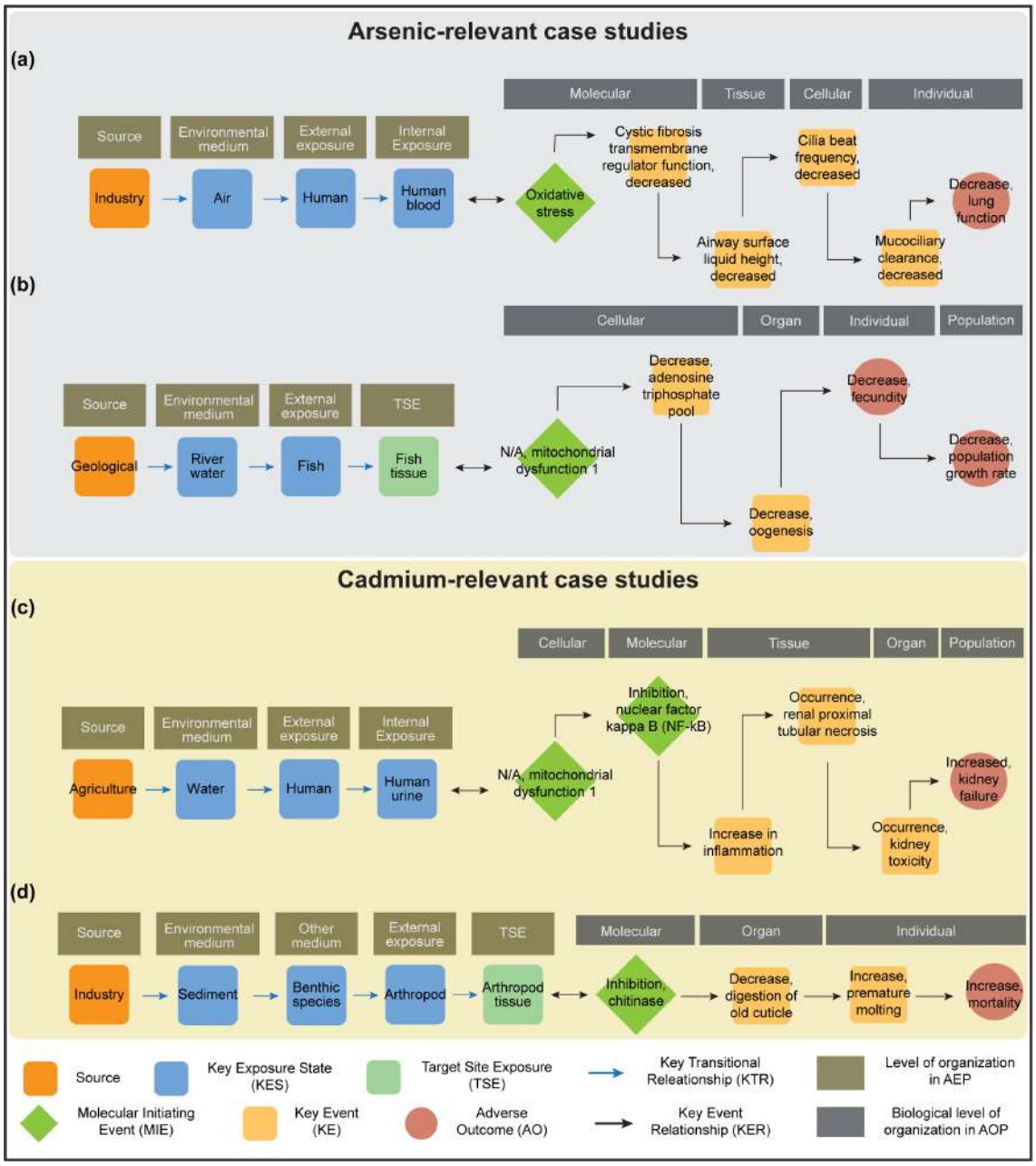


Figure 4.14: Aggregate Exposure Pathway (AEP)-Adverse Outcome Pathway (AOP) constructs explored in case studies of arsenic- and cadmium-induced toxicities. (a) AEP-AOP construct illustrating the pathway of arsenic exposure originating from industry and leading to decreased lung function in humans. (b) AEP-AOP construct illustrating the pathway of arsenic exposure originating from geological source and leading to decline in population growth rate in aquatic organisms. (c) AEP-AOP construct illustrating the pathway of cadmium exposure originating from agricultural source and leading to kidney failure in humans. (d) AEP-AOP construct illustrating the pathway of cadmium exposure originating from industry and leading to increased mortality in aquatic organisms.

AEP-AOP construct linking arsenic exposure to decreased population growth in aquatic organisms

From the AEP network, we observed that arsenic originating from geological sources can leach into river systems, where it accumulates in aquatic organisms such as fish and becomes detectable in their tissues, providing evidence of bioaccumulation along the exposure pathway (Figure 4.4). Building on this exposure context, arsenic exposure has been shown to disrupt mitochondrial membrane potential in PLHC-1 fish cells, with significant damage observed at various concentrations and exposure durations [272]. Additionally, arsenic exposure has been shown to impair ATP synthesis, resulting in a reduced ATP pool in Mediterranean polychaete species [273]. Arsenic exposure has also been found to impair oocyte development in freshwater fish from India, with an increased number of atretic follicles observed, ultimately leading to reduced fecundity [274]. Moreover, Zhang *et al.* [275] highlighted the detrimental effects of arsenic on the survival of various aquatic organisms through a comprehensive literature review. Thus, we were able to explore a potential source-to-outcome pathway underlying arsenic-induced decline in population growth among aquatic organisms by leveraging various published evidence (Figure 4.14b).

AEP-AOP construct linking cadmium exposure to kidney failure in humans

In India, chronic kidney disease (CKD) has been reported in several agricultural belts where pesticide use is widespread [255, 276]. From the AEP network, we observed that agricultural practices, particularly the use of pesticides, contribute to the release of cadmium into the environment, where it settles onto water surfaces that are subsequently used for drinking purposes (Figure 4.5). In addition, farmers are continuously exposed during pesticide application, increasing their body burden of cadmium (Supplementary Table S4.2). Although direct measurements of cadmium in kidney tissue were not available, elevated urinary cadmium levels in CKD patients indicated significant cadmium accumulation in the kidney [255, 276], consistent with its role as a critical biomarker of renal

burden [277]. These findings suggest that cadmium bioaccumulation is contributing to kidney dysfunction. Therefore, we selected the AOP leading to kidney failure as a natural extension of the AEP, to gain more mechanistic insights into the molecular and cellular processes underlying cadmium-induced nephrotoxicity.

Cadmium has been shown to induce mitochondrial dysfunction in HEK293 cells by increasing membrane permeability, reducing membrane potential, and causing mitochondrial swelling [278]. Cadmium-exposed rat kidney proximal tubular epithelial cells showed inhibition of the NF-κB signaling pathway, leading to impaired inflammatory responses [279]. Furthermore, several epidemiological studies have demonstrated a strong correlation between cadmium exposure and renal damage in humans [280, 281]. Thus, we were able to explore a potential source-to-outcome pathway underlying cadmium-induced kidney failure in humans by leveraging various published evidence (Figure 4.14c).

AEP-AOP construct linking cadmium exposure to increased mortality in aquatic organisms

From the AEP network, we observed that cadmium is released into aquatic environments from industrial activities, where it accumulates in sediments and is subsequently taken up by benthic species (Figure 4.5). This uptake facilitates its transfer through the food web to arthropods and other aquatic organisms, providing the exposure context for observed biological effects. These exposure conditions align with experimental observations, where cadmium-exposed fresh water crabs, *Sinopotamon henanense*, showed a dose-dependent inhibition of chitinase activity in their epidermal layers [282]. The chitinase is responsible for dissolving the chitin in the old exoskeleton, which can then be partially reabsorbed into the body and subsequently used to synthesize the new exoskeleton in crustaceans [283]. Since chitinase activity is inhibited, it leads to reduction in digestion of the old cuticle, as observed by Luo *et al.* [282]. Bodar *et al.* [284] observed that cadmium-exposed *Daphnia magna* showed altered ecdysteroid levels, leading to impairment in the molting process. Additionally, multiple studies have reported increased mortality due to cadmium exposure

in different crustacean species [284–286]. Thus, we were able to explore a potential source-to-outcome pathway underlying cadmium-induced mortality among aquatic organisms, especially crustaceans, by leveraging various published evidence (Figure 4.14d).

4.2.6 Stressor-species networks associating impacts of inorganic arsenic and cadmium compounds with ecologically relevant species

A stressor-species network reveals the intricate interactions between environmental stressors and their effects on ecosystems [57,60,61]. Here, we used the data within the ECOTOX knowledgebase [63] to construct three types of stressor-species networks: one based on acute toxicity data (LC_{50} and EC_{50}), one based on chronic toxicity data (NOEC) and one using BCF data to explore the impacts of inorganic arsenic and cadmium on ecosystems and identify vulnerable species.

We observed that 167 species (across 9 ECOTOX species groups) are linked to 15 inorganic arsenic compounds and 644 species (across 12 ECOTOX species groups) are linked to 8 inorganic cadmium compounds (Supplementary Table S4.11) in the acute toxicity network. We further observed that species within the 'Fish' ECOTOX group were most frequently tested for inorganic arsenic-induced acute toxicity, while 'Crustaceans' were most frequently tested for inorganic cadmium-induced acute toxicity (Figure 4.15a; Supplementary Table S4.11). Notably, sodium arsenite (CAS:7784-46-5) and cadmium chloride (CAS:10108-64-2) were documented as being toxic to the highest number of species (Supplementary Table S4.11). Additionally, we observed that *Daphnia magna* has been tested with the highest number of both arsenic and cadmium compounds (Supplementary Table S4.11).

In the stressor-species network based on chronic toxicity data, we observed that 130 species (across 12 ECOTOX species groups) are linked to 16 inorganic arsenic compounds and 264 species (across 14 ECOTOX species groups) are linked to 4 inorganic cadmium compounds (Supplementary Table S4.11). We further observed that the species within

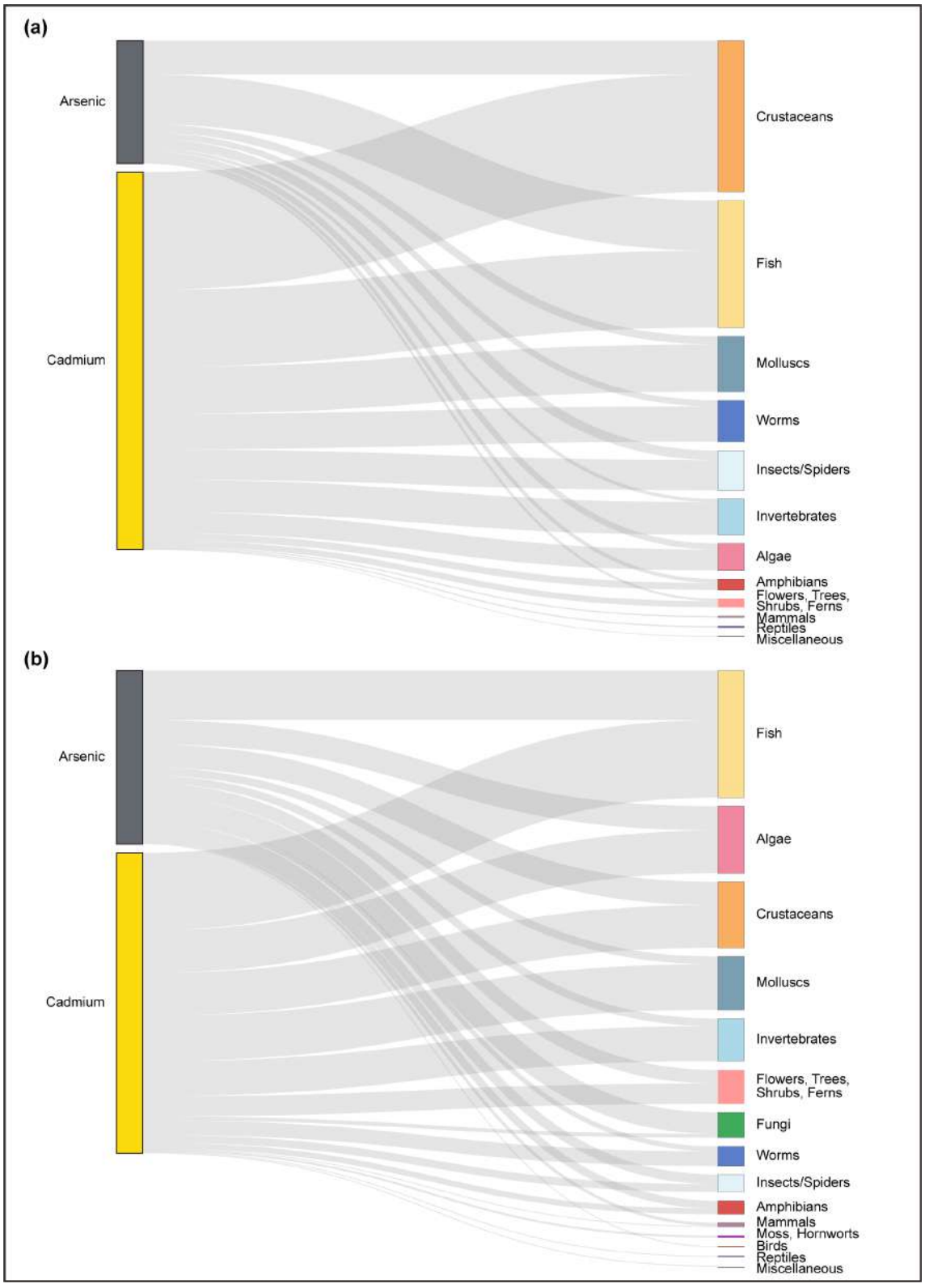


Figure 4.15: Sankey plot depicting associations between inorganic arsenic and cadmium compounds, and ECOTOX species groups through their: (a) acute toxicity concentration data (LC50 and EC50); and (b) chronic toxicity concentration data (NOEC). The plot provides associations between inorganic arsenic and cadmium with 15 ECOTOX species groups.

the ‘Fish’ ECOTOX group were most frequently tested for both inorganic arsenic-induced and cadmium-induced chronic toxicity (Figure 4.15b; Supplementary Table S4.11). Notably, sodium arsenite (CAS:7784-46-5) and cadmium chloride (CAS:10108-64-2) were documented as being toxic to the highest number of species (Supplementary Table S4.11). Additionally, we observed that *Danio rerio* has been tested with the highest number of arsenic compounds, while *Cyprinus carpio* and *Oncorhynchus mykiss* have been tested with the highest number of cadmium compounds (Supplementary Table S4.11).

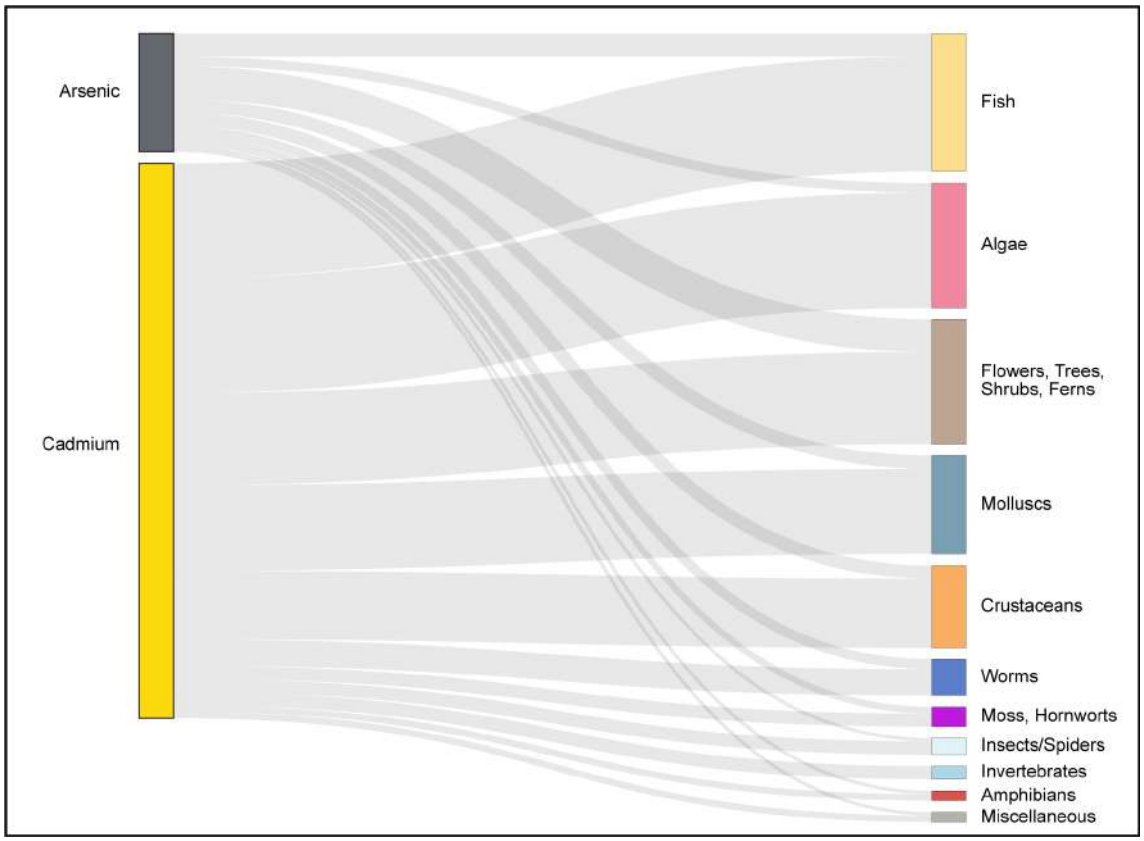


Figure 4.16: Sankey plot depicting associations between inorganic arsenic and cadmium compounds, and ECOTOX species groups through their bioconcentration factors (BCFs). The plot provides associations between inorganic arsenic and cadmium with 11 ECOTOX species groups.

Similarly, in the BCF-based network, we observed that 32 species (across 10 ECOTOX species groups) are linked to 5 inorganic arsenic compounds and 147 species (across 11 ECOTOX species groups) to 4 inorganic cadmium compounds (Supplementary Table S4.12). We further observed that species within the ‘Flowers, Trees, Shrubs, Ferns’ ECOTOX group were documented to bioaccumulate most arsenic compounds, while

'Algae' were documented to bioaccumulate most cadmium compounds (Figure 4.16; Supplementary Table S4.12). Notably, disodium arsenate (CAS:7778-43-0) and cadmium chloride (CAS:10108-64-2) were documented as having bioconcentration potential in the largest number of species (Supplementary Table S4.12). Additionally, we observed that *Daphnia magna* has been documented to absorb the highest number of arsenic compounds, while *Lemna minor* has been documented to absorb the highest number of cadmium compounds (Supplementary Table S4.12).

4.2.7 SSDs for inorganic arsenic and cadmium compounds

The SSDs are statistical tools used in ecological risk assessments to evaluate the sensitivity of various species to toxic environmental substances [57, 64, 287, 288]. They are utilized to derive hazard concentration values like HC05 [57, 64, 287, 288]. Additionally, the SSDn is a novel approach that normalizes chemical toxicity data based on a reference species, thereby improving the comparison of chemicals with similar toxicological profiles [68, 69]. Here, we computed both SSD and SSDn for inorganic arsenic and cadmium compounds based on acute and chronic toxicity data compiled in the ECOTOX knowledgebase. Further, we integrated the derived HC05 values with the toxicity concentrations in the stressor-species network, and annotated a species as sensitive to a chemical if the corresponding toxicity value was below the HC05 value. We further highlighted these sensitive species associations within the stressor-species network to provide a comprehensive view of all the sensitive species associated with inorganic arsenic and cadmium compounds. In comparison with our previously constructed stressor-species networks [57], this approach offers a novel perspective, as it integrates hazard concentrations into network visualization, thereby enhancing our understanding of the toxicity impacts across different species. We made these stressor-species visualizations accessible on the associated website: https://cb.imsc.res.in/heavymetaltox/. Figure 4.17 shows an example of the stressor-species network constructed based on chronic toxicity values associated with inorganic arsenic compounds.

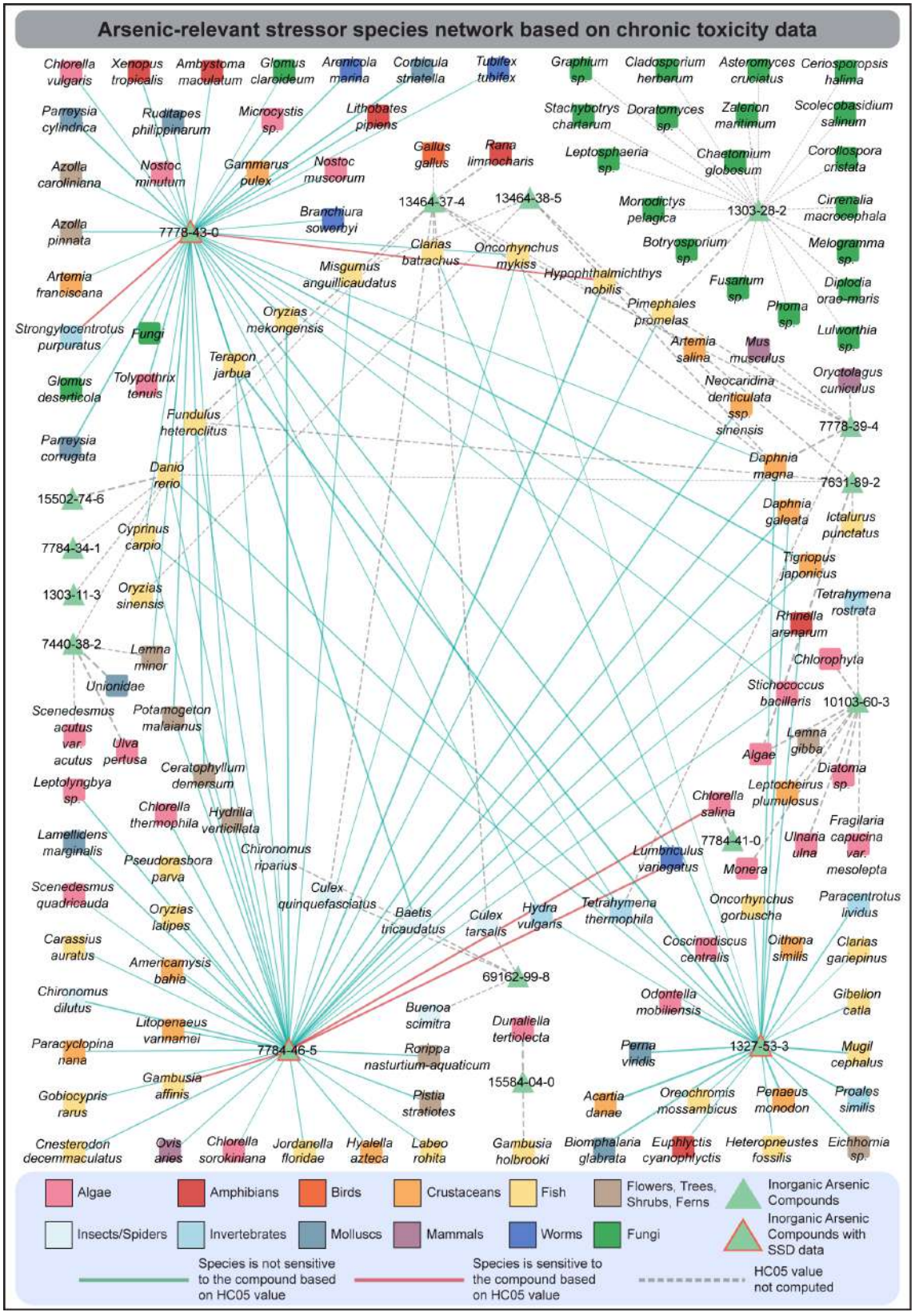

Arsenic-relevant stressor species network based on chronic toxicity data
Chlorella vulgaris
Xenopus tropicalis
Ambystoma maculatum
Glomus claroideum
Aranicola marina
Corbicula striatella
Tubifex tubifex
Graphium sp.
Cladosporium herbarum
Asteromyces cruciatus
Ceriosporopsis halima
Parreysia cylindrica
Ruditapes philippinarum
Microcystis sp.
Lithobates pipiens
Stachybotrys chartarum
Doratomyces sp.
Zalerion maritimum
Scolecobasidium salinum
Azolla caroliniana
Nostoc minutum
Gammarus pulex
Nostoc muscorum
Gallus gallus
Rana limnocharis
Leptosphaeria sp.
Chaetomium globosum
Corollospora cristata
Azolla pinnata
7778-43-0
Branchiura sowerbyi
13464-37-4
13464-38-5
Monodictys pelagica
1303-28-2
Cirrenalia macrocephala
Artemia franciscana
Clarias batrachus
Oncorhynchus mykiss
Misgurnus anguillicaudatus
Hypophthalmichthys nobilis
Botryosporium sp.
Melogramma sp.
Fusarium sp.
Diplodia orae-maris
Strongylocentrotus purpuratus
Fungi
Oryzias mekongensis
Pimephales promelas
Phoma sp.
Lulworthia sp.
Artemia salina
Mus musculus
Terapon jarbua
Glomus deserticola
Tolypothrix tenuis
Neocaridina denticulata ssp. sinensis
Oryctolagus cuniculus
Fundulus heteroclitus
7778-39-4
Parreysia corrugata
Danio rerio
Daphnia magna
7631-89-2
15502-74-6
Daphnia galeata
Ictalurus punctatus
Cyprinus carpio
7784-34-1
Tigriopus japonicus
Oryzias sinensis
1303-11-3
Tetrahymena rostrata
Rhinella arenarum
7440-38-2
Lemna minor
Chlorophyta
Unionidae
Stichococcus bacillaris
10103-60-3
Scenedesmus acutus var. acutus
Potamogeton malaianus
Lemna gibba
Algae
Ulva pertusa
Diatoma sp.
Ceratophyllum demersum
Leptocheirus plumulosus
Leptolyngbya sp.
Chlorella salina
Chlorella thermophila
Hydrilla verticillata
Fragilaria capucina var. mesolepta
Lamellidens marginalis
7784-41-0
Ulnaria ulna
Chironomus riparius
Lumbriculus variegatus
Pseudorasbora parva
Monera
Scenedesmus quadricauda
Culex quinquefasciatus
Oncorhynchus gorbuscha
Paracentrotus lividus
Oryzias latipes
Baetis tricaudatus
Culex tarsalis
Hydra vulgaris
Tetrahymena thermophila
Carassius auratus
Coscinodiscus centralis
Oithona similis
Clarias ganiepinus
Americamysis bahia
69162-99-8
Chironomus dilutus
Odontella mobiliensis
Gibelion catla
Buenoa scimitra
Litopenaeus vannamei
Dunaliella tertiolecta
Perna viridis
1327-53-3
Mugil cephalus
Paracyclopina nana
7784-46-5
Rorippa nasturtium-aquaticum
Gambusia affinis
15584-04-0
Acartia danae
Oreochromis mossambicus
Penaeus monodon
Proales similis
Gobiocypris rarus
Pistia stratiotes
Cnesterodon decemmaculatus
Ovis aries
Chlorella sorokiniana
Jordanella floridae
Hyalella azteca
Labeo rohita
Gambusia holbrooki
Biomphalaria glabrata
Euphlyctis cyanophlyctis
Heteropneustes fossilis
Eichhornia sp.
Algae
Amphibians
Birds
Crustaceans
Fish
Flowers, Trees, Shrubs, Ferns
Inorganic Arsenic Compounds
Insects/Spiders
Invertebrates
Molluscs
Mammals
Worms
Fungi
Inorganic Arsenic Compounds with SSD data
Species is not sensitive to the compound based on HC05 value
Species is sensitive to the compound based on HC05 value
HC05 value not computed

Figure 4.17 ***(previous page)*****:** Stressor-species network constructed for inorganic arsenic compounds using the chronic toxicity concentrations as edge weights. The network comprises 16 inorganic arsenic compounds, 130 species belonging to 12 ECOTOX species groups, and 179 stressor-species links. The four inorganic arsenic compounds for which SSDs were computed are highlighted with a 'red' border. A stressor-species link is shown in 'red' if the species is sensitive to the corresponding stressor, otherwise it is shown in 'blue'. A 'gray dotted' link corresponds to a stressor-species pair, where the HC05 values were not derived for the stressor.

SSDs based on acute toxicity endpoints

To construct SSDs, we used the acute toxicity endpoints for aquatic species from the ECOTOX database and derived the HC05 values for four inorganic arsenic compounds, namely arsenic oxide (CAS:1327-53-3), disodium arsenate (CAS:7778-43-0), sodium arsenite (CAS:7784-46-5), and trisodium arsenate (CAS:13464-38-5), and four inorganic cadmium compounds, namely cadmium (CAS:7440-43-9), cadmium chloride (CAS:10108-64-2), cadmium sulfate (CAS:10124-36-4), and cadmium nitrate (CAS:10325-94-7), using both the USEPA SSD Toolbox and the R-based ssdtools package (Supplementary Table S4.23). We observed that both tools consistently identified the same best-fit models, and derived similar HC05 values for all chemicals (Supplementary Table S4.23). Figures 4.18-4.25 present the SSD plots for each chemical generated using the corresponding best-fit models from both tools.

The HC05 values indicate the toxic effects of chemicals across various species, with lower HC05 values signifying a higher toxic potential [289]. Based on the model-averaged HC05 values, we observed that cadmium compounds are in general more toxic than the arsenic compounds (Supplementary Table S4.23), aligning with the findings of Karthikeyan *et al.* [287] regarding sea water quality criterion for heavy metals. Moreover, within the arsenic compounds, we observed that trivalent arsenic compounds (or arsenites), namely arsenic oxide (CAS:1327-53-3) and sodium arsenite (CAS:7784-46-5), to be more toxic than pentavalent arsenic compounds (or arsenates), namely disodium arsenate (CAS:7778-43-0) and trisodium arsenate (CAS:13464-38-5) (Supplementary Table S4.23), aligning with earlier observations [247, 290, 291].

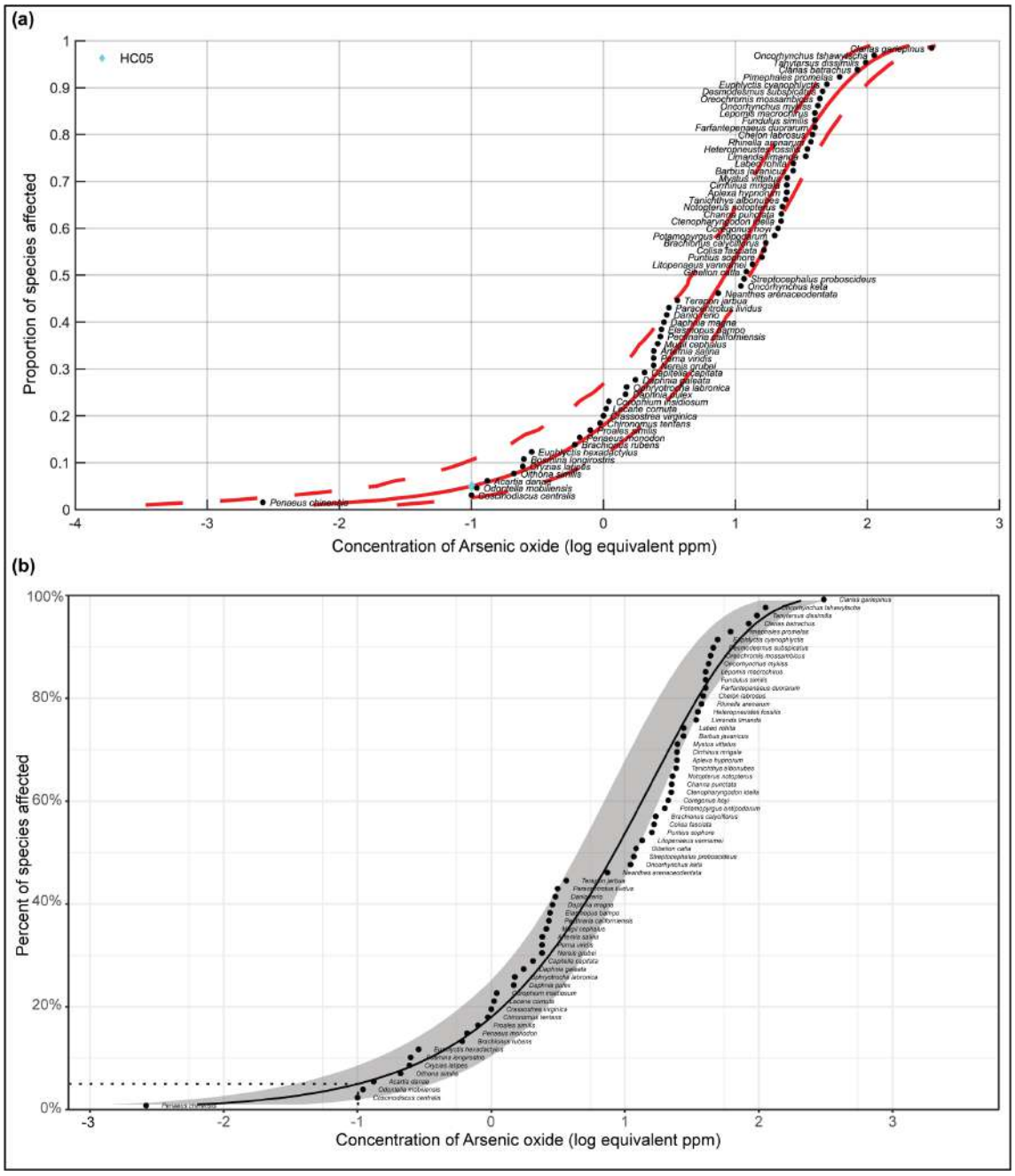


Figure 4.18: The plots of SSD for arsenic oxide (CAS:1327-53-3) based on acute toxicity data and computed using the best-fit Weibull model. (a) As determined by USEPA SSD Toolbox where the HC05 value is denoted by cyan colored diamond. (b) As determined by ssdtools where the HC05 value is denoted by a dotted line.

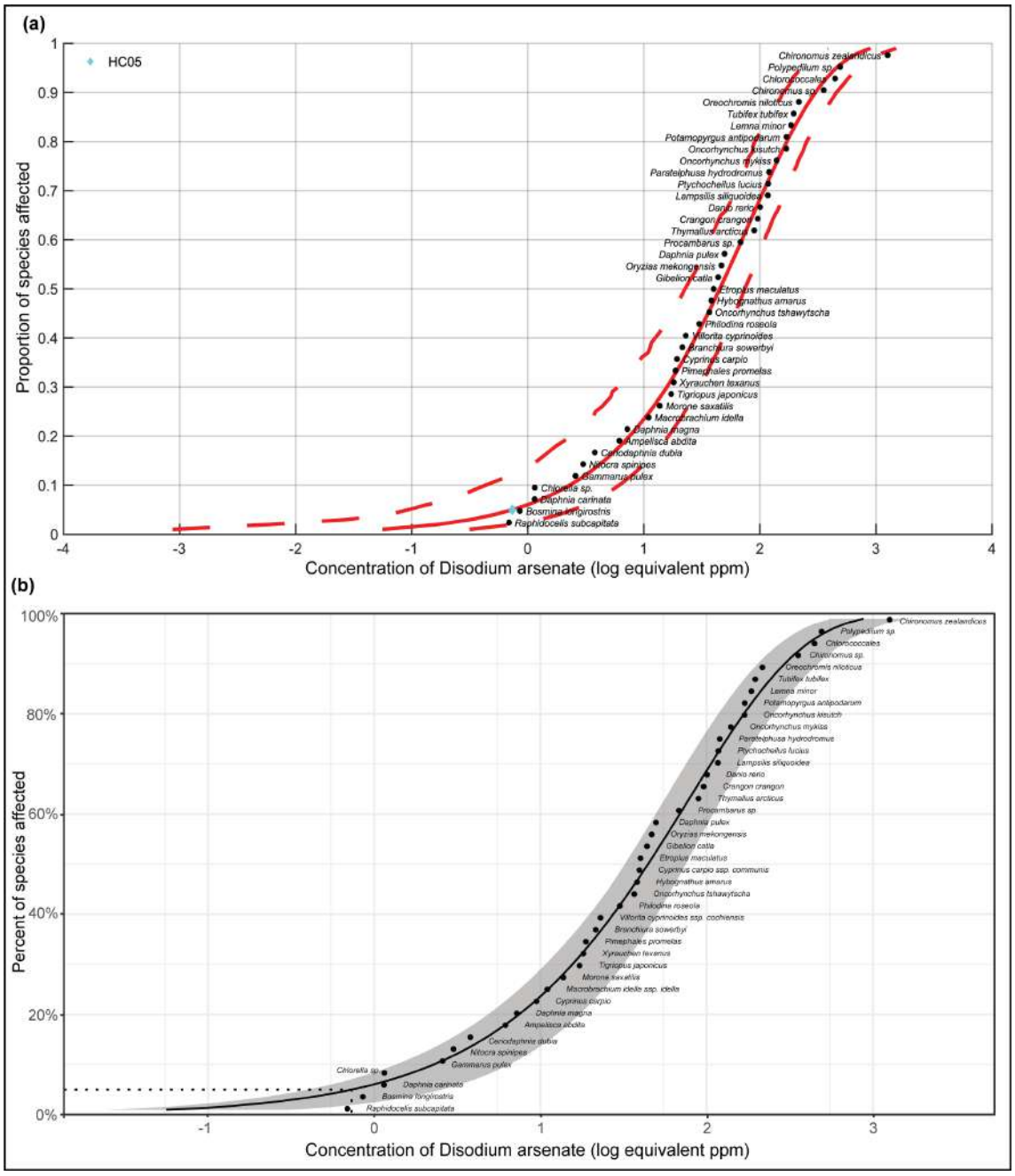


Figure 4.19: The plots of SSD for disodium arsenate (CAS:7778-43-0) based on acute toxicity data and computed using the best-fit Weibull model. (a) As determined by USEPA SSD Toolbox where the HC05 value is denoted by cyan colored diamond. (b) As determined by ssdtools where the HC05 value is denoted by a dotted line.

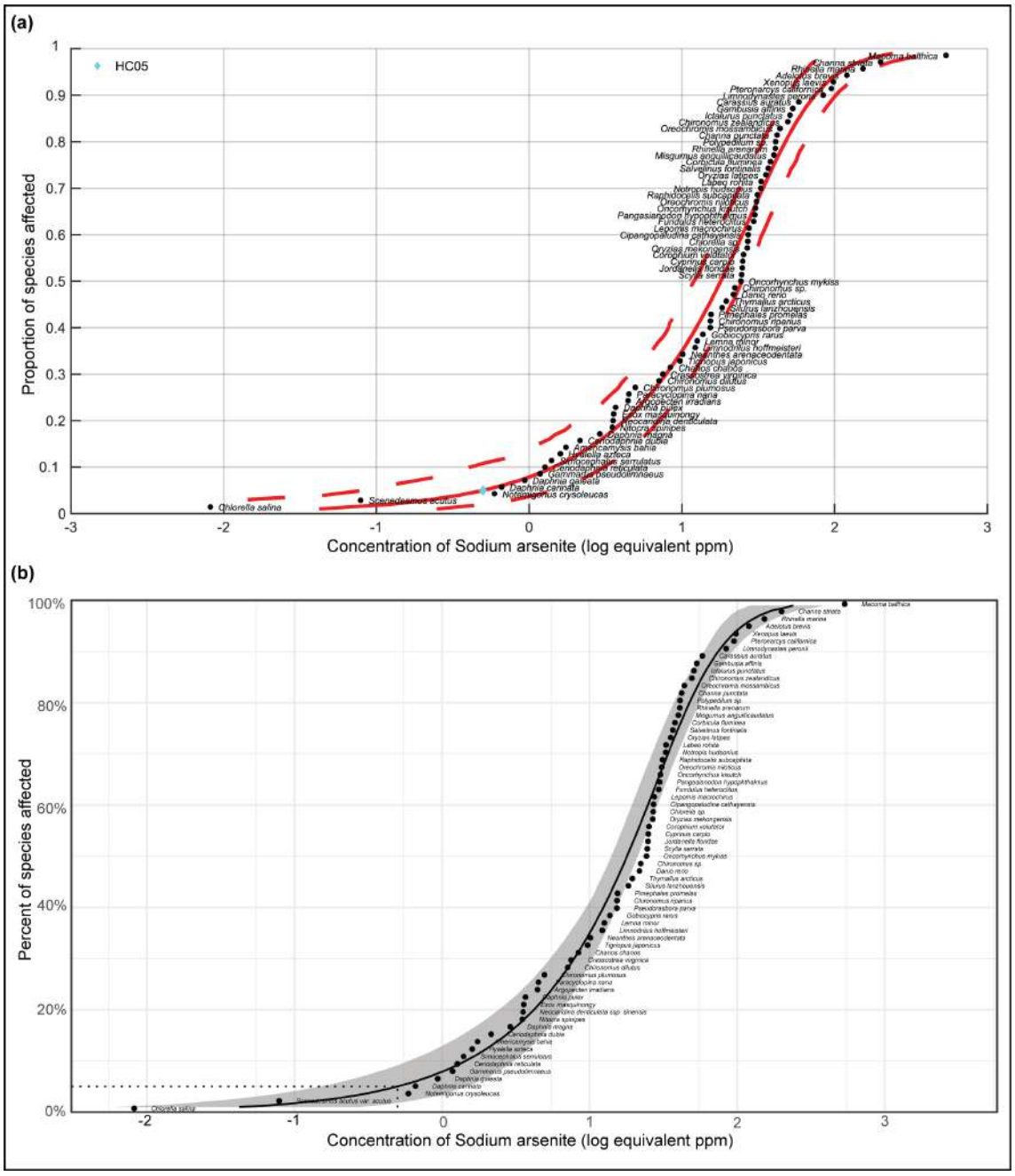


Figure 4.20: The plots of SSD for sodium arsenite (CAS:7784-46-5) based on acute toxicity data and computed using the best-fit Burr Type III model. (a) As determined by USEPA SSD Toolbox where the HC05 value is denoted by cyan colored diamond. (b) As determined by ssdtools where the HC05 value is denoted by a dotted line.

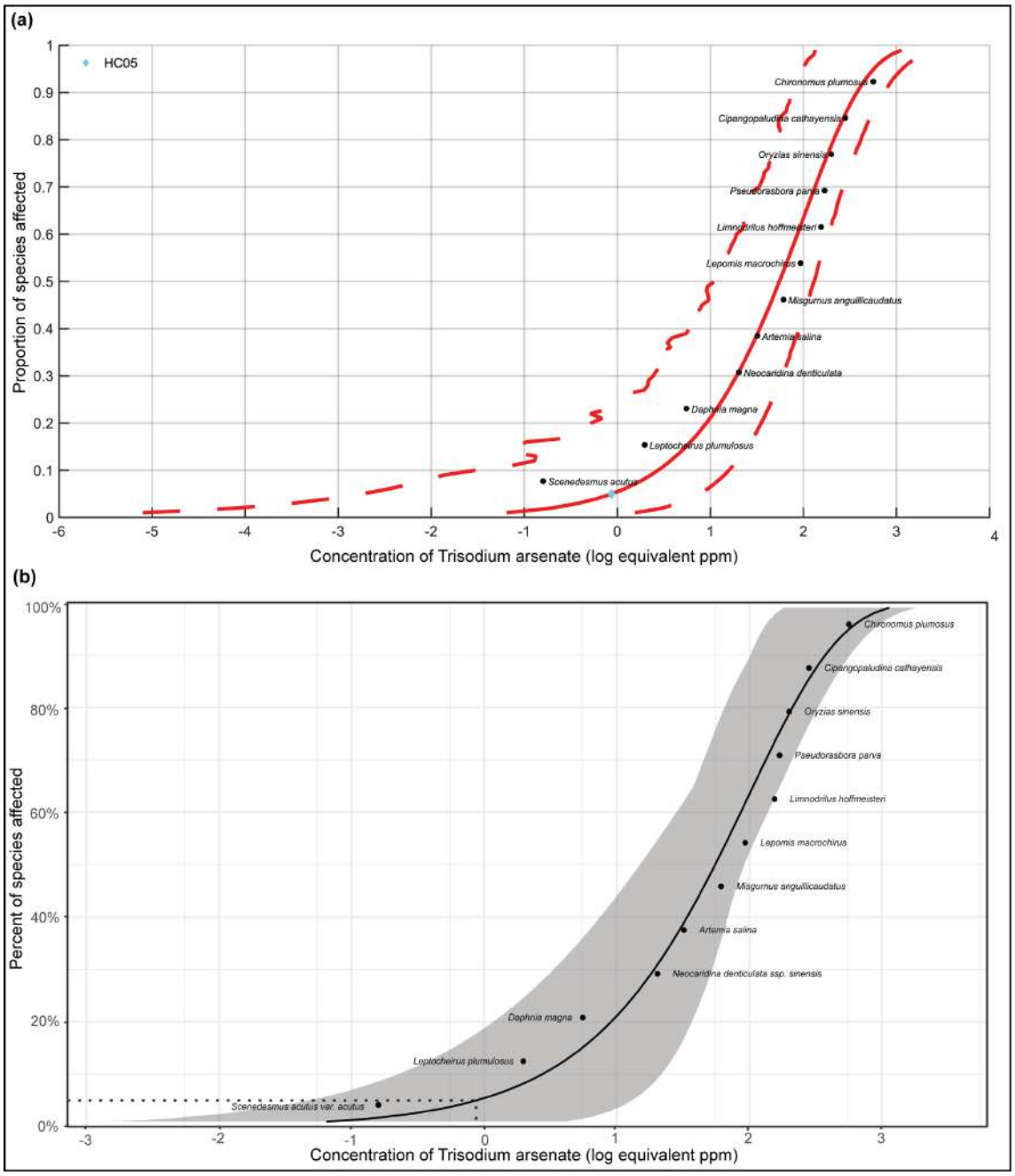


Figure 4.21: The plots of SSD for trisodium arsenate (CAS:13464-38-5) based on acute toxicity data and computed using the best-fit Weibull model. (a) As determined by USEPA SSD Toolbox where the HC05 value is denoted by cyan colored diamond. (b) As determined by ssdtools where the HC05 value is denoted by a dotted line.

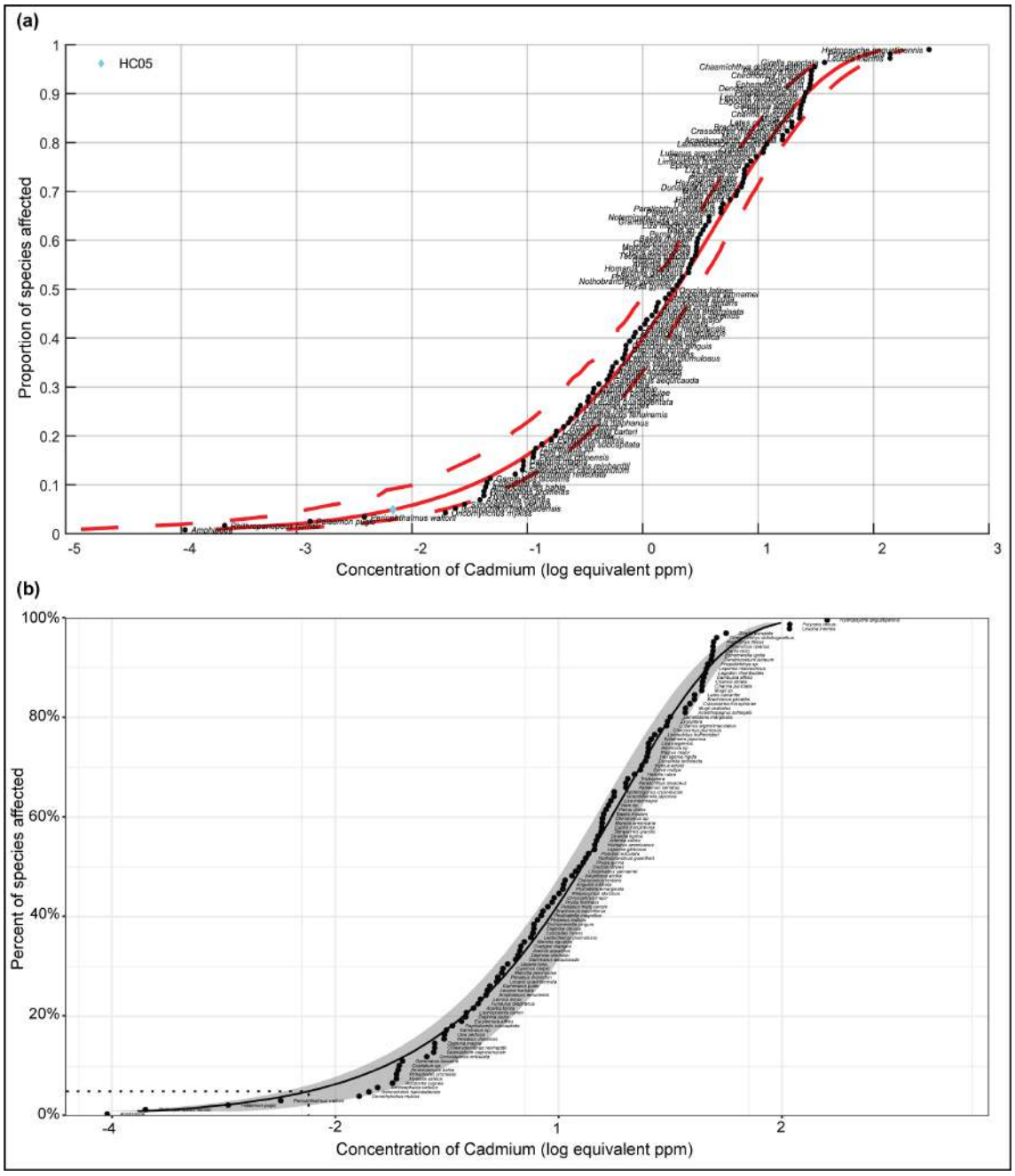


Figure 4.22: The plots of SSD for cadmium (CAS:7440-43-9) based on acute toxicity data and computed using the best-fit Weibull model. (a) As determined by USEPA SSD Toolbox where the HC05 value is denoted by cyan colored diamond. (b) As determined by ssdtools where the HC05 value is denoted by a dotted line.

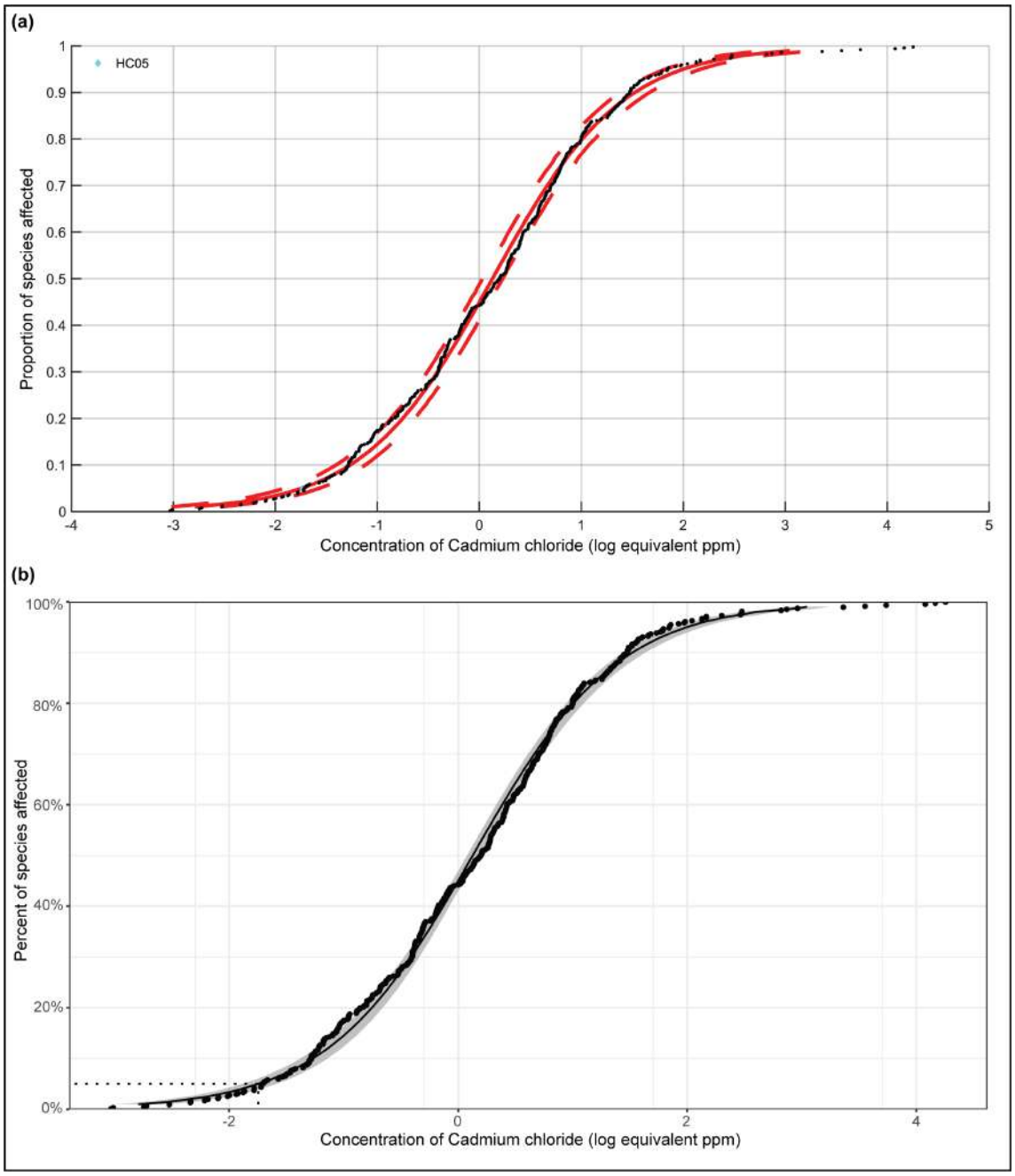


Figure 4.23: The plots of SSD for cadmium chloride (CAS:10108-64-2) based on acute toxicity data and computed using the best-fit Log-Logistic model. (a) As determined by USEPA SSD Toolbox where the HC05 value is denoted by cyan colored diamond. (b) As determined by ssdtools where the HC05 value is denoted by a dotted line.

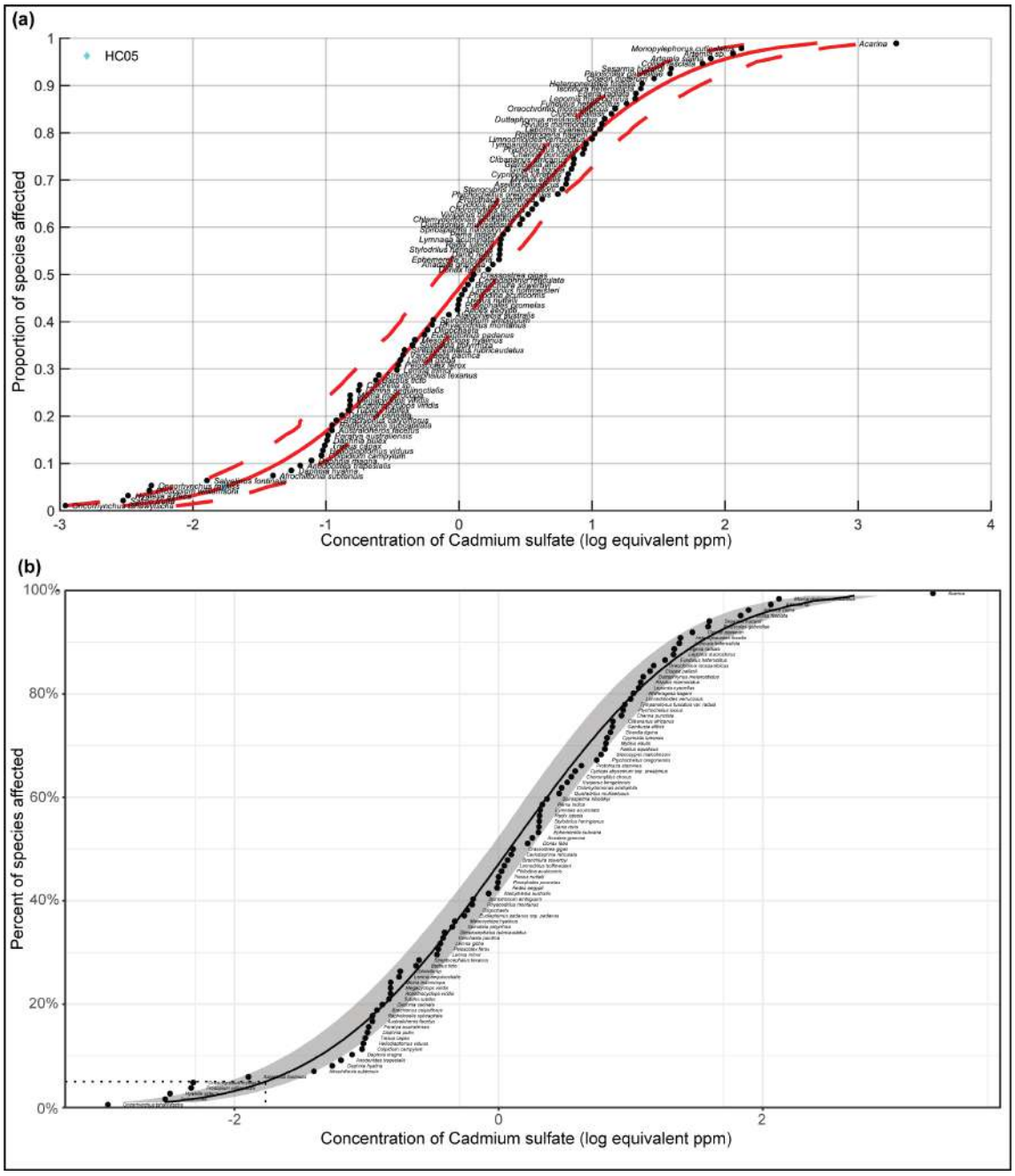


Figure 4.24: The plots of SSD for cadmium sulfate (CAS:10124-36-4) based on acute toxicity data and computed using the best-fit Log-Normal model. (a) As determined by USEPA SSD Toolbox where the HC05 value is denoted by cyan colored diamond. (b) As determined by ssdtools where the HC05 value is denoted by a dotted line.

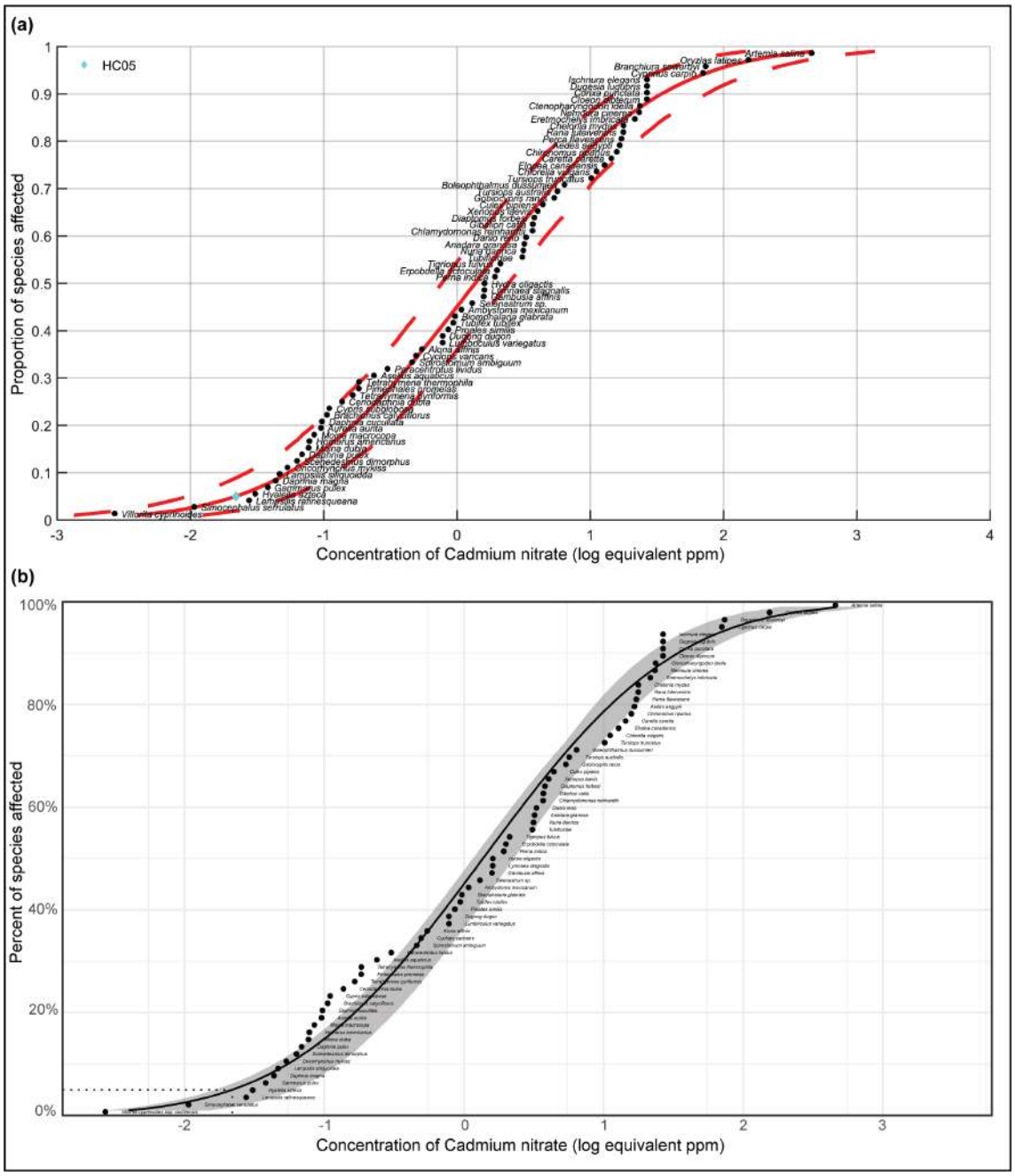


Figure 4.25: The plots of SSD for cadmium nitrate (CAS:10325-94-7) based on acute toxicity data and computed using the best-fit Weibull model. (a) As determined by USEPA SSD Toolbox where the HC05 value is denoted by cyan colored diamond. (b) As determined by ssdtools where the HC05 value is denoted by a dotted line.

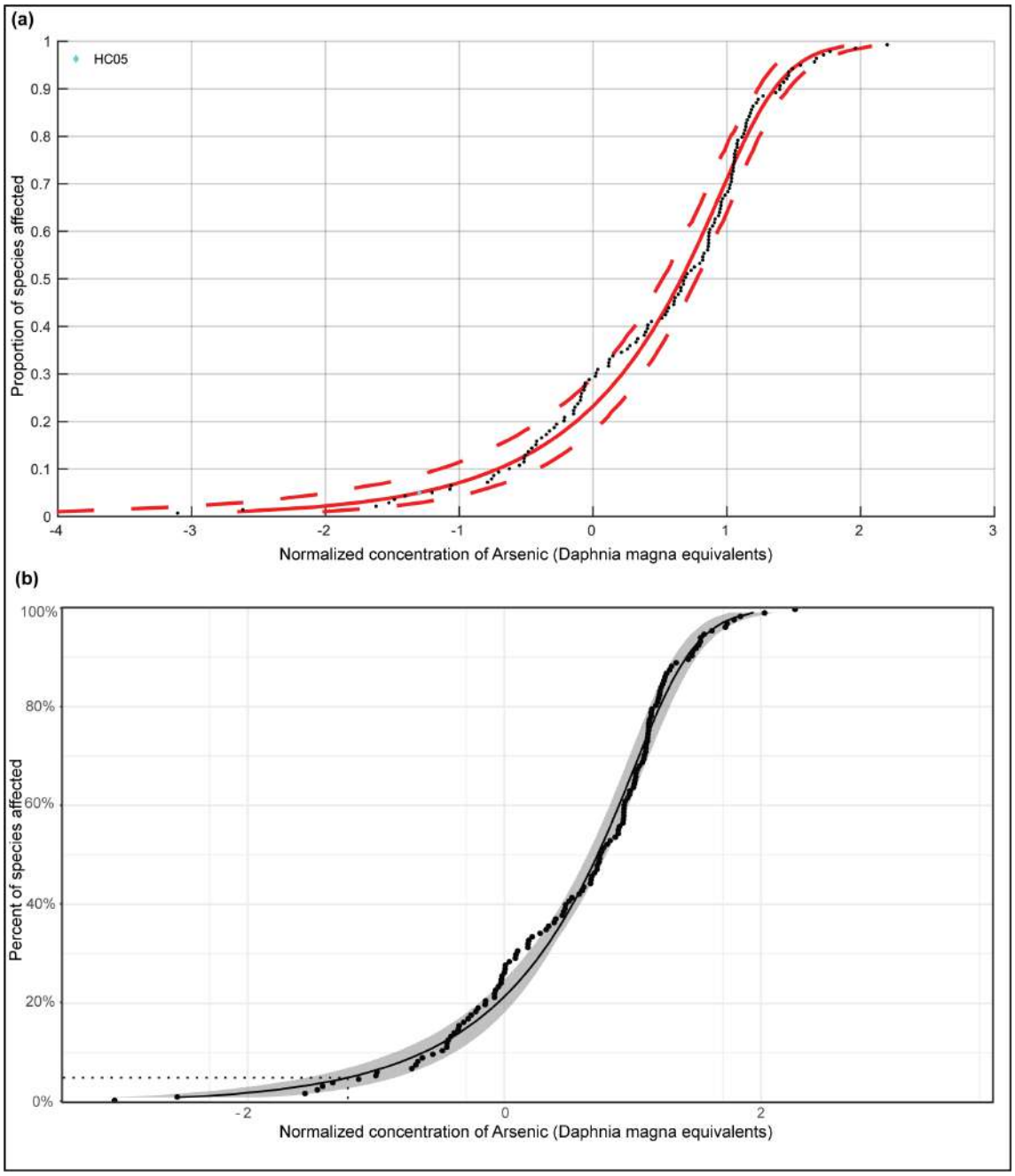


Figure 4.26: The plots of toxicity-normalized SSD (SSDn) for inorganic arsenic based on the acute toxicity data and computed using the best-fit Burr Type III model. (a) As determined by USEPA SSD Toolbox where the HC05 value is denoted by cyan colored diamond. (b) As determined by ssdtools where the HC05 value is denoted by a dotted line.

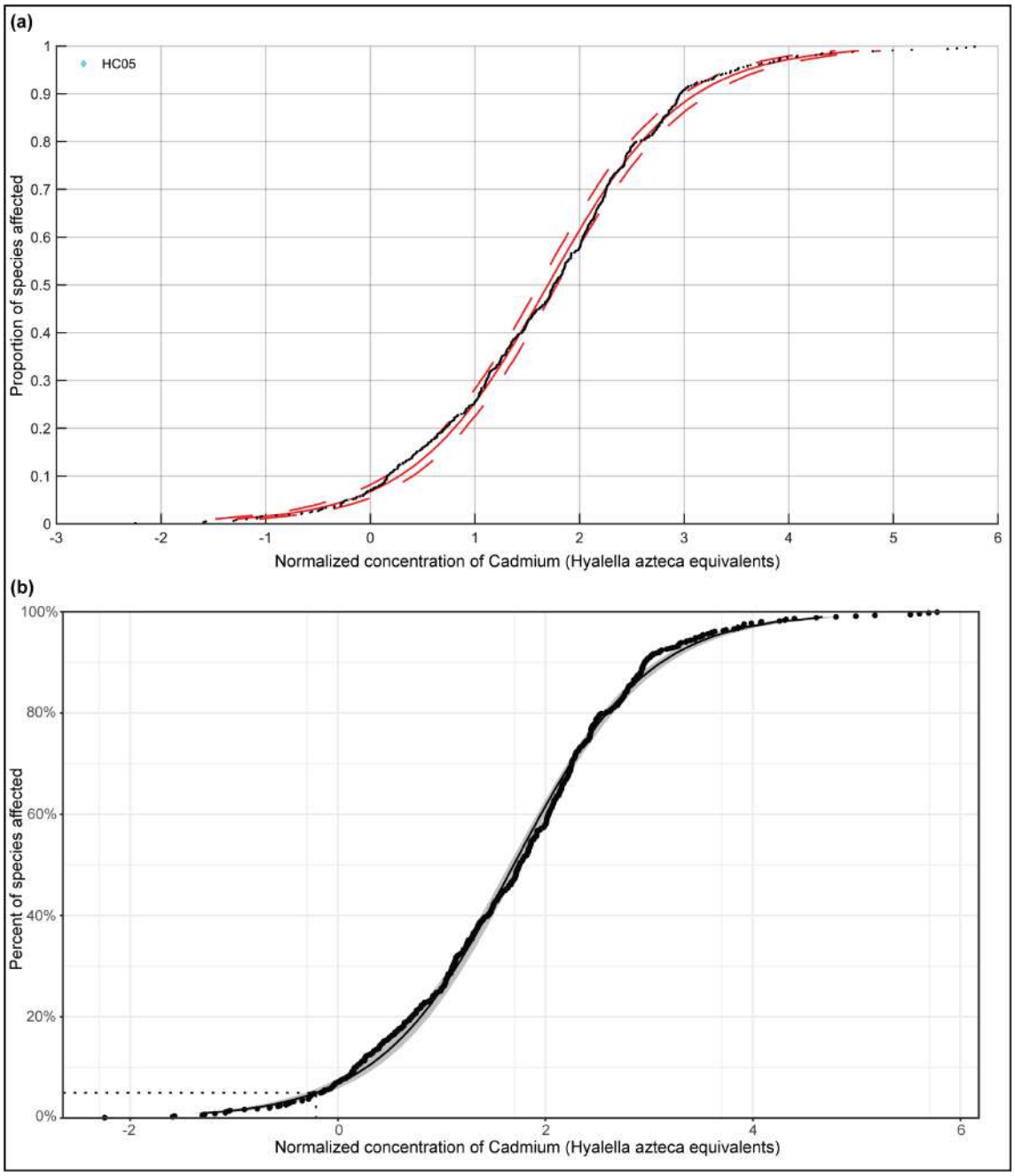


Figure 4.27: The plots of toxicity-normalized SSD (SSDn) for inorganic cadmium based on the acute toxicity data and computed using the best-fit Log-Logistic model. (a) As determined by USEPA SSD Toolbox where the HC05 value is denoted by cyan colored diamond. (b) As determined by ssdtools where the HC05 value is denoted by a dotted line.

Furthermore, SSDs can help identify species that are particularly sensitive to specific chemicals [287]. We observed that species in the ECOTOX species groups 'Crustaceans' and 'Algae' were commonly sensitive to both inorganic arsenic and cadmium compounds. Based on the annotated stressor-species network (Supplementary Table S4.11), we observed that *Scenedesmus acutus* var. acutus was sensitive to more than one inorganic arsenic compound, while *Eisenia fetida*, *Oncorhynchus tshawytscha*, and *Salmo trutta* were sensitive to more than one inorganic cadmium compound (Supplementary Table S4.11). Moreover, we observed that *Oncorhynchus kisutch*, which is sensitive to cadmium chloride (CAS:10108-64-2), and *Periophthalmus waltoni*, which is sensitive to cadmium (CAS:7440-43-9), were also documented to bioaccumulate these chemicals (Supplementary Tables S4.11-S4.12).

Additionally, we derived the HC05 values using the SSDn approach for each individual chemical (Figures 4.26-4.27; Supplementary Table S4.24). We observed that the SSDn method provided consistent estimates of HC05 values for all these chemicals (Supplementary Table S4.24), highlighting its utility as a reliable tool for assessing the acute toxicity of inorganic arsenic and cadmium.

SSDs based on chronic toxicity endpoints

Similarly, we constructed SSDs by utilizing the chronic toxicity endpoints for aquatic species from the ECOTOX database for three inorganic arsenic compounds, namely arsenic oxide (CAS:1327-53-3), disodium arsenate (CAS:7778-43-0), and sodium arsenite (CAS:7784-46-5) and four inorganic cadmium compounds, namely cadmium (CAS:7440-43-9), cadmium chloride (CAS:10108-64-2), cadmium sulfate (CAS:10124-36-4), and cadmium nitrate (CAS:10325-94-7). We observed that the results were consistent across both the tools (Supplementary Table S4.25). Figures 4.28-4.34 present the SSD plots for each chemical generated using the corresponding best-fit models from both tools. Thereafter, we derived the individual HC05 values, and observed that cadmium compounds were more toxic than the arsenic compounds (Supplementary Table S4.25). Moreover,

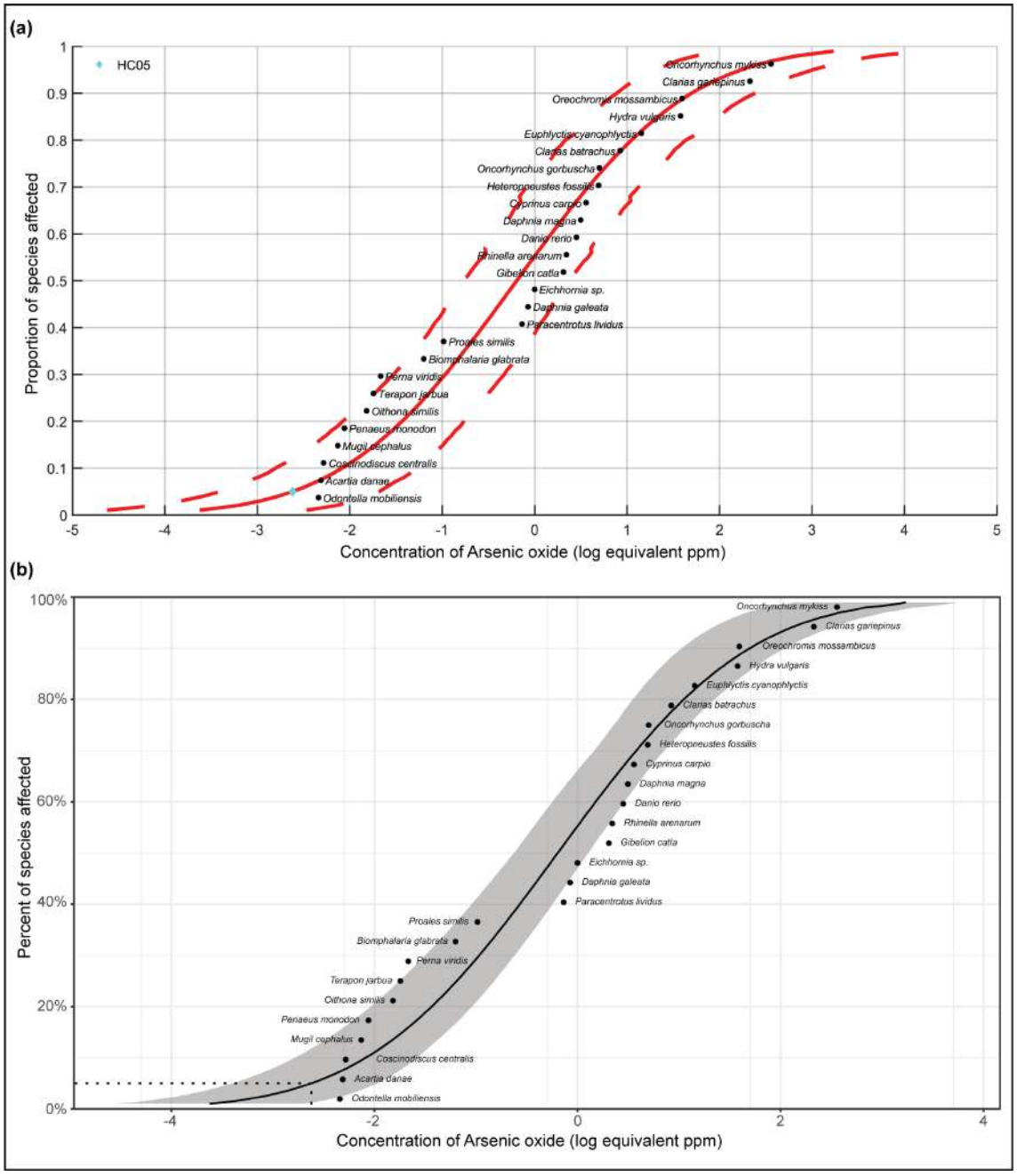


Figure 4.28: The plots of SSD for arsenic oxide (CAS:1327-53-3) based on chronic toxicity data and computed using the best-fit Log-Normal model. (a) As determined by USEPA SSD Toolbox where the HC05 value is denoted by cyan colored diamond. (b) As determined by ssdtools where the HC05 value is denoted by a dotted line.

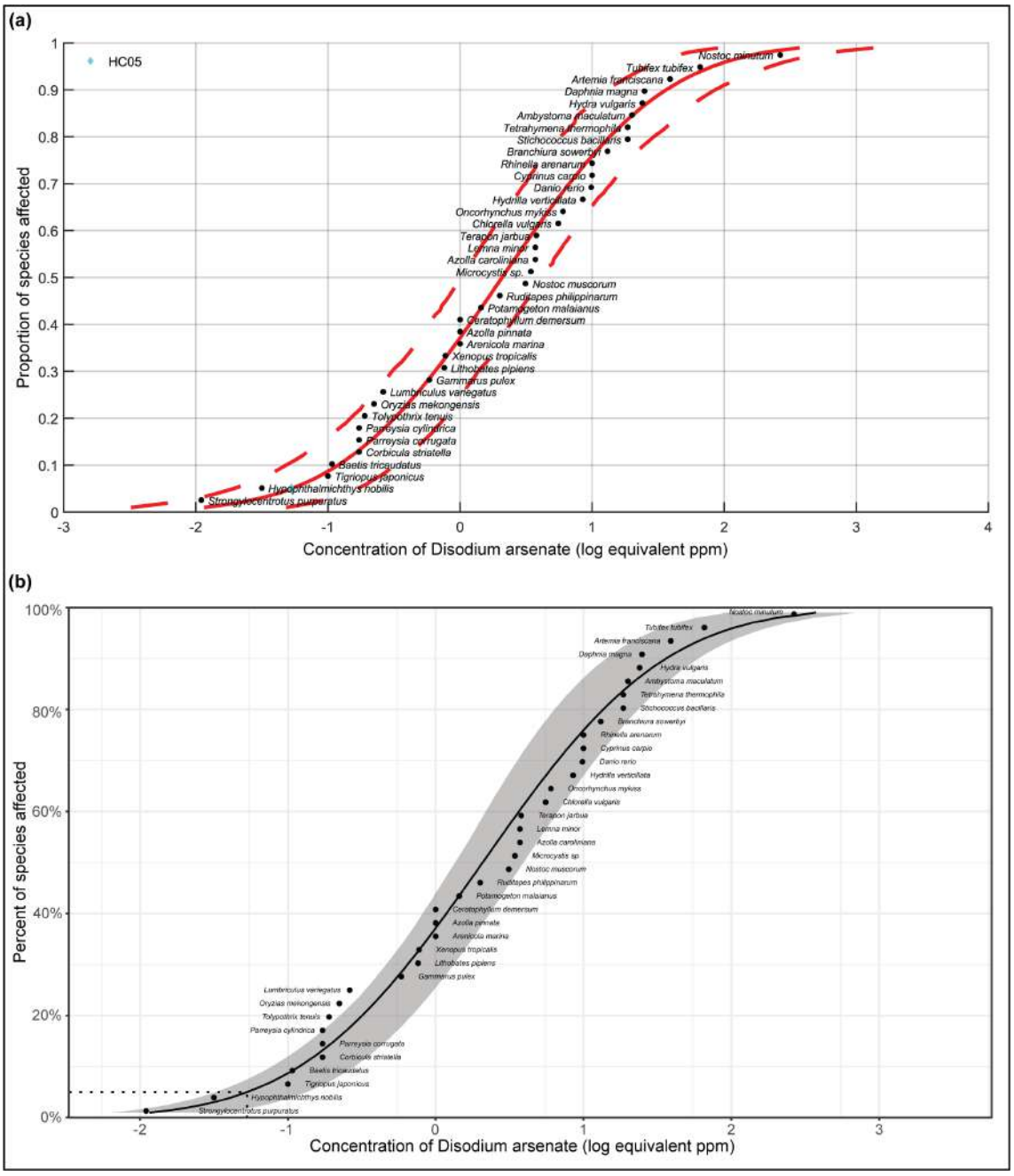


Figure 4.29: The plots of SSD for disodium arsenate (CAS:7778-43-0) based on chronic toxicity data and computed using the best-fit Log-Normal model. (a) As determined by USEPA SSD Toolbox where the HC05 value is denoted by cyan colored diamond. (b) As determined by ssdtools where the HC05 value is denoted by a dotted line.

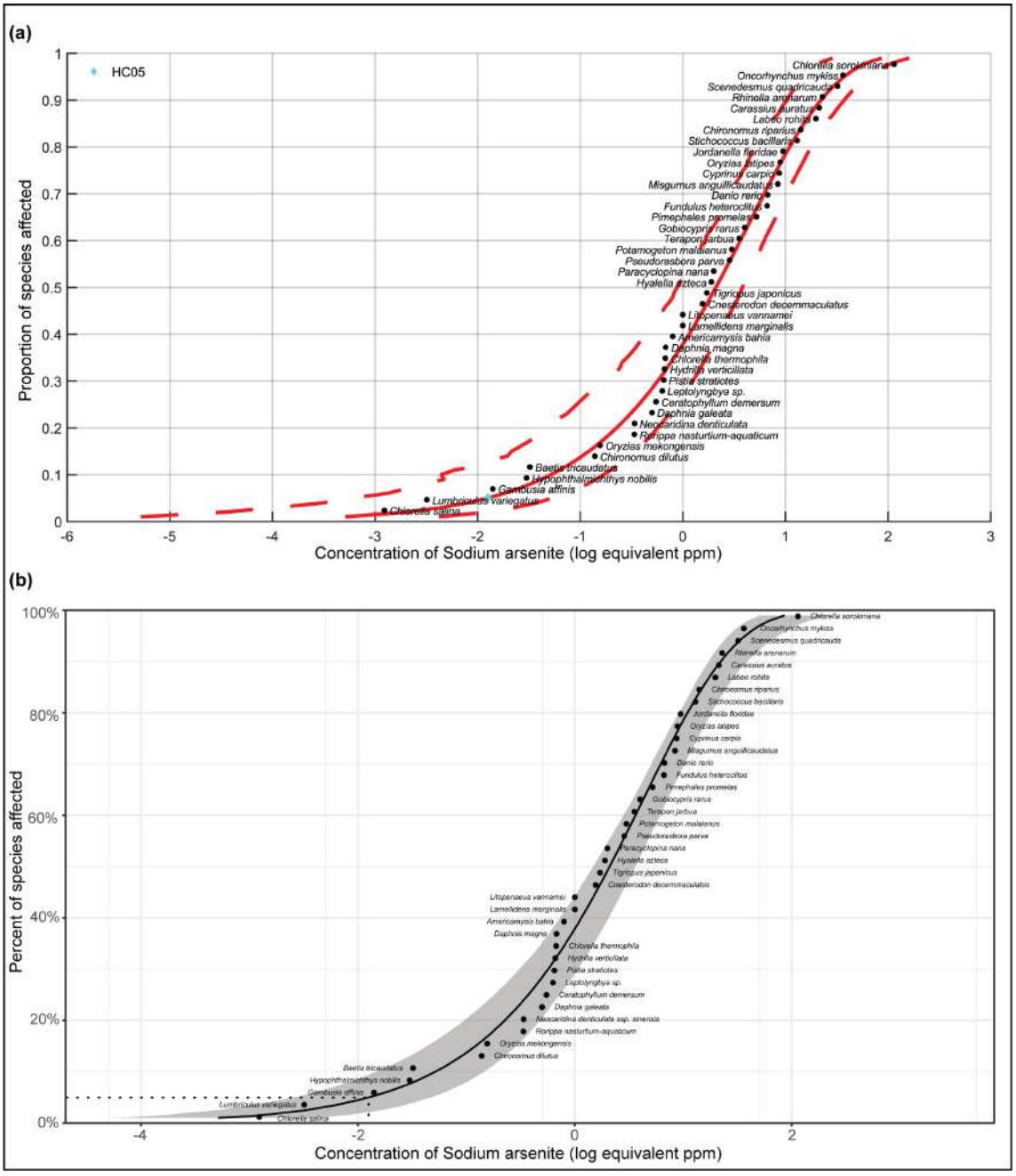


Figure 4.30: The plots of SSD for sodium arsenite (CAS:7784-46-5) based on chronic toxicity data and computed using the best-fit Burr Type III model. (a) As determined by USEPA SSD Toolbox where the HC05 value is denoted by cyan colored diamond. (b) As determined by ssdtools where the HC05 value is denoted by a dotted line.

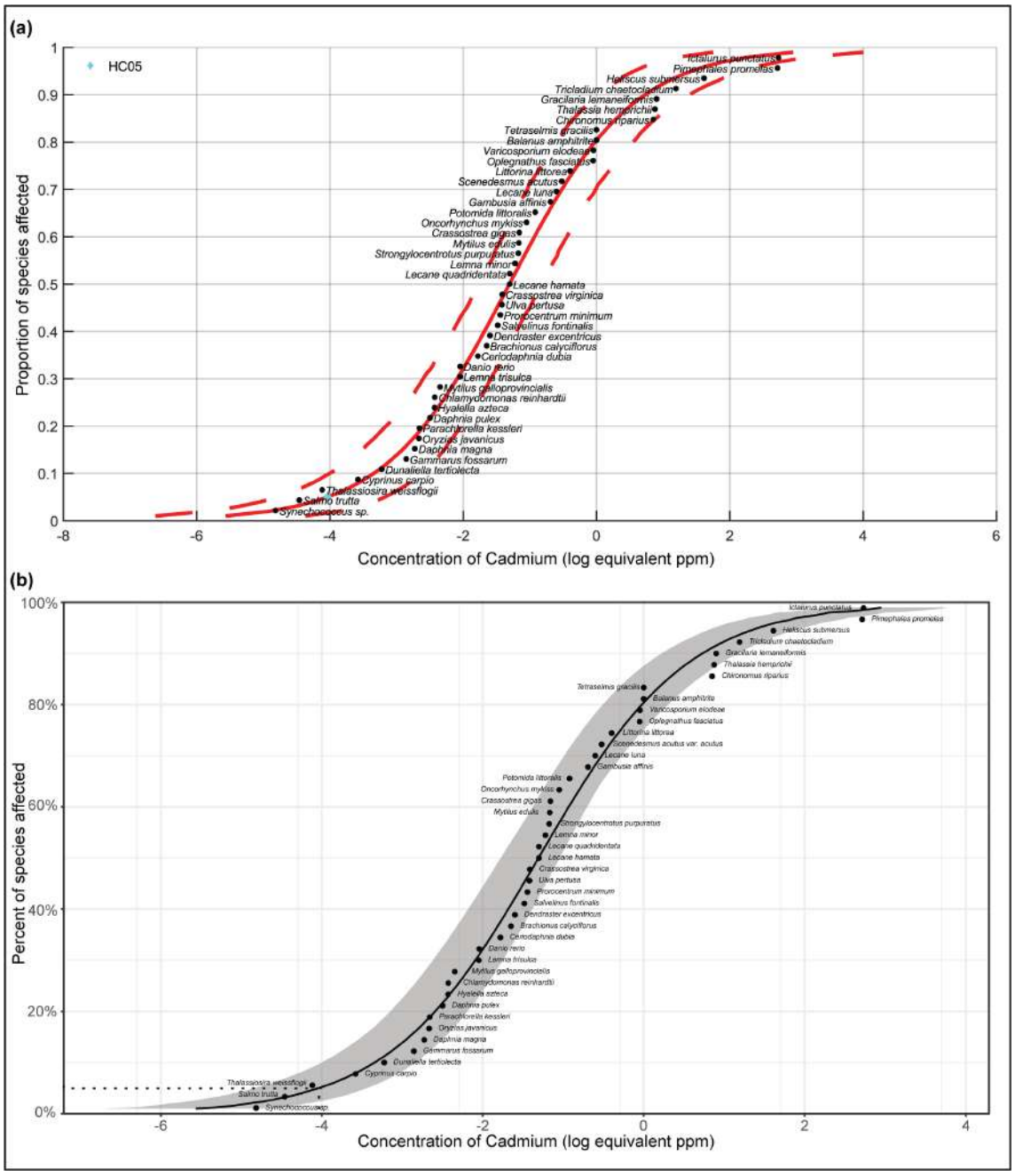


Figure 4.31: The plots of SSD for cadmium (CAS:7440-43-9) based on chronic toxicity data and computed using the best-fit Log-Logistic model. (a) As determined by USEPA SSD Toolbox where the HC05 value is denoted by cyan colored diamond. (b) As determined by ssdtools where the HC05 value is denoted by a dotted line.

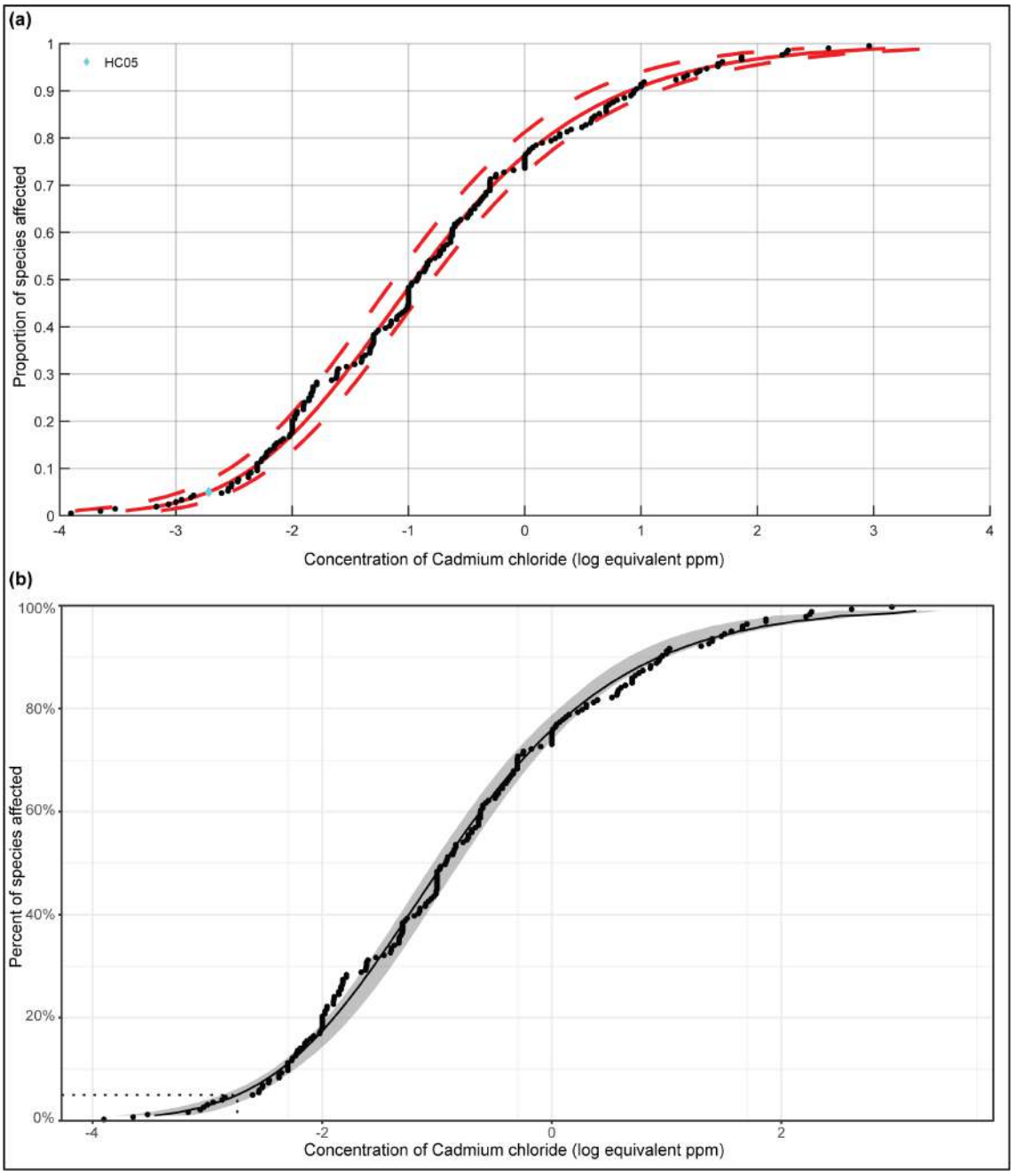


Figure 4.32: The plots of SSD for cadmium chloride (CAS:10108-64-2) based on chronic toxicity data and computed using the best-fit Burr Type III model. (a) As determined by USEPA SSD Toolbox where the HC05 value is denoted by cyan colored diamond. (b) As determined by ssdtools where the HC05 value is denoted by a dotted line.

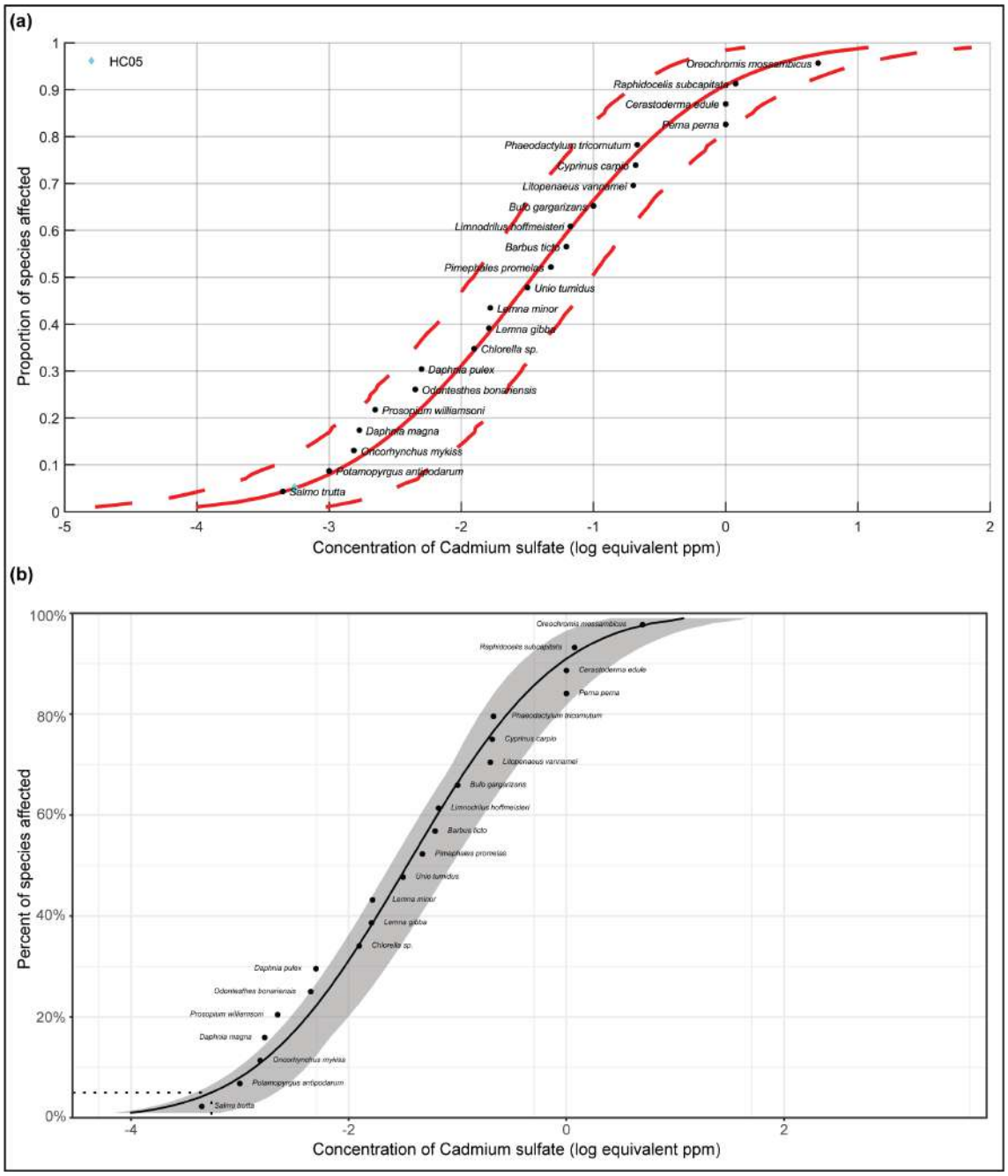


Figure 4.33: The plots of SSD for cadmium sulfate (CAS:10124-36-4) based on chronic toxicity data and computed using the best-fit Log-Normal model. (a) As determined by USEPA SSD Toolbox where the HC05 value is denoted by cyan colored diamond. (b) As determined by ssdtools where the HC05 value is denoted by a dotted line.

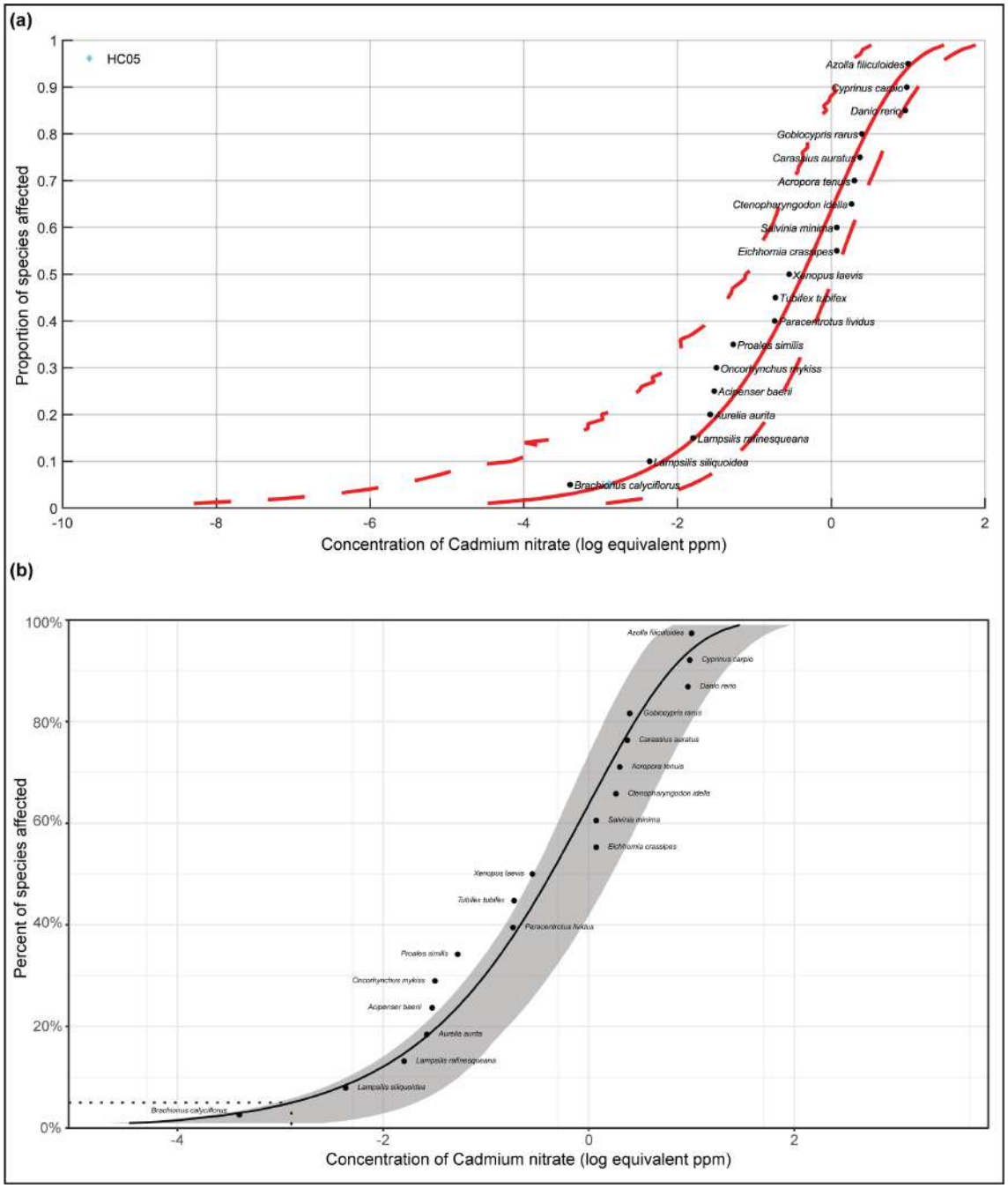


Figure 4.34: The plots of SSD for cadmium nitrate (CAS:10325-94-7) based on chronic toxicity data and computed using the best-fit Weibull model. (a) As determined by USEPA SSD Toolbox where the HC05 value is denoted by cyan colored diamond. (b) As determined by ssdtools where the HC05 value is denoted by a dotted line.

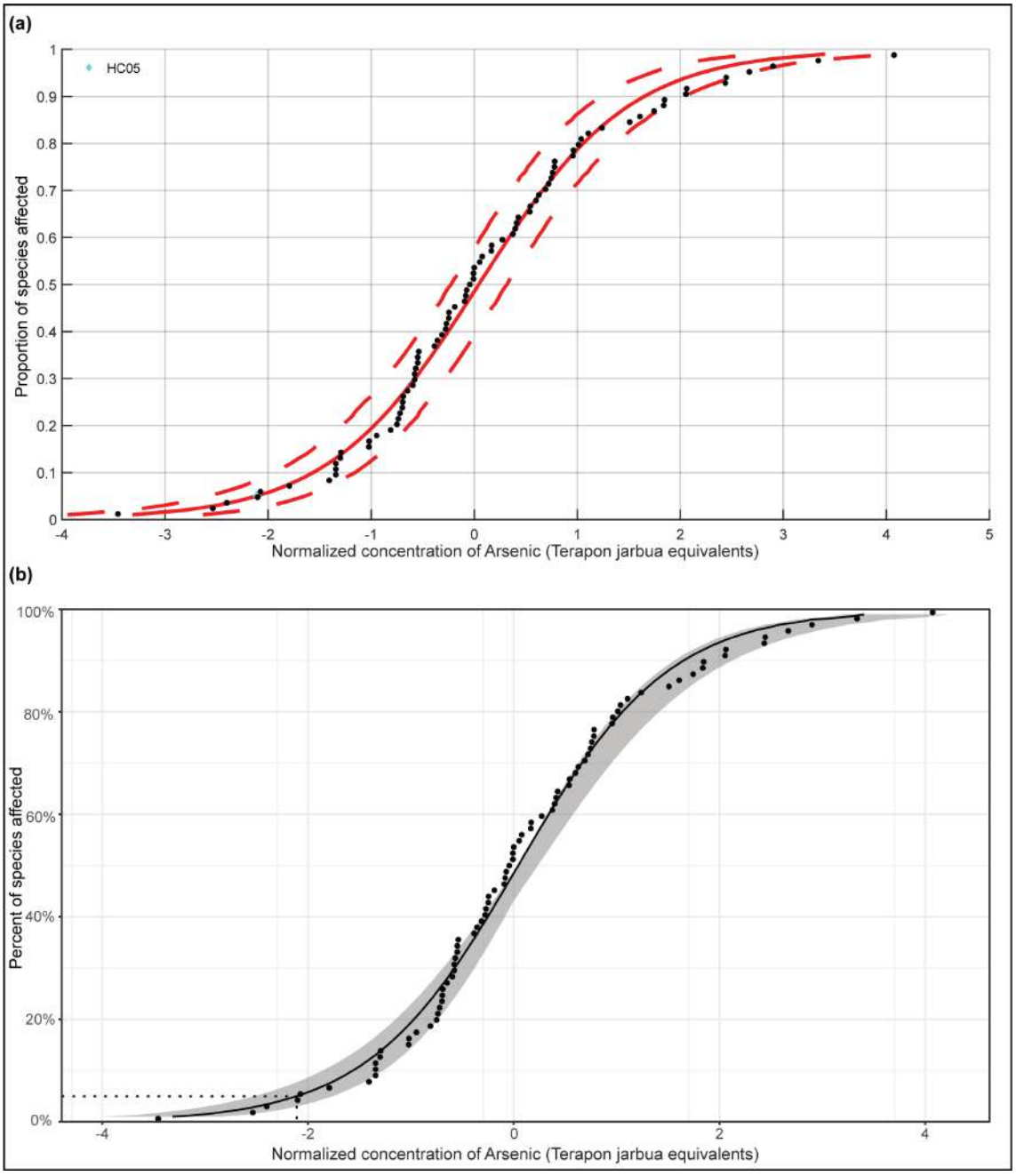


Figure 4.35: The plots of toxicity-normalized SSD (SSDn) for inorganic arsenic based on the chronic toxicity data and computed using the best-fit Log-Logistic model. (a) As determined by USEPA SSD Toolbox where the HC05 value is denoted by cyan colored diamond. (b) As determined by ssdtools where the HC05 value is denoted by a dotted line.

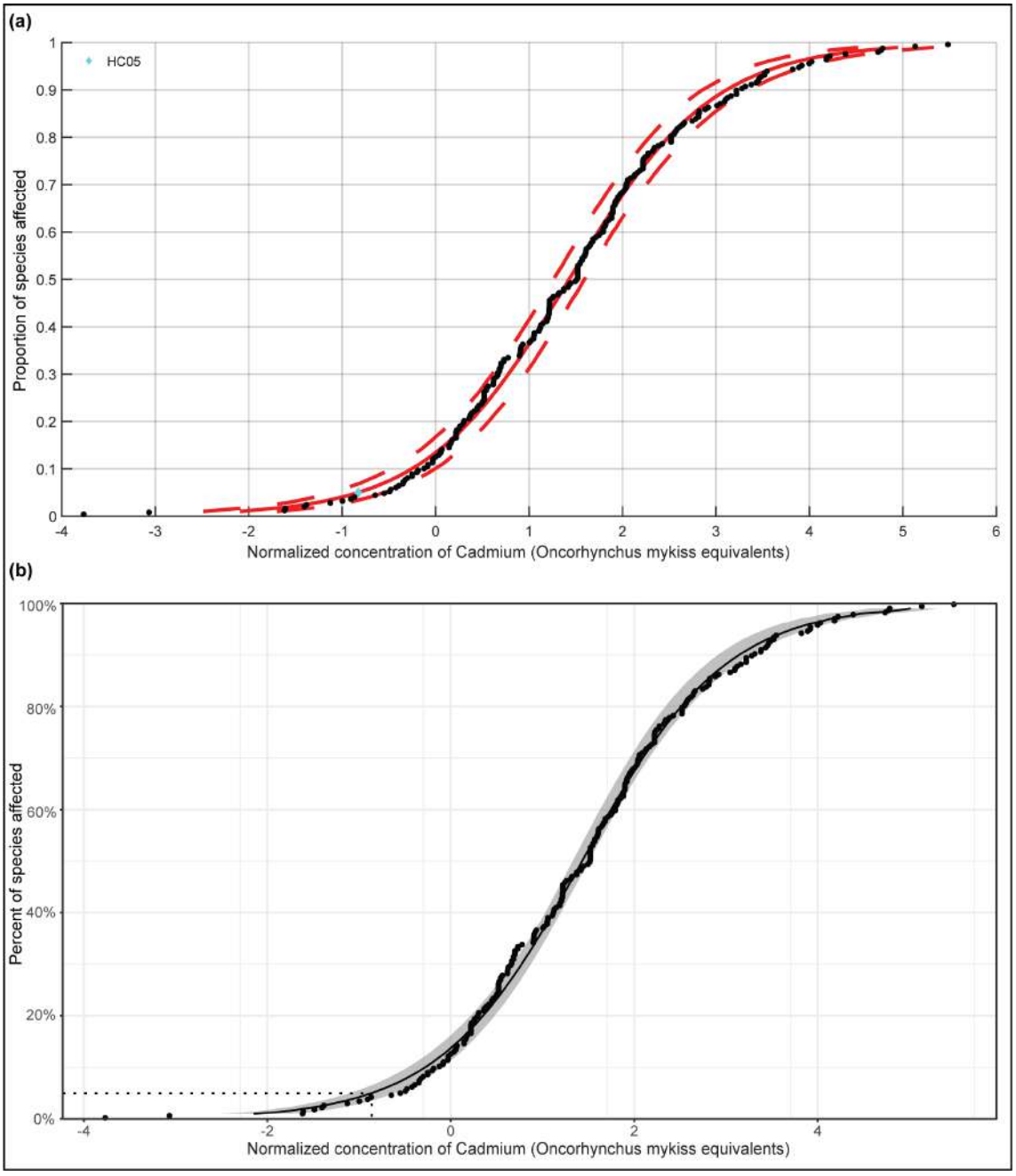


Figure 4.36: The plots of toxicity-normalized SSD (SSDn) for inorganic cadmium based on the chronic toxicity data and computed using the best-fit Log-Logistic model. (a) As determined by USEPA SSD Toolbox where the HC05 value is denoted by cyan colored diamond. (b) As determined by ssdtools where the HC05 value is denoted by a dotted line.

we observed that arsenites, namely arsenic oxide (CAS:1327-53-3) and sodium arsenite (CAS:7784-46-5), were more toxic than disodium arsenate (CAS:7778-43-0) (Supplementary Table S4.25).

Furthermore, we observed that species in the ECOTOX species groups 'Crustaceans' and 'Fish' were commonly sensitive to both inorganic arsenic and cadmium compounds (Supplementary Table S4.11). Based on the annotated stressor-species network (Supplementary Table S4.11), we observed that *Salmo trutta* was sensitive to more than one inorganic cadmium compound, while there were no species that were sensitive to more than one inorganic arsenic compound. Moreover, we observed that *Hyalella azteca*, which is sensitive to cadmium chloride (CAS:10108-64-2), was also documented to bioaccumulate this chemical (Supplementary Tables S4.11-S4.12). Additionally, we derived the HC05 values using the SSDn approach (Figures 4.35-4.36; Supplementary Table S4.26), and observed that this method provided consistent estimates of HC05, other than that for cadmium.

In conclusion, we utilized both the acute and chronic toxicity endpoints for aquatic organisms from the ECOTOX knowledgebase to construct SSDs and derive HC05 values for the inorganic arsenic and cadmium compounds. Additionally, we employed the SSDn approach and found that it provided consistent HC05 estimates for these chemicals.

4.2.8 RQ-based analysis revealed ecotoxicity of arsenic and cadmium in Indian rivers

Water quality guidelines are legislative instruments that define safe and desirable water conditions by setting regulatory concentration limits as jurisdiction-specific risk management tools [292–294]. Globally, these guidelines are applied either to protect human health through drinking water standards or to safeguard aquatic ecosystems through environmental water quality criteria (WQC) [294]. Here, we compiled available drinking water standards and WQC for arsenic and cadmium from multiple regulatory bodies, including those in

India, the United States, the EU, South Korea, China, Australia, and New Zealand (Supplementary Tables S4.27). We observed that India has drinking water quality guidelines of 10 μg/L for arsenic and 3 μg/L for cadmium, as outlined by Bureau of Indian Standards (BIS), and these values are comparable with the WHO standards (Supplementary Table S4.27). However, we noted that these standards apply only to human consumption, and India currently lacks environmental WQC, suggesting the need to develop such standards for aquatic environments.

The PNEC values are often used to derive water quality standards as well as to assess the risk posed by a particular chemical [295]. We observed that the SSDn approach provided normalized unitless toxicity values for arsenic and cadmium that cannot be directly used to derive PNEC values. Therefore, we computed the geometric mean of PNEC values from corresponding arsenic and cadmium compound-relevant SSDs, and considered them as representative PNEC values for total arsenic and total cadmium, respectively. Further, we compared these PNEC values with existing environmental WQCs across the globe (Supplementary Table S4.27), and found them to be comparable with, or in some cases, more stringent than existing WQC as implemented by different regulatory bodies (Supplementary Table S4.27). Thus, we used the calculated PNEC values to derive RQs for arsenic and cadmium, and subsequently utilized them to understand their ecotoxicological effects in Indian river systems.

The Central Water Commission report provided the total arsenic and total cadmium concentrations observed across different rivers of India [232]. Of the 409 sampling stations spanning 24 states and 238 rivers that had data on arsenic concentrations in water, we observed that 225 sampling stations spanning 22 states and 148 rivers had RQ > 1 (Supplementary Table S4.28). Similarly, of the 424 sampling stations spanning 24 states and 247 rivers that had data on cadmium concentrations in water, we observed that 256 sampling stations spanning 24 states and 161 rivers had RQ > 1 (Supplementary Table S4.28). An RQ > 1 potentially indicates ecological risk due to arsenic and cadmium contamination in these rivers [70].

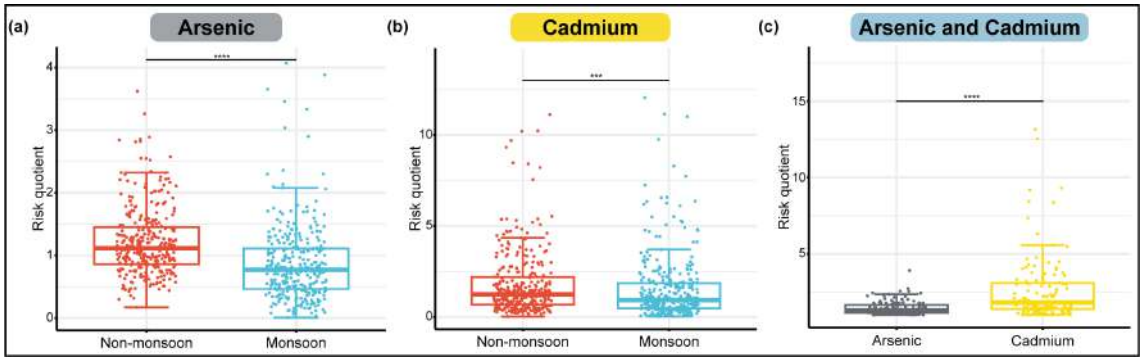


Figure 4.37: Comparative analysis of RQs of arsenic and cadmium computed in this chapter. (a) Comparison of non-monsoon and monsoon RQ data for arsenic (significantly different at $p < 0.0001$). (b) Comparison of non-monsoon and monsoon RQ data for cadmium (significantly different at $p < 0.001$). This box-plot shows data within the 99^{th} quantile for better visualization. (c) Comparison of overall RQ data for arsenic and cadmium (significantly different at $p < 0.0001$). This box-plot shows data within the 99^{th} quantile for better visualization.

S. No.	Test variables		Paired sample size	Shapiro-Wilks test *p*-value	Wilcoxon paired signed-rank Test		Effect size	
	Variable 1	Variable 2			Statistic	*p*-value	Wilcoxon effect size	Magnitude
1	Arsenic monsoon RQ	Arsenic non-monsoon RQ	318	1.64E-07	11274	7.46E-18	0.478	Moderate
2	Cadmium monsoon RQ	Cadmium non-monsoon RQ	326	9.14E-32	20675.5	2.26E-04	0.194	Low
3	Arsenic RQ	Cadmium RQ	135	1.01E-21	1453	1.01E-21	0.5929	Large

Table 4.2: Comparison of RQ data for arsenic and cadmium in the Indian rivers.

Next, we compared the seasonal RQ data (monsoon and non-monsoon) for both arsenic and cadmium using Wilcoxon paired signed-rank test as Shapiro-Wilks test suggested non-normality of the distributions of the corresponding differences (Table 4.2). For both the chemicals, we found that monsoon RQ was significantly lower than non-monsoon RQ (Table 2; Figures 4.37a-4.37b). This can be attributed to the dilution effect observed during the monsoon season, where increased river flows reduce metal concentrations [296].

Finally, we compared arsenic RQs with cadmium RQs after checking the non-normality of the distributions of the differences (Table 4.2). We used the Wilcoxon paired signed-rank

test and observed that arsenic RQs were significantly lower than cadmium RQs (Table 4.2; Figure 4.37c). This result suggests that the ecological risk posed by cadmium to Indian rivers is potentially higher than that of arsenic.

4.3 Discussion

In this chapter, network toxicology-based frameworks provided crucial insights into the exposure pathways and complex biological responses associated with inorganic arsenic and cadmium, enhancing our understanding of their systemic impacts across species and ecosystems in India. We constructed AEPs for the first time for both inorganic arsenic and cadmium in the Indian context, revealing novel source-to-target site pathways. Notably, compared with previous work on AEP construction for cadmium and nephrotoxicity in China [297], we expanded coverage to all possible biological and environmental states in this chapter.

Further, we expanded the associations between these chemicals and ecotoxicologically-relevant AOPs by leveraging the ECOTOX knowledgebase. Notably, compared to previous work on the arsenic-associated AOP network [174], we extended the associations to include all human- and ecotoxicologically-relevant AOPs associated with arsenic-induced toxicity. Moreover, compared with our earlier work on AOPs associated with cadmium-induced toxicity [55], we retained previously identified AOPs and expanded coverage to include additional ecotoxicologically-relevant AOPs.

Moreover, our AEP-AOP based case studies highlighted critical knowledge gaps regarding human exposures and tissue bioaccumulation. We identified particularly vulnerable species through the constructed stressor-species networks, underscoring the ecological relevance of this approach. We further substantiated the associated ecological risks through the RQ analysis conducted for Indian river systems, which showed that many regions exhibited elevated risks. Overall, we leveraged diverse network toxicological approaches to comprehensively assess the toxicological impacts of inorganic arsenic and

cadmium exposure from an India-specific One Health perspective. We made the networks and related data generated in this chapter freely accessible for further research at https://cb.imsc.res.in/heavymetaltox/.

However, our approach is constrained by the limitations of the various underlying data. The limited information on environmental factors like pH and temperature limits the scope of these findings in broader natural settings. The constructed AEP networks were further constrained by the geographical diversity of sampling sites across India and the use of diverse measurement techniques, making it difficult to harmonize the chemical concentrations. Moreover, the limited India-specific toxicity data in ECOTOX restricts the representation of ecological groups covered in this chapter. For SSDs, the lack of geographic specificity limits the applicability of inferences to specific locations. Additionally, the higher fold change observed when comparing HC05 values derived from SSDn and single-compound SSDs for cadmium may indicate that, despite selecting the most sensitive species available, the species may not be sensitive enough to assess this chemical. Furthermore, the approaches used in this chapter are limited in their ability to assess the bioaccumulation of arsenic and cadmium compounds across trophic levels. Finally, the data used for the RQ analysis may not fully represent current contaminant levels in these rivers, as the sampling was conducted before 2019.

Nonetheless, the network-based approaches employed in this chapter have enhanced our understanding of the toxicities induced by inorganic arsenic and cadmium. The inclusion of ecotoxicologically-relevant endpoints from the ECOTOX database, alongside mammalian toxicity endpoints, has broadened the scope and applicability of associated AOPs in advancing our understanding of arsenic- and cadmium-induced toxicities. A novel computational approach was developed to infer potential taxonomic applicability for AOPs lacking such information, thereby expanding their scope and relevance. The AEP-AOP constructs provided a mechanistic perspective by highlighting source-to-outcome pathways associated with inorganic arsenic and cadmium in India. Additionally, the SSD and SSDn approaches facilitated comparisons among these compounds, offering a framework for

prioritizing them in further toxicity assessment studies. Importantly, the SSDn approach enabled pooling of arsenic and cadmium-relevant toxicity information, and a comparison with the results from compound-specific SSDs showed good concordance, thereby supporting our approach. Moreover, by integrating SSD data with the stressor-species networks, species groups particularly sensitive to exposure from these compounds were identified, thereby strengthening the utility of stressor-species networks in ecological risk assessment. Additionally, the taxonomic applicability inference methodology can help identify ecotoxicologically-relevant AOPs that were previously unrecognized, thereby enabling the selection of relevant endpoints for which SSDs can be constructed. The calculated PNECs for total inorganic arsenic and cadmium were comparable to environmental WQCs implemented by regulatory bodies worldwide, thereby corroborating the utility of this data-centric approach. Thus, the RQ analysis based on these computed PNEC values can potentially aid in formulating better water quality criteria in India.

In sum, this chapter provides a comprehensive understanding of the exposure pathways, toxicity mechanisms, and impacts of inorganic arsenic and cadmium compounds on human and ecosystem health, thereby supporting a One Health-focused approach towards enhanced regulatory and mitigation strategies for these compounds in India.

Supplementary Information

Supplementary Tables S4.1-S4.28 associated with this chapter are available for download from the GitHub repository: https://github.com/asamallab/PhDThesis-Nikhil_C/blob/main/SI/ST_Chapter4.xlsx.

Chapter 5

Summary and future outlook

The balance of nature is not a status quo; it is fluid, ever shifting, in a constant state of adjustment. Man, too, is part of this balance.

- Rachel Carson

In this thesis, we investigated the links between exposome, particularly chemical exposome, and human health using diverse *in silico* methods. By employing network toxicology and computational approaches, we characterized chemical exposures, identified underlying toxicity mechanisms, and assessed ecological and human health risks associated with environmental chemical exposures. In the following section, we provide a summary of the research reported across different chapters of this thesis. Thereafter, we conclude with a short discussion of possible future directions based on the research reported in this thesis.

5.1 Summary

ViCEKb: Vitiligo-linked Chemical Exposome Knowledgebase

In Chapter 2 of this thesis, we focused on characterizing the chemical exposome of vitiligo (Figure 5.1). We initially systematically curated 113 chemical triggers of vitiligo, along

with their sources of exposure and associated evidence, from 224 published articles. We made these chemicals, along with their structures and toxicological metrics, accessible through a user-friendly online resource, ViCEKb, accessible at https://cb.imsc.res.in/vicekb/.

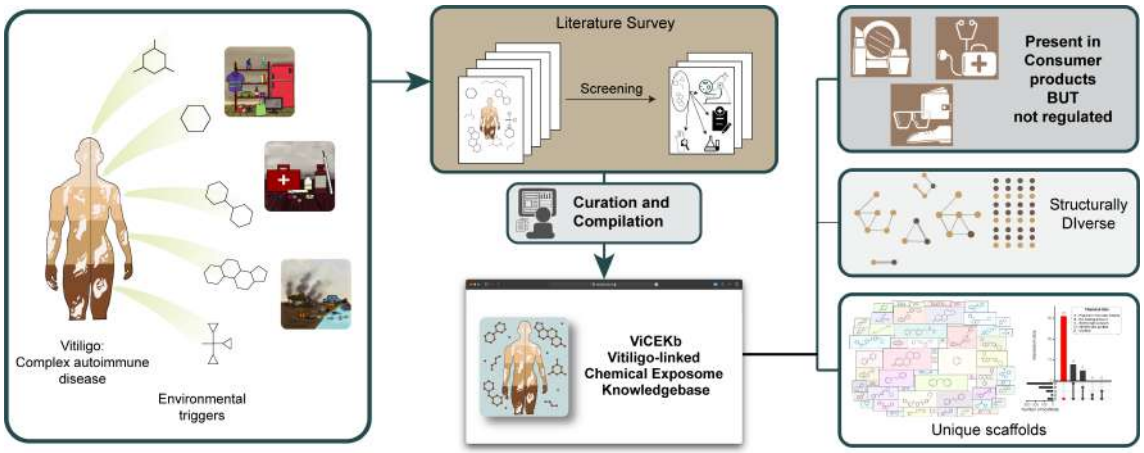


Figure 5.1: Summary of work done in characterizing the chemical exposome of vitiligo, as reported in Chapter 2.

We cross-referenced these chemicals with the United States Environmental Protection Agency's (USEPA's) Chemical and Products Database (CPDat) and identified that 35 of the 113 chemicals are present in daily-use consumer products, including household care and cosmetics. We further observed that the majority of ViCEKb chemicals had not been tested or flagged for skin-sensitizing effects under existing skin-specific regulations across jurisdictions, and most were absent from other exposome databases, underscoring the need for disease-specific chemical exposome characterization.

Through a cheminformatics-based chemical space analysis, we observed that only 28 of 113 chemicals shared greater than 50% structural similarity with another chemical, indicating high chemical diversity. We further identified that 94 of the 113 chemicals share 56 unique scaffolds, of which 31 scaffolds were not captured by any skin-specific regulations, revealing a clear regulatory gap. Finally, we obtained skin-specific transcriptomic signatures for 18 of the 113 chemicals, of which 10 chemicals showed significant enrichment of inflammatory pathways, demonstrating a shared biological mechanism despite

chemical diversity. Overall, we presented the first comprehensive effort in compilation and exploration of various chemical triggers of vitiligo. We believe such a resource will enable deciphering the complex etiology of vitiligo and aid in the characterization of human chemical exposome.

DEDuCT 3.0: An enhanced and expanded FAIR-compliant resource and toxicology knowledge graph for endocrine disrupting chemicals

In Chapter 3 of this thesis, we presented DEDuCTv3.0, an updated resource and the first large-scale toxicology knowledge graph for potential endocrine disrupting chemicals (EDCs) (Figure 5.2). We systematically screened more than 14000 newly published articles through a four-stage workflow and identified 479 new potential EDCs from 1051 papers, associated with 573 endocrine-mediated effects. We merged these with prior versions to obtain a comprehensive resource covering 1043 chemicals, 796 endocrine-mediated effects, and 3269 published articles.

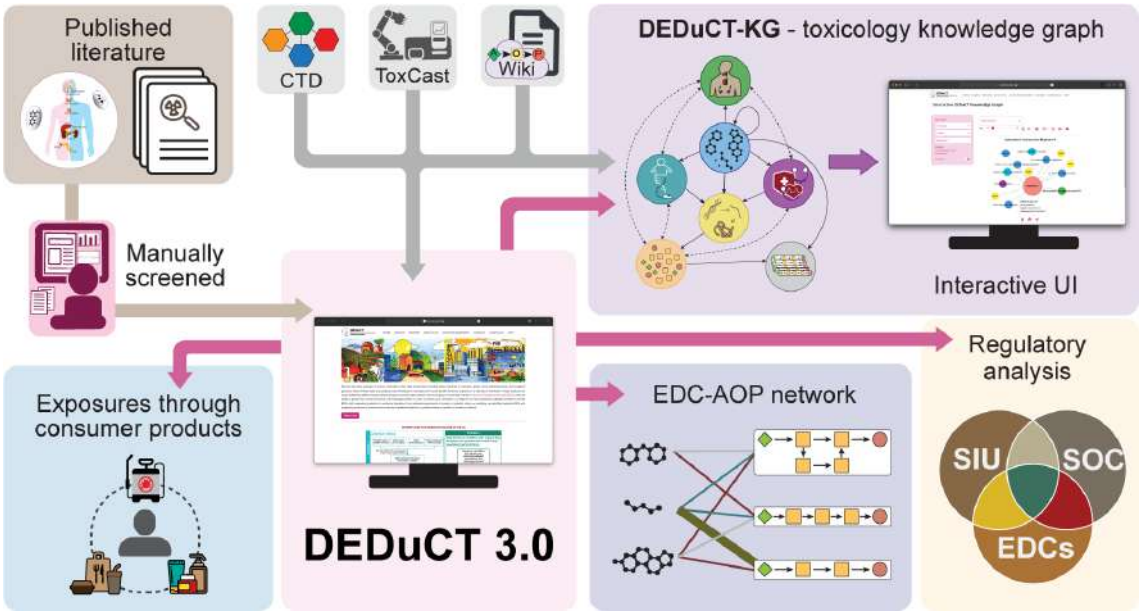


Figure 5.2: Summary of work done in updating and gaining mechanistic insights from DEDuCTv3.0, as reported in Chapter 3.

We curated human- and rodent-specific gene targets from ToxCast invitrodb v4.3 and Comparative Toxicogenomics Database (CTD), alongside phenotype and disease information. We further curated 385 high-confidence Adverse Outcome Pathways (AOPs)

from AOP-Wiki and systematically mapped EDCs to construct a stressor-AOP network linking 949 chemicals to 381 AOPs. We then organized all compiled data into a toxicology knowledge graph comprising 75106 nodes, categorized as chemicals, endocrine-mediated effects, genes, phenotypes, diseases, key events (KEs), and AOPs, and 1065949 edges across 31 edge types. We demonstrated the utility of this knowledge graph by identifying potential mechanisms of action underlying EDC-associated obesity and neurodegenerative diseases. Additionally, we assessed regulatory coverage using multiple chemical lists, and explored exposure sources using the USEPA's CPDat. We made all compiled data accessible through an updated, interactive, FAIR-compliant interface at the DEDuCTv3.0 webserver, accessible at https://cb.imsc.res.in/deduct/.

In sum, we presented a much-needed update to the resource and, most importantly, organized diverse toxicological data as a knowledge graph to enable future research. We provided an expanded list of 1043 potential EDCs in DEDuCTv3.0, which could assist academia, industry, and regulatory agencies in developing safer consumer products and better chemical regulation strategies for EDCs.

Network toxicology focused investigation on the impacts of inorganic arsenic and cadmium on human and ecosystem health in India

In Chapter 4 of this thesis, we applied diverse network toxicology frameworks to comprehensively assess the impacts of inorganic arsenic and cadmium from an India-specific One Health perspective (Figure 5.3). We identified 62 inorganic arsenic and 18 inorganic cadmium compounds from six exposome-relevant resources and extracted their corresponding toxicological endpoints. For the first time, we constructed Aggregate Exposure Pathway (AEP) networks using India-specific data across different environmental and biological states, and identified source-to-biological site of action pathways for these heavy metals.

We curated 342 complete and connected AOPs from AOP-Wiki and integrated them with extracted endpoints to construct stressor-AOP networks for both arsenic and cadmium, with varying levels of relevance and coverage scores. We also analyzed AOP networks and

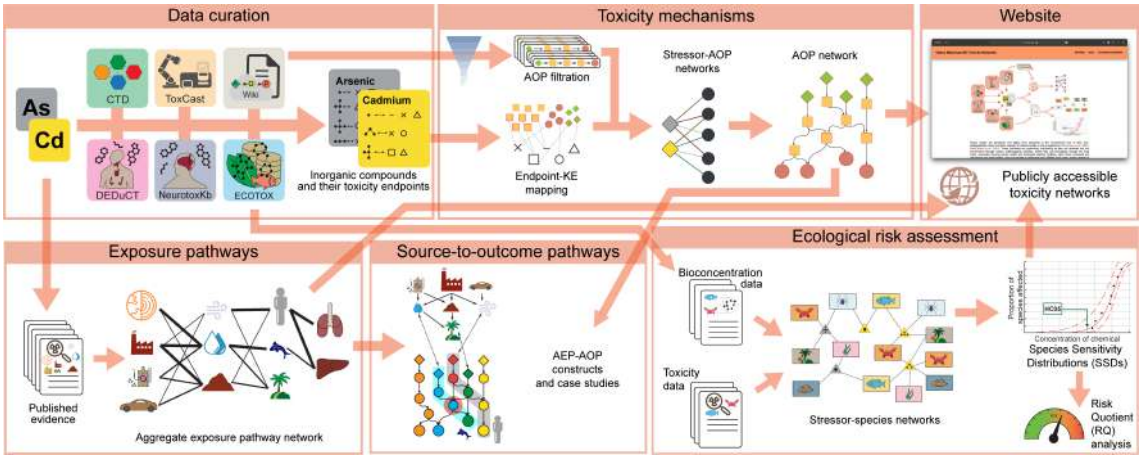


Figure 5.3: Summary of work done in network toxicology-based investigation of inorganic arsenic- and cadmium-induced adverse effects on human and ecosystem health in India, as reported in Chapter 4.

obtained novel taxonomic applicability information for the mapped AOPs based on the compiled evidence. Further, we merged the AEPs and AOPs to obtain complete source-to-outcome pathways for inorganic arsenic and cadmium, and thereafter, relied on published evidence to support the identified pathways.

We utilized acute and chronic toxicity data, along with bioconcentration factor (BCF) data from the ECOTOX knowledgebase, to construct stressor-species networks to assess ecological effects. We computed Species sensitivity distributions (SSDs) from aquatic toxicity data to estimate hazard concentrations (HCs) for these chemicals in aquatic environments. We further applied a novel approach by integrating hazard concentration data into the stressor-species network, enabling the identification of highly sensitive species. Finally, we computed risk quotients (RQs) for arsenic and cadmium in Indian rivers, contributing to ecological risk assessment. Notably, we compiled all the generated data and results and made it accessible at https://cb.imsc.res.in/heavymetaltox/. Overall, we integrated multiple network toxicology approaches across the One Health framework for inorganic arsenic and cadmium in the Indian context, providing valuable insights for risk assessment and mitigation strategies.

5.2 Future outlook

Human health is not only influenced by an individual's genetics, but is also shaped by the exposome, the totality of environmental exposures that an individual encounters throughout their lifetime [1]. Research on the human exposome is increasingly taking precedence to build a healthier world, with global efforts such as the Human Exposome Project that aims to systematically chart lifelong environmental exposures and their biological consequences, much like how the Human Genome Project did for genetics [7, 8]. Towards this, *in silico* approaches have emerged as powerful tools, allowing for faster testing strategies and guiding hypothesis generation for quicker solutions [8]. In this thesis, we explored the applications of different computational approaches, including network toxicology, to systematically characterize the chemical exposome and gain mechanistic insights into their induced adverse effects. Based on the research reported in this thesis, we envision several future directions that can further advance this understanding, and potentially assist in the broader goals of the Human Exposome Project in characterizing the chemical exposome and linking it to human health.

In Chapter 2, we characterized the chemical exposome of vitiligo through ViCEKb. We observed that many chemicals were reported to be present in medicine and health care products, while medical exposure was the most reported source of exposure for chemical triggers (Figure 2.2). This is particularly relevant in the context of adverse drug reactions (ADRs), which refers to the unintended or harmful effects attributed to the use of medicines, and are vital for pharmacovigilance across health sectors [298, 299]. Thus, such targeted characterization of drug-like compounds can help uncover their side effects, and can eventually aid in the development of ADR prediction strategies [300–302]. Although skin-specific transcriptomics studies were limited for these chemicals, we were still able to gain some insights into the potential mechanisms through which these chemicals may contribute to the etiology of vitiligo. Therefore, focused omics-based characterization of these chemicals can further enrich our understanding of the chemical-induced etiology of

this disease. Furthermore, ViCEKb has already served as a positive dataset to characterize other potential triggers of vitiligo [303]. Such approaches can further aid in the better screening and prioritization of chemicals, especially cosmetics, and assist in designing personalized recommendation systems regarding the vitiligo-triggering effects of chemicals on individuals.

In Chapter 3, we presented DEDuCTv3.0, an updated FAIR-compliant version of DEDuCT that integrates toxicological data from diverse sources, providing a single point of access to all information associated with EDCs. Such an integrated resource can serve as a valuable tool to drive future research towards their management and mitigation. The construction of EDC-AOP networks can provide a basis to organize known mechanisms of toxicity associated with such EDCs, and help design better testing strategies. Although in this chapter, we limited ourselves to developed AOPs within AOP-Wiki, computational approaches can also be utilized to integrate EDC-specific toxicity data within DEDuCT, and generate putative or computational AOPs, which can help capture more EDC-specific mechanisms that are otherwise not present in AOP-Wiki [304, 305]. In this update, we have also constructed the first large-scale toxicological knowledge graph for EDCs by integrating EDC-specific toxicological data from diverse sources. Such a construction also makes the data well-suited for artificial intelligence (AI)-based research on EDCs [306, 307]. For instance, the graph embedding-based deep learning techniques can be utilized to expand the scope of research, generate hypotheses, and predict newer associations that were otherwise not evident from the data itself [308].

In Chapter 4, we utilized different network toxicology frameworks to systematically organize inorganic arsenic- and cadmium-relevant data, gain an understanding of their complete source-to-outcome pathways, and investigate their effects on human and ecosystem health, thereby realizing the potential of computational approaches to go beyond human health and aid in the characterization of One Health. Importantly, the developed workflow can be extended to other environmental stressors, and help gain similar insights from a One Health perspective. Integration of bioaccumulation information from stressor-species net-

works, and subsequent analysis of biomagnification through food web networks [309, 310], can further highlight potential exposure pathways and risks associated with environmental stressors. Moreover, enriching AEP networks with additional exposure data can potentially help uncover organism-specific co-exposure patterns of various environmental stressors.

Overall, the systematic characterization and subsequent development of resources presented in this thesis have made the results and insights easily accessible to the scientific community, and support future efforts towards a more complete characterization of the chemical exposome. The structured network toxicology investigation provides a basis for characterizing environmental chemical stressors, and can be extended to other chemicals to build a more comprehensive understanding of the chemical exposome and its links to human and ecosystem health from a One Health perspective [30].

References

[1] Wild, C. P. Complementing the Genome with an "Exposome": The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology. *Cancer Epidemiology, Biomarkers & Prevention* **14**, 1847–1850 (2005).

[2] Miller, G. W. & Jones, D. P. The Nature of Nurture: Refining the Definition of the Exposome. *Toxicological Sciences* **137**, 1–2 (2014).

[3] Vrijheid, M. The exposome: a new paradigm to study the impact of environment on health. *Thorax* **69**, 876 (2014).

[4] Cecchi, L., D'Amato, G. & Annesi-Maesano, I. External exposome and allergic respiratory and skin diseases. *Journal of Allergy and Clinical Immunology* **141**, 846–857 (2018).

[5] Rappaport, S. M. & Smith, M. T. Environment and disease risks. *Science* **330**, 460–461 (2010).

[6] Handakas, E., Robinson, O. & Laine, J. E. The exposome approach to study children's health. *Current Opinion in Environmental Science Health* **32**, 100455 (2023).

[7] Burki, T. The human exposome project. *The Lancet* **406**, 16–17 (2025).

[8] Sillé, F. C. M., Belkadi, M., Koehler, K., Ali, J., Vasiliou, V., Sarigiannis, D. & Hartung, T. Charting exposomethics: a roadmap for the ethical foundations of the human exposome project. *Human Genomics* **20**, 34 (2026).

[9] Brander, S. M. Rethinking our chemical legacy and reclaiming our planet. *One Earth* **5**, 316–319 (2022).

[10] Wang, Z., Walker, G. W., Muir, D. C. G. & Nagatani-Yoshida, K. Toward a Global Understanding of Chemical Pollution: A First Comprehensive Analysis of National and Regional Chemical Inventories. *Environmental Science & Technology* **54**, 2575–2584 (2020).

[11] PubChem. https://pubchem.ncbi.nlm.nih.gov.

[12] Samanipour, S., Barron, L. P., van Herwerden, D., Praetorius, A., Thomas, K. V. & O'Brien, J. W. Exploring the Chemical Space of the Exposome: How Far Have We Gone? *JACS Au* **4**, 2412–2425 (2024).

[13] National Research Council. Toxicity Testing in the 21st Century: A Vision and a Strategy. *The National Academies Press, Washington, DC* (2007).

[14] Sewell, F. *et al.* New approach methodologies (NAMs): identifying and overcoming hurdles to accelerated adoption. *Toxicology Research* **13**, tfae044 (2024).

[15] National Research Council. Using 21st Century Science to Improve Risk-Related Evaluations. *The National Academies Press, Washington, DC* (2017).

[16] Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**, 160018 (2016).

[17] Huang, R. *et al.* Modelling the Tox21 10K chemical profiles for in vivo toxicity prediction and mechanism characterization. *Nature Communications* **7**, 10425 (2016).

[18] Dix, D. J., Houck, K. A., Martin, M. T., Richard, A. M., Setzer, R. W. & Kavlock, R. J. The ToxCast Program for Prioritizing Toxicity Testing of Environmental Chemicals. *Toxicological Sciences* **95**, 5–12 (2007).

[19] Feshuk, M., Kolaczkowski, L., Dunham, K., Davidson-Fritz, S. E., Carstens, K. E., Brown, J., Judson, R. S. & Paul Friedman, K. The ToxCast pipeline: updates to curve-fitting approaches and database structure. *Frontiers in Toxicology* **5**, 1275980 (2023).

[20] ToxCast & Tox21 Summary Files from invitrodb_v4.3, retrieved from https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data on November 12, 2025. Data relased August 2025.

[21] CompTox Chemicals Dashboard. https://comptox.epa.gov/dashboard/.

[22] Williams, A. J. *et al.* The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. *Journal of Cheminformatics* **9**, 61 (2017).

[23] Williams, A. J., Lambert, J. C., Thayer, K. & Dorne, J.-L. C. Sourcing data on chemical properties and hazard data from the US-EPA CompTox Chemicals Dashboard: A practical guide for human risk assessment. *Environment International* **154**, 106566 (2021).

[24] Davis, A. P. *et al.* Comparative Toxicogenomics Database's 20th anniversary: update 2025. *Nucleic Acids Research* **53**, D1328–D1334 (2025).

[25] CTD: Comparative Toxicogenomics Database. https://ctdbase.org.

[26] Keenan, A. B. *et al.* The Library of Integrated Network-Based Cellular Signatures NIH Program: System-Level Cataloging of Human Cells Response to Perturbations. *Cell Systems* **6**, 13–24 (2018).

[27] Pilarczyk, M. *et al.* Connecting omics signatures and revealing biological mechanisms with iLINCS. *Nature Communications* **13**, 4678 (2022).

[28] Vermeulen, R., Schymanski, E. L., Barabási, A.-L. & Miller, G. W. The exposome and health: Where chemistry meets biology. *Science* **367**, 392–396 (2020).

[29] Rusyn, I. & Daston, G. P. Computational Toxicology: Realizing the Promise of the Toxicity Testing in the 21st Century. *Environmental Health Perspectives* **118**, 1047–1050 (2010).

[30] Gao, P. The Exposome in the Era of One Health. *Environmental Science & Technology* **55**, 2790–2799 (2021).

[31] Ibanez, A. *et al.* Neuroecological links of the exposome and One Health. *Neuron* **112**, 1905–1910 (2024).

[32] Zhang, W. Network toxicology: A new science. *Computational Ecology and Software* **6**, 31–40 (2016).

[33] del Giudice, G. *et al.* A Network Toxicology Approach for Mechanistic Modelling of Nanomaterial Hazard and Adverse Outcomes. *Advanced Science* **11**, 2400389 (2024).

[34] Buttke, D. E. Toxicology, Environmental Health, and the "One Health" Concept. *Journal of Medical Toxicology* **7**, 329–332 (2011).

[35] Hogan, A. *et al.* Knowledge Graphs. *ACM Computing Surveys* **54** (2021).

[36] Min, W., Liu, C., Xu, L. & Jiang, S. Applications of knowledge graphs for food science and industry. *Patterns* **3**, 100484 (2022).

[37] Peng, C., Xia, F., Naseriparsa, M. & Osborne, F. Knowledge Graphs: Opportunities and Challenges. *Artificial Intelligence Review* **56**, 13071–13102 (2023).

[38] Delmas, M. *et al.* FORUM: building a Knowledge Graph from public databases and scientific literature to extract associations between chemicals and diseases. *Bioinformatics* **37**, 3896–3904 (2021).

[39] Evangelista, J. E. *et al.* Toxicology knowledge graph for structural birth defects. *Communications Medicine* **3**, 98 (2023).

[40] Hao, Y., Romano, J. D. & Moore, J. H. Knowledge graph aids comprehensive explanation of drug and chemical toxicity. *CPT: Pharmacometrics & Systems Pharmacology* **12**, 1072–1079 (2023).

[41] Sepehri, S. *et al.* The TOXIN knowledge graph: supporting animal-free risk assessment of cosmetics. *Database* **2025**, baae121 (2025).

[42] Teeguarden, J. G. *et al.* Completing the Link between Exposure Science and Toxicology for Improved Environmental Health Decision Making: The Aggregate Exposure Pathway Framework. *Environmental Science & Technology* **50**, 4579–4586 (2016).

[43] Tan, Y.-M., Leonard, J. A., Edwards, S., Teeguarden, J. & Egeghy, P. Refining the aggregate exposure pathway. *Environmental Science: Processes & Impacts* **20**, 428–436 (2018).

[44] Peng, G., Lin, Y., van Bavel, B., Li, D., Ni, J. & Song, Y. Aggregate exposure pathways for microplastics (mpAEP): An evidence-based framework to identify research and regulatory needs. *Water Research* **209**, 117873 (2022).

[45] Jeong, J., Im, J. & Choi, J. Integrating aggregate exposure pathway and adverse outcome pathway for micro/nanoplastics: A review on exposure, toxicokinetics, and toxicity studies. *Ecotoxicology and Environmental Safety* **272**, 116022 (2024).

[46] Wambaugh, J. F. *et al.* New approach methodologies for exposure science. *Current Opinion in Toxicology* **15**, 76–92 (2019).

[47] Ankley, G. T. *et al.* Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment. *Environmental Toxicology and Chemistry* **29**, 730–741 (2010).

[48] Villeneuve, D. L. *et al.* Adverse Outcome Pathway (AOP) Development I: Strategies and Principles. *Toxicological Sciences* **142**, 312–320 (2014).

[49] Villeneuve, D. L. *et al.* Adverse Outcome Pathway Development II: Best Practices. *Toxicological Sciences* **142**, 321–330 (2014).

[50] AOP-Wiki. https://aopwiki.org.

[51] AOP-Wiki Developers' Handbook. https://aopwiki.org/handbooks/5 (2024).

[52] Knapen, D. *et al.* Adverse outcome pathway networks I: Development and applications. *Environmental Toxicology and Chemistry* **37**, 1723–1733 (2018).

[53] Villeneuve, D. L. *et al.* Adverse outcome pathway networks II: Network analytics. *Environmental Toxicology and Chemistry* **37**, 1734–1748 (2018).

[54] Ravichandran, J., Karthikeyan, B. S. & Samal, A. Investigation of a derived adverse outcome pathway (AOP) network for endocrine-mediated perturbations. *Science of the Total Environment* **826**, 154112 (2022).

[55] Sahoo, A. K., Chivukula, N., Ramesh, K., Singha, J., Marigoudar, S. R., Sharma, K. V. & Samal, A. An integrative data-centric approach to derivation and characterization of an adverse outcome pathway network for cadmium-induced toxicity. *Science of The Total Environment* **920**, 170968 (2024).

[56] Sahoo, A. K., Chivukula, N., Madgaonkar, S. R., Ramesh, K., Marigoudar, S. R., Sharma, K. V. & Samal, A. Leveraging integrative toxicogenomic approach towards development of stressor-centric adverse outcome pathway networks for plastic additives. *Archives of Toxicology* **98**, 3299–3321 (2024).

[57] Sahoo, A. K., Madgaonkar, S. R., Chivukula, N., Karthikeyan, P., Ramesh, K., Marigoudar, S. R., Sharma, K. V. & Samal, A. Network-based investigation of petroleum hydrocarbons-induced ecotoxicological effects and their risk assessment. *Environment International* **194**, 109163 (2024).

[58] Hines, D. E., Conolly, R. B. & Jarabek, A. M. A Quantitative Source-to-Outcome Case Study To Demonstrate the Integration of Human Health and Ecological End

Points Using the Aggregate Exposure Pathway and Adverse Outcome Pathway Frameworks. *Environmental Science & Technology* **53**, 11002–11012 (2019).

[59] Price, P. S., Jarabek, A. M. & Burgoon, L. D. Organizing mechanism-related information on chemical interactions using a framework based on the aggregate exposure and adverse outcome pathways. *Environment International* **138**, 105673 (2020).

[60] Bruder, A., Frainer, A., Rota, T. & Primicerio, R. The Importance of Ecological Networks in Multiple-Stressor Research and Management. *Frontiers in Environmental Science* **7**, 59 (2019).

[61] Wang, S., Li, C., Zhang, L., Chen, Q. & Wang, S. Assessing the ecological impacts of polycyclic aromatic hydrocarbons petroleum pollutants using a network toxicity model. *Environmental Research* **245**, 117901 (2024).

[62] Wedding, L. M. *et al.* Linking multiple stressor science to policy opportunities through network modeling. *Marine Policy* **146**, 105307 (2022).

[63] Olker, J. H. *et al.* The ECOTOXicology Knowledgebase: A Curated Database of Ecologically Relevant Toxicity Tests to Support Environmental Research and Risk Assessment. *Environmental Toxicology and Chemistry* **41**, 1520–1539 (2022).

[64] Posthuma, L., Suter II, G. W. & Traas, T. P. (eds.) *Species Sensitivity Distributions in Ecotoxicology* (CRC Press, Boca Raton, 2001).

[65] Dowse, R., Tang, D., Palmer, C. G. & Kefford, B. J. Risk assessment using the species sensitivity distribution method: Data quality versus data quantity. *Environmental Toxicology and Chemistry* **32**, 1360–1369 (2013).

[66] Kamo, M. Species Sensitivity Distribution in Ecological Risk Assessment. In Kamo, M. (ed.) *Theories in Ecological Risk Assessment*, 103–134 (Springer Nature Singapore, Singapore, 2023).

[67] US EPA. Guidelines for Ecological Risk Assessment. https://www.epa.gov/risk/guidelines-ecological-risk-assessment (1998).

[68] Giddings, J. M., Wirtz, J., Campana, D. & Dobbs, M. Derivation of combined species sensitivity distributions for acute toxicity of pyrethroids to aquatic animals. *Ecotoxicology* **28**, 242–250 (2019).

[69] Lambert, F. N., Raimondo, S. & Barron, M. G. Assessment of a New Approach Method for Grouped Chemical Hazard Estimation: The Toxicity-Normalized Species Sensitivity Distribution (SSDn). *Environmental Science & Technology* **56**, 8278–8289 (2022).

[70] European Commission. Technical Guidance Document on Risk Assessment Part II. Tech. Rep., Office for Official Publications of the European Communities L - 2985 Luxembourg (2003).

[71] Cristale, J., Katsoyiannis, A., Sweetman, A. J., Jones, K. C. & Lacorte, S. Occurrence and risk assessment of organophosphorus and brominated flame retardants in the River Aire (UK). *Environmental Pollution* **179**, 194–200 (2013).

[72] Gao, P., Li, Z., Gibson, M. & Gao, H. Ecological risk assessment of nonylphenol in coastal waters of China based on species sensitivity distribution model. *Chemosphere* **104**, 113–119 (2014).

[73] Chivukula, N., Ramesh, K., Subbaroyan, A., Sahoo, A. K., Dhanakoti, G. B., Ravichandran, J. & Samal, A. ViCEKb: Vitiligo-linked Chemical Exposome Knowledgebase. *Science of The Total Environment* **913**, 169711 (2024).

[74] Chivukula, N., Vashishth, S., Kandasamy, P., Madgaonkar, S. R. & Samal, A. A fair-compliant resource and toxicology knowledge graph for mechanistic exploration of endocrine disrupting chemicals associated toxicity. *Journal of Hazardous Materials Advances* **24**, 101660 (2026).

[75] Chivukula, N., Madgaonkar, S. R., Ramesh, K., Mangot, S., Karthikeyan, P., Marigoudar, S. R., Sharma, K. V. & Samal, A. Network toxicology focused investigation on the impacts of inorganic arsenic and cadmium on human and ecosystem health. *Environmental Toxicology and Chemistry* **45**, 468–489 (2025).

[76] Rodrigues, M., Ezzedine, K., Hamzavi, I., Pandya, A. G. & Harris, J. E. New discoveries in the pathogenesis and classification of vitiligo. *Journal of the American Academy of Dermatology* **77**, 1–13 (2017).

[77] Richmond, J. M. *et al.* Antibody blockade of IL-15 signaling has the potential to durably reverse vitiligo. *Science Translational Medicine* **10**, eaam7710 (2018).

[78] Alkhateeb, A., Fain, P. R., Thody, A., Bennett, D. C. & Spritz, R. A. Epidemiology of Vitiligo and Associated Autoimmune Diseases in Caucasian Probands and Their Families. *Pigment Cell Research* **16**, 208–214 (2003).

[79] Laberge, G., Mailloux, C. M., Gowan, K., Holland, P., Bennett, D. C., Fain, P. R. & Spritz, R. A. Early disease onset and increased risk of other autoimmune diseases in familial generalized vitiligo. *Pigment Cell Research* **18**, 300–305 (2005).

[80] Porter, J. R., Beuf, A. H., Lerner, A. & Nordlund, J. Psychosocial effect of vitiligo: A comparison of vitiligo patients with "normal" control subjects, with psoriasis patients, and with patients with other pigmentary disorders. *Journal of the American Academy of Dermatology* **15**, 220–224 (1986).

[81] Lilly, E., Lu, P. D., Borovicka, J. H., Victorson, D., Kwasny, M. J., West, D. P. & Kundu, R. V. Development and validation of a vitiligo-specific quality-of-life instrument (VitiQoL). *Journal of the American Academy of Dermatology* **69**, e11–e18 (2013).

[82] Spritz, R. A. & Andersen, G. H. Genetics of Vitiligo. *Vitiligo* **35**, 245–255 (2017).

[83] Jin, Y. *et al.* Genome-wide association analyses identify 13 new susceptibility loci for generalized vitiligo. *Nature Genetics* **44**, 676–680 (2012).

[84] Shen, C. *et al.* Genetic Susceptibility to Vitiligo: GWAS Approaches for Identifying Vitiligo Susceptibility Genes and Loci. *Frontiers in Genetics* **7**, 3 (2016).

[85] Lu, G., Hao, X., Chen, W.-H. & Mu, S. GAAD: A Gene and Autoimmiune Disease Association Database. *Bioinformatics Commons (I)* **16**, 252–261 (2018).

[86] Gupta, I., Narang, A., Singh, P., Manchanda, V., Khanna, S., Mukerji, M., Natarajan, V. T. & Dash, D. Vitivar: A locus specific database of vitiligo associated genes and variations. *Gene* **721**, 100018 (2019).

[87] Boissy, R. E. & Manga, P. On the Etiology of Contact/Occupational Vitiligo. *Pigment Cell Research* **17**, 208–214 (2004).

[88] Vrijman, C., Hosseinpour, D., Bakker, J. G., Wolkerstorfer, A., Bos, J. D., van der Veen, J. P. W. & Luiten, R. M. Provoking factors, including chemicals, in Dutch patients with vitiligo. *British Journal of Dermatology* **168**, 1003–1011 (2013).

[89] Odedra, S. & Yoo, J. The risk of chemical leucoderma with skin-lightening therapies. *Clinical and Experimental Dermatology* **46**, 1391–1393 (2021).

[90] Rmadi, N. *et al.* Role of chemical exposure in the incidence of vitiligo: a case–control study in Tunisia. *Libyan Journal of Medicine* **18**, 2132628 (2023).

[91] PubMed. https://pubmed.ncbi.nlm.nih.gov/.

[92] Page, M. J. *et al.* The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* **372**, n71 (2021).

[93] Chemical Abstracts Service Registry. https://www.cas.org/cas-data/cas-registry.

[94] Grulke, C. M., Williams, A. J., Thillanadarajah, I. & Richard, A. M. EPA's DSSTox database: History of development of a curated chemistry resource supporting computational toxicology research. *Computational Toxicology* **12**, 100096 (2019).

[95] DSSTox: Distributed Structure-Searchable Toxicity Database. https://www.epa.gov/comptox-tools/distributed-structure-searchable-toxicity-dsstox-database.

[96] Djoumbou Feunang, Y. *et al.* ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *Journal of Cheminformatics* **8**, 61 (2016).

[97] ClassyFire. http://classyfire.wishartlab.com/.

[98] Ghosh, S. & Mukhopadhyay, S. Chemical leucoderma: a clinico-aetiological study of 864 cases in the perspective of a developing country. *British Journal of Dermatology* **160**, 40–47 (2009).

[99] Lee, C. S. *et al.* Different effects of five depigmentary compounds, rhododendrol, raspberry ketone, monobenzone, rucinol and AP736 on melanogenesis and viability of human epidermal melanocytes. *Experimental Dermatology* **25**, 44–49 (2016).

[100] Herman, A., de Montjoye, L., Marot, L. & Baeck, M. Induction of leukoderma following allergic contact dermatitis to FreeStyle Libre. *Contact Dermatitis* **81**, 456–458 (2019).

[101] Gawkrodger, D. J., Cork, M. J. & Bleehen, S. S. Ocupational vitiligo and contact sensitivity to para-tertiary butyl catechol. *Contact Dermatitis* **25**, 200–201 (1991).

[102] Brown, T., Zirvi, M., Cotsarelis, G. & Gelfand, J. M. Vitiligo-like hypopigmentation associated with imiquimod treatment of genital warts. *Journal of the American Academy of Dermatology* **52**, 715–716 (2005).

[103] Anthony, N., Bourneau-martin, D., Ghamrawi, S., Lagarce, L., Babin, M. & Briet, M. Drug-induced vitiligo: a case/non-case study in Vigibase®, the WHO phar-

macovigilance database. *Fundamental & Clinical Pharmacology* **34**, 736–742 (2020).

[104] PubChem data sources. https://pubchem.ncbi.nlm.nih.gov/sources/.

[105] Karthikeyan, B. S., Ravichandran, J., Mohanraj, K., Vivek-Ananth, R. P. & Samal, A. A curated knowledgebase on endocrine disrupting chemicals and their biological systems-level perturbations. *Science of The Total Environment* **692**, 281–296 (2019).

[106] O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T. & Hutchison, G. R. Open Babel: An open chemical toolbox. *Journal of Cheminformatics* **3**, 33 (2011).

[107] RDKit: Open-Source Cheminformatics Software. https://www.rdkit.org/.

[108] Bemis, G. W. & Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *Journal of Medicinal Chemistry* **39**, 2887–2893 (1996).

[109] Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry* **32**, 1466–1474 (2011).

[110] PaDEL-Descriptor. http://www.yapcwsoft.com/dd/padeldescriptor/.

[111] Shen, J., Cheng, F., Xu, Y., Li, W. & Tang, Y. Estimation of ADME Properties with Substructure Pattern Recognition. *Journal of Chemical Information and Modeling* **50**, 1034–1041 (2010).

[112] Yang, H. *et al.* admetSAR 2.0: web-service for prediction and optimization of chemical ADMET properties. *Bioinformatics* **35**, 1067–1069 (2019).

[113] Daina, A., Michielin, O. & Zoete, V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Scientific Reports* **7**, 42717 (2017).

[114] Pires, D. E. V., Blundell, T. L. & Ascher, D. B. pkCSM: Predicting Small-Molecule Pharmacokinetic and Toxicity Properties Using Graph-Based Signatures. *Journal of Medicinal Chemistry* **58**, 4066–4072 (2015).

[115] Schyman, P., Liu, R., Desai, V. & Wallqvist, A. vNN Web Server for ADMET Predictions. *Frontiers in Pharmacology* **8**, 889 (2017).

[116] Dionisio, K. L., Phillips, K., Price, P. S., Grulke, C. M., Williams, A., Biryol, D., Hong, T. & Isaacs, K. K. The Chemical and Products Database, a resource for exposure-relevant data on chemicals in consumer products. *Scientific Data* **5**, 180125 (2018).

[117] CPDat: Chemical and Products Database. https://www.epa.gov/chemical-research/chemical-and-products-database-cpdat.

[118] Organisation for Economic Co-operation and Development (OECD). https://www.oecd.org/.

[119] Isaacs, K. K., Dionisio, K., Phillips, K., Bevington, C., Egeghy, P. & Price, P. S. Establishing a system of consumer product use categories to support rapid modeling of human exposure. *Journal of Exposure Science & Environmental Epidemiology* **30**, 171–183 (2020).

[120] Karthikeyan, B. S., Ravichandran, J., Aparna, S. R. & Samal, A. DEDuCT 2.0: An updated knowledgebase and an exploration of the current regulations and guidelines from the perspective of endocrine disrupting chemicals. *Chemosphere* **267**, 128898 (2021).

[121] DEDuCT: Database of Endocrine Disrupting Chemicals and their Toxicity Profiles. https://cb.imsc.res.in/deduct/.

[122] Ravichandran, J., Karthikeyan, B. S., Jost, J. & Samal, A. An atlas of fragrance chemicals in children's products. *Science of The Total Environment* **818**, 151682 (2022).

[123] FCCP. https://cb.imsc.res.in/fccp/.

[124] Karthikeyan, B. S., Ravichandran, J., Aparna, S. R. & Samal, A. ExHuMId: A curated resource and analysis of Exposome of Human Milk across India. *Chemosphere* **271**, 129583 (2021).

[125] ExHuMId. https://cb.imsc.res.in/exhumid/.

[126] Ravichandran, J., Karthikeyan, B. S., Singla, P., Aparna, S. R. & Samal, A. NeurotoxKb 1.0: Compilation, curation and exploration of a knowledgebase of environmental neurotoxicants specific to mammals. *Chemosphere* **278**, 130387 (2021).

[127] NeurotoxKb. https://cb.imsc.res.in/neurotoxkb/.

[128] Richard, A. M. *et al.* ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. *Chemical Research in Toxicology* **29**, 1225–1251 (2016).

[129] ToxCast & Tox21 Summary Files from invitrodb_v3. Retrieved from https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data on April 4, 2023. Data released August 2022.

[130] Tanimoto, T. T. IBM internal report. *Nov* **17** (1957).

[131] Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics* **7**, 20 (2015).

[132] Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research* **13**, 2498–2504 (2003).

[133] Yang, Z.-Y., Yang, Z.-J., Lu, A.-P., Hou, T.-J. & Cao, D.-S. Scopy: an integrated negative design python library for desirable HTS/VS database design. *Briefings in Bioinformatics* **22**, bbaa194 (2021).

[134] Vivek-Ananth, R., Sahoo, A. K., Baskaran, S. P. & Samal, A. Scaffold and Structural Diversity of the Secondary Metabolite Space of Medicinal Fungi. *ACS Omega* **8**, 3102–3113 (2023).

[135] Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. & Pfister, H. UpSet: Visualization of Intersecting Sets. *IEEE Transactions on Visualization and Computer Graphics* **20**, 1983–1992 (2014).

[136] iLINCS. https://www.ilincs.org/ilincs/.

[137] Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R. & Ma'ayan, A. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).

[138] Enrichr. https://maayanlab.cloud/Enrichr/.

[139] Gillespie, M. *et al.* The reactome pathway knowledgebase 2022. *Nucleic Acids Research* **50**, D687–D692 (2022).

[140] The Gene Ontology Consortium *et al.* The Gene Ontology knowledgebase in 2023. *Genetics* **224**, iyad031 (2023).

[141] MariaDB. https://mariadb.org.

[142] php. https://www.php.net.

[143] jQuery. https://jquery.com.

[144] Bootstrap v5.0. https://getbootstrap.com/docs/5.0/getting-started/introduction/.

[145] Google Charts. https://developers.google.com/chart/.

[146] Apache web server. https://www.apache.org.

[147] Lin, D. *et al.* The TRUST Principles for digital repositories. *Scientific Data* **7**, 144 (2020).

[148] Donner, Y., Kazmierczak, S. & Fortney, K. Drug Repurposing Using Deep Embeddings of Gene Expression Profiles. *Molecular Pharmaceutics* **15**, 4314–4325 (2018).

[149] Szilveszter, K. P., Németh, T. & Mócsai, A. Tyrosine Kinases in Autoimmune and Inflammatory Skin Diseases. *Frontiers in Immunology* **10**, 1862 (2019).

[150] Haz-map: Information on hazardous chemicals and occupational diseases. https://haz-map.com.

[151] Davis, A. P., Wiegers, T. C., Johnson, R. J., Sciaky, D., Wiegers, J. & Mattingly, C. Comparative Toxicogenomics Database (CTD): update 2023. *Nucleic Acids Research* **51**, D1257–D1262 (2023).

[152] IFRA: The International Fragrance Association. https://ifrafragrance.org.

[153] CSCP: California Safe Cosmetics Program Product Database. https://cscpsearch.cdph.ca.gov/search/publicsearch.

[154] Solecki, R. *et al.* Scientific principles for the identification of endocrine-disrupting chemicals: a consensus statement. *Archives of Toxicology* **91**, 1001–1006 (2017).

[155] Bertram, M. G., Gore, A. C., Tyler, C. R. & Brodin, T. Endocrine-disrupting chemicals. *Current Biology* **32**, R727–R730 (2022).

[156] Svingen, T. *et al.* Enhanced identification of endocrine disruptors through integration of science-based regulatory practices and innovative methodologies: The MERLON Project. *Open Research Europe* **4**, 68 (2024).

[157] ANSES. Elaboration of a list of substances of interest as regards to a potential endocrine activity and prioritisation strategy for assessment. Tech. Rep. 2019-SA-0179, ANSES, France (2021).

[158] Sakhteman, A., Failli, M., Kublbeck, J., Levonen, A. & Fortino, V. A toxicogenomic data space for system-level understanding and prediction of EDC-induced toxicity. *Environment International* **156**, 106751 (2021).

[159] Yu, Z., Wu, Z., Zhou, M., Cao, K., Li, W., Liu, G. & Tang, Y. EDC-Predictor: A Novel Strategy for Prediction of Endocrine-Disrupting Chemicals by Integrating Pharmacological and Toxicological Profiles. *Environmental Science & Technology* **57**, 18013–18025 (2023).

[160] WHO/UNEP. *State of the Science of Endocrine Disrupting Chemicals 2012. An assessment of the State of the Science of Endocrine Disruptors Prepared by a. Group of Experts for the United Nations Environment Programme (UNEP) and WHO, Geneva, Switzerland).* (2013).

[161] TEDx: The endocrine disruption exchange. https://endocrinedisruption.org/.

[162] Montes-Grajales, D. & Olivero-Verbel, J. EDCs DataBank: 3D-Structure database of endocrine disrupting chemicals. *Toxicology* **327**, 87–94 (2015).

[163] Handa, S. *et al.* The Chemical and Products Database v4.0, an updated resource supporting chemical exposure evaluations. *Scientific Data* **12**, 950 (2025).

[164] Judson, R. *et al.* Analysis of the Effects of Cell Stress and Cytotoxicity on In Vitro Assay Activity Across a Diverse Chemical and Assay Space. *Toxicological Sciences* **152**, 323–339 (2016).

[165] Jaccard, P. The Distribution of the Flora in the Alpine Zone. *New Phytologist* **11**, 37–50 (1912).

[166] Aguayo-Orozco, A., Audouze, K., Siggaard, T., Barouki, R., Brunak, S. & Taboureau, O. sAOP: linking chemical stressors to adverse outcomes pathway networks. *Bioinformatics* **35**, 5391–5392 (2019).

[167] OECD. Revised Guidance Document on Developing and Assessing Adverse Outcome Pathways, Series on Testing and Assessment No. 184. *OECD Publishing, Paris* (2017).

[168] Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring Network Structure, Dynamics, and Function using NetworkX. In Varoquaux, G., Vaught, T. & Millman, J. (eds.) *Proceedings of the 7th Python in Science Conference (SciPy 2008)*, 11–15 (2008).

[169] Schriml, L. M. *et al.* Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Research* **47**, D955–D962 (2019).

[170] DO: Disease ontology. https://disease-ontology.org/do/.

[171] GO: evidence codes. https://geneontology.org/docs/guide-go-evidence-codes/.

[172] Fröhlich, H., Speer, N., Poustka, A. & Beißbarth, T. GOSim – an R-package for computation of information theoretic GO similarities between terms and gene products. *BMC Bioinformatics* **8**, 166 (2007).

[173] MeSH Broswer. https://meshb.nlm.nih.gov/.

[174] Chai, Z. *et al.* Generating adverse outcome pathway (AOP) of inorganic arsenic-induced adult male reproductive impairment via integration of phenotypic analysis in comparative toxicogenomics database (CTD) and AOP wiki. *Toxicology and Applied Pharmacology* **411**, 115370 (2021).

[175] NCBI Taxonomy browser. https://www.ncbi.nlm.nih.gov/taxonomy.

[176] CTD: Glossary. https://ctdbase.org/help/glossary.jsp.

[177] NCBI Gene - FTP. https://ftp.ncbi.nih.gov/gene.

[178] Saarimäki, L. A., Fratello, M., Pavel, A., Korpilähde, S., Leppänen, J., Serra, A. & Greco, D. A curated gene and biological system annotation of adverse outcome pathways related to human health. *Scientific Data* **10**, 409 (2023).

[179] BioMart. https://asia.ensembl.org/info/data/biomart/index.html?

[180] Neo4j - graph database. https://neo4j.com.

[181] Evangelista, J. E., Lutsky, A. D., Byrd, A. I., Clarke, D. J. B., Prabhakaran, A., Jenkins, S. L. & Ma'ayan, A. Creating an Interactive Web Interface for Networks Stored in Knowledge Graph Databases. *Current Protocols* **5**, e70200 (2025).

[182] Neo4j - Cypher manual. https://neo4j.com/docs/cypher-manual/current/introduction/.

[183] Knowledge Graph UI. https://github.com/MaayanLab/Knowledge-Graph-UI/.

[184] Davis, A. P., Wiegers, T. C., Grondin, C. J., Johnson, R. J., Sciaky, D., Wiegers, J. & Mattingly, C. J. Leveraging the Comparative Toxicogenomics Database to Fill in Knowledge Gaps for Environmental Health: A Test Case for Air Pollution-induced Cardiovascular Disease. *Toxicological Sciences* **177**, 392–404 (2020).

[185] Heindel, J. J. *et al.* Obesity II: Establishing causal links between chemical exposures and obesity. *Biochemical Pharmacology* **199**, 115015 (2022).

[186] Pharos. https://pharos.nih.gov.

[187] Kelleher, K. J. *et al.* Pharos 2023: an integrated resource for the understudied human proteome. *Nucleic Acids Research* **51**, D1405–D1416 (2023).

[188] Hotamisligil, G. S., Arner, P., Caro, J. F., Atkinson, R. L. & Spiegelman, B. M. Increased adipose tissue expression of tumor necrosis factor-alpha in human obesity and insulin resistance. *The Journal of Clinical Investigation* **95**, 2409–2415 (1995).

[189] Samad, F., Uysal, K. T., Wiesbrock, S. M., Pandey, M., Hotamisligil, G. S. & Loskutoff, D. J. Tumor necrosis factor α is a key component in the obesity-linked elevation of plasminogen activator inhibitor 1. *Proceedings of the National Academy of Sciences* **96**, 6902–6907 (1999).

[190] Tzanavari, T., Giannogonas, P. & Karalis, K. P. TNF-α and Obesity. In Kollias, G. & Sfikakis, P. (eds.) *TNF Pathophysiology: Molecular and Cellular Mechanisms*, vol. 11 (S.Karger AG, 2010).

[191] Li, X., Wang, S., Mao, X., Fang, M., Liu, X., Jiang, J. & Jin, W. AhR in biological processes of adipocytes and lipid metabolism in obesity: Friend and foe. *Life Sciences* **381**, 123996 (2025).

[192] Arsenescu, V., Arsenescu, R. I., King, V., Swanson, H. & Cassis, L. A. Polychlorinated biphenyl-77 induces adipocyte differentiation and proinflammatory adipokines and promotes obesity and atherosclerosis. *Environmental Health Perspectives* **116**, 761–768 (2008).

[193] Luo, X. *et al.* Obesity induces preadipocyte CD36 expression promoting inflammation via the disruption of lysosomal calcium homeostasis and lysosome function. *EBioMedicine* **56**, 102797 (2020).

[194] Xu, W.-H., Qu, Y.-Y., Wang, J., Wang, H.-K., Wan, F.-N., Zhao, J.-Y., Zhang, H.-L. & Ye, D.-W. Elevated CD36 expression correlates with increased visceral adipose tissue and predicts poor prognosis in ccRCC patients. *Journal of Cancer* **10**, 4522–4531 (2019).

[195] Błażejewska, W., Dąbrowska, J., Michałowska, J. & Bogdański, P. The Role of Adiponectin and ADIPOQ Variation in Metabolic Syndrome: A Narrative Review. *Genes* **16**, 699 (2025).

[196] Hu, E., Liang, P. & Spiegelman, B. M. AdipoQ Is a Novel Adipose-specific Gene Dysregulated in Obesity. *Journal of Biological Chemistry* **271**, 10697–10703 (1996).

[197] Wang, C., Xu, J., Pan, K., Xu, Y. & Yu, J. Relationship between endocrine disruptors and neurodegenerative diseases: Systematic review and meta-analysis. *iScience* **28**, 112779 (2025).

[198] Chin-Chan, M., Navarro-Yepes, J. & Quintanilla-Vega, B. Environmental pollutants as risk factors for neurodegenerative disorders: Alzheimer and Parkinson diseases. *Frontiers in Cellular Neuroscience* **9**, 124 (2015).

[199] Wójcik, P., Jastrzębski, M. K., Zięba, A., Matosiuk, D. & Kaczor, A. A. Caspases in Alzheimer's Disease: Mechanism of Activation, Role, and Potential Treatment. *Molecular Neurobiology* **61**, 4834–4853 (2024).

[200] Du, F. *et al.* PINK1 signalling rescues amyloid pathology and mitochondrial dysfunction in Alzheimer's disease. *Brain* **140**, 3233–3251 (2017).

[201] Zhu, Y., Zhang, J. & Zeng, Y. Overview of Tyrosine Hydroxylase in Parkinson's Disease. *CNS & Neurological Disorders - Drug Targets* **11**, 350–358 (2012).

[202] Lohr, K. M. & Miller, G. W. VMAT2 and Parkinson's disease: harnessing the dopamine vesicle. *Expert Review of Neurotherapeutics* **14**, 1115–1117 (2014).

[203] Pickrell, A. M. & Youle, R. J. The Roles of PINK1, Parkin, and Mitochondrial Fidelity in Parkinson's Disease. *Neuron* **85**, 257–273 (2015).

[204] Clausen, L., Okarmus, J., Voutsinos, V., Meyer, M., Lindorff-Larsen, K. & Hartmann-Petersen, R. PRKN-linked familial Parkinson's disease: cellular and molecular mechanisms of disease-linked variants. *Cellular and Molecular Life Sciences* **81**, 223 (2024).

[205] United States Agency for Toxic Substances and Disease Registry (ATSDR) - Substance Priority List. https://www.atsdr.cdc.gov/programs/substance-priority-list.html.

[206] Balali-Mood, M., Naseri, K., Tahergorabi, Z., Khazdair, M. R. & Sadeghi, M. Toxic Mechanisms of Five Heavy Metals: Mercury, Lead, Chromium, Cadmium, and Arsenic. *Frontiers in Pharmacology* **12**, 643972 (2021).

[207] Camacho, J., de Conti, A., Pogribny, I. P., Sprando, R. L. & Hunt, P. R. Assessment of the effects of organic vs. inorganic arsenic and mercury in Caenorhabditis elegans. *Current Research in Toxicology* **3**, 100071 (2022).

[208] Hispard, F. *et al.* Effects of subchronic digestive exposure to organic or inorganic cadmium on biomarkers in rat tissues. *Ecotoxicology and Environmental Safety* **70**, 490–498 (2008).

[209] Tchounwou, P. B., Yedjou, C. G., Patlolla, A. K. & Sutton, D. J. Heavy Metal Toxicity and the Environment. In Luch, A. (ed.) *Molecular, Clinical and Environmental Toxicology: Volume 3: Environmental Toxicology*, 133–164 (Springer Basel, Basel, 2012).

[210] Chakraborti, D. *et al.* Groundwater arsenic contamination in Ganga–Meghna–Brahmaputra plain, its health effects and an approach for mitigation. *Environmental Earth Sciences* **70**, 1993–2008 (2013).

[211] Gupta, A. & Singh, E. J. Arsenic–Iron Relationships in Aquifers of North East India: Implications for Public Health and the Environment. *Environmental Management* **63**, 437–443 (2019).

[212] Rahman, A., Mondal, N. C. & Fauzia, F. Arsenic enrichment and its natural background in groundwater at the proximity of active floodplains of Ganga River, northern India. *Chemosphere* **265**, 129096 (2021).

[213] Saha, D. & Sahu, S. A decade of investigations on groundwater arsenic contamination in Middle Ganga Plain, India. *Environmental Geochemistry and Health* **38**, 315–337 (2016).

[214] Bhardwaj, S., Soni, R., Gupta, S. K. & Shukla, D. P. Mercury, arsenic, lead and cadmium in waters of the Singrauli coal mining and power plants industrial zone, Central East India. *Environmental Monitoring and Assessment* **192**, 251 (2020).

[215] Sharma, S., Nagpal, A. K. & Kaur, I. Heavy metal contamination in soil, food crops and associated health risks for residents of Ropar wetland, Punjab, India and its environs. *Food Chemistry* **255**, 15–22 (2018).

[216] Sudhakar, G., Jyothi, B. & Venkateswarlu, V. Metal pollution and its impact on algae in flowing waters in India. *Archives of Environmental Contamination and Toxicology* **21**, 556–566 (1991).

[217] ECOTOX Knowledgebase. https://cfpub.epa.gov/ecotox/.

[218] AOP-Wiki: Project Downloads. https://aopwiki.org/downloads.

[219] Jeong, J. & Choi, J. Development of AOP relevant to microplastics based on toxicity mechanisms of chemical additives using ToxCast™ and deep learning models combined approach. *Environment International* **137**, 105557 (2020).

[220] ToxCast & Tox21 Summary Files from invitrodb_v4.1, retrieved from https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data on June 25, 2024. Data relased Septmeber 2023.

[221] Davis, A. P., Wiegers, T. C., Grondin, C. J., Johnson, R. J., Sciaky, D., Wiegers, J. & Mattingly, C. J. Leveraging the Comparative Toxicogenomics Database to Fill in Knowledge Gaps for Environmental Health: A Test Case for Air Pollution-induced Cardiovascular Disease. *Toxicological Sciences* **177**, 392–404 (2020).

[222] Jeong, J., Kim, D. & Choi, J. Integrative Data Mining Approach: Case Study with Adverse Outcome Pathway Network Leading to Pulmonary Fibrosis. *Chemical Research in Toxicology* **36**, 838–847 (2023).

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

[224] Methods for Measuring the Acute Toxicity of Effluents and Receiving Waters to Freshwater and Marine Organisms. Fifth Edition. Tech. Rep. 821-R-02-012, U.S. Environmental Protection Agency (2002).

[225] Schwarz, C. J. & Tillmanns, A. R. Improving statistical methods to derive species sensitivity distribution. Tech. Rep. Water Science Series, WSS2019-07, Province of British Columbia, Victoria (2019).

[226] Wheeler, J. R., Grist, E. P. M., Leung, K. M. Y., Morritt, D. & Crane, M. Species sensitivity distributions: data and model choice. *Marine Pollution Bulletin* **45**, 192–202 (2002).

[227] Belanger, S. *et al.* Future needs and recommendations in the development of species sensitivity distributions: Estimating toxicity thresholds for aquatic ecological communities and assessing impacts of chemical exposures. *Integrated Environmental Assessment and Management* **13**, 664–674 (2017).

[228] U.S. EPA Species Sensitivity Distribution (SSD) Toolbox. https://www.epa.gov/comptox-tools/species-sensitivity-distribution-ssd-toolbox.

[229] Cavanaugh, J. E. & Neath, A. A. The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. *WIREs Computational Statistics* **11**, e1460 (2019).

[230] Etterson, M. *Technical Manual: SSD Toolbox Version 1.0* (US Environmental Protection Agency, 2020).

[231] Fox, D. R. *et al.* Recent Developments in Species Sensitivity Distribution Modeling. *Environmental Toxicology and Chemistry* **40**, 293–308 (2021).

[232] Central Water Commission. Status of Trace & Toxic Metals in Indian Rivers (2019). https://cwc.gov.in/sites/default/files/status-trace-toxic-metals-indian-rivers-2019-2.pdf.

[233] Gredelj, A., Barausse, A., Grechi, L. & Palmeri, L. Deriving predicted no-effect concentrations (PNECs) for emerging contaminants in the river Po, Italy, using three approaches: Assessment factor, species sensitivity distribution and AQUATOX ecosystem modelling. *Environment International* **119**, 66–78 (2018).

[234] Li, H., Meng, F. & Li, A. Ecological risk assessment for xylenes and propylbenzenes in aquatic environment using a species sensitivity distribution approach. *Ecotoxicology and Environmental Safety* **261**, 115106 (2023).

[235] Park, J. *et al.* Probabilistic ecological risk assessment of heavy metals using the sensitivity of resident organisms in four Korean rivers. *Ecotoxicology and Environmental Safety* **183**, 109483 (2019).

[236] Park, J. & Kim, S. D. Derivation of Predicted No Effect Concentrations (PNECs) for Heavy Metals in Freshwater Organisms in Korea Using Species Sensitivity Distributions (SSDs). *Minerals* **10**, 697 (2020).

[237] Rizzi, C., Villa, S., Cuzzeri, A. S. & Finizio, A. Use of the Species Sensitivity Distribution Approach to Derive Ecological Threshold of Toxicological Concern (eco-TTC) for Pesticides. *International Journal of Environmental Research and Public Health* **18**, 12078 (2021).

[238] Shapiro, S. S. & Wilk, M. B. An analysis of variance test for normality (complete samples). *Biometrika* **52**, 591–611 (1965).

[239] Hsu, H. & Lachenbruch, P. A. Paired t Test. In *Wiley StatsRef: Statistics Reference Online* (2014).

[240] Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* **1**, 80–83 (1945).

[241] ECHA: Substances restricted under REACH. https://echa.europa.eu/substances-restricted-under-reach.

[242] ECHA - Candidate List of substances of very high concern for Authorisation. https://echa.europa.eu/candidate-list-table.

[243] Neveu, V., Nicolas, G., Salek, R. M., Wishart, D. S. & Scalbert, A. Exposome-Explorer 2.0: an update incorporating candidate dietary biomarkers and dietary associations with cancer risk. *Nucleic Acids Research* **48**, D908–D912 (2020).

[244] Ravichandran, J., Karthikeyan, B. S., Aparna, S. R. & Samal, A. Network biology approach to human tissue-specific chemical exposome. *The Journal of Steroid Biochemistry and Molecular Biology* **214**, 105998 (2021).

[245] Allen, H. E. The significance of trace metal speciation for water, sediment and soil quality criteria and standards. *Science of The Total Environment* **134**, 23–45 (1993).

[246] Pierrot, D. & Millero, F. J. The Speciation of Metals in Natural Waters. *Aquatic Geochemistry* **23**, 1–20 (2017).

[247] Hughes, M. F., Beck, B. D., Chen, Y., Lewis, A. S. & Thomas, D. J. Arsenic Exposure and Toxicology: A Historical Perspective. *Toxicological Sciences* **123**, 305–332 (2011).

[248] McGeer, J. C., Niyogi, S. & Scott Smith, D. 3 - Cadmium. In Wood, C. M., Farrell, A. P. & Brauner, C. J. (eds.) *Fish Physiology*, vol. 31, 125–184 (Academic Press, 2011).

[249] Sunda, W. G., Engel, D. W. & Thuotte, R. M. Effect of chemical speciation on toxicity of cadmium to grass shrimp, Palaemonetes pugio: importance of free cadmium ion. *Environmental Science & Technology* **12**, 409–413 (1978).

[250] Banerjee, M. *et al.* High arsenic in rice is associated with elevated genotoxic effects in humans. *Scientific Reports* **3**, 2195 (2013).

[251] Liao, C.-M., Lin, T.-L., Hsieh, N.-H. & Chen, W.-Y. Assessing the arsenic-contaminated rice (*Oryza sativa*) associated children skin lesions. *Journal of Hazardous Materials* **176**, 239–251 (2010).

[252] Pal, A., Chowdhury, U. K., Mondal, D., Das, B., Nayak, B., Ghosh, A., Maity, S. & Chakraborti, D. Arsenic Burden from Cooked Rice in the Populations of Arsenic Affected and Nonaffected Areas and Kolkata City in West-Bengal, India. *Environmental Science & Technology* **43**, 3349–3355 (2009).

[253] Samal, A. C., Kar, S., Bhattacharya, P. & Santra, S. C. Human exposure to arsenic through foodstuffs cultivated using arsenic contaminated groundwater in areas of West Bengal, India. *Journal of Environmental Science and Health, Part A* **46**, 1259–1265 (2011).

[254] Signes-Pastor, A. J., Mitra, K., Sarkhel, S., Hobbes, M., Burló, F., de Groot, W. T. & Carbonell-Barrachina, A. A. Arsenic Speciation in Food and Estimation of the Dietary Intake of Inorganic Arsenic in a Rural Village of West Bengal, India. *Journal of Agricultural and Food Chemistry* **56**, 9469–9474 (2008).

[255] Biswas, P., Sahu, A. K., Sahoo, S. K., Ahmed, S. N., Shristi, S., Naik, P. K. & Baitharu, I. Prevalence of chronic kidney disease and anemia in Hirakud Command

Area, Odisha, India: unveiling the role of environmental toxicants. *Journal of Nephrology* **38**, 563–577 (2025).

[256] Das, S., Nath, M., Laskar, A. K., DebRoy, S., Deb, S., Barhai, A. & Choudhury, A. P. Lead and cadmium exposure network in children in a periurban area in India: susceptibility and health risk. *Environmental Science and Pollution Research* **28**, 28133–28145 (2021).

[257] Ghosh, S., Mondal, S., Mandal, J., Mukherjee, A. & Bhattacharyya, P. Effect of metal fractions on rice grain metal uptake and biological parameters in mica mines waste contaminated soils. *Journal of Environmental Sciences* **136**, 313–324 (2024).

[258] Takes, F. W. & Kosters, W. A. Determining the Diameter of Small World Networks. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, 1191–1196 (Association for Computing Machinery, New York, NY, USA, 2011).

[259] Baker, N., Knudsen, T. & Williams, A. Abstract Sifter: a comprehensive front-end system to PubMed. *F1000Research* **6**, 2164 (2017).

[260] Jaylet, T., Coustillet, T., Jornod, F., Margaritte-Jeannin, P. & Audouze, K. AOP-helpFinder 2.0: Integration of an event-event searches module. *Environment International* **177**, 108017 (2023).

[261] Jornod, F., Jaylet, T., Blaha, L., Sarigiannis, D., Tamisier, L. & Audouze, K. AOP-helpFinder webserver: a tool for comprehensive analysis of the literature to support adverse outcome pathways development. *Bioinformatics* **38**, 1173–1175 (2022).

[262] Jensen, M. A., Blatz, D. J. & LaLone, C. A. Defining the Biologically Plausible Taxonomic Domain of Applicability of an Adverse Outcome Pathway: A Case Study Linking Nicotinic Acetylcholine Receptor Activation to Colony Death. *Environmental Toxicology and Chemistry* **42**, 71–87 (2023).

[263] Zhang, L., Gao, Y., Wu, S., Zhang, S., Smith, K. R., Yao, X. & Gao, H. Global impact of atmospheric arsenic on health risk: 2005 to 2015. *Proceedings of the National Academy of Sciences* **117**, 13975–13982 (2020).

[264] Agency for Toxic Substances and Disease Registry. Toxicological profile for arsenic (2007). https://www.atsdr.cdc.gov/toxprofiles/tp2.pdf.

[265] Sanchez, T. R., Perzanowski, M. & Graziano, J. H. Inorganic arsenic and respiratory health, from early life exposure to sex-specific effects: A systematic review. *Environmental Research* **147**, 537–555 (2016).

[266] Liu, Y.-F., Tang, M.-M., Sun, J., Li, J.-F., Jiang, Y.-L., Zhao, H. & Fu, L. Arsenic exposure and lung function decline in chronic obstructive pulmonary disease patients: The mediating influence of systematic inflammation and oxidative stress. *Food and Chemical Toxicology* **181**, 114044 (2023).

[267] Hosseini, M.-J., Shaki, F., Ghazi-Khansari, M. & Pourahmad, J. Toxicity of Arsenic (III) on Isolated Liver Mitochondria: A New Mechanistic Approach. *Iranian Journal of Pharmaceutical Research* **12**, 121 – 138 (2013).

[268] Stevens, J. J., Graham, B., Dugo, E. B., Berhaneselassie-Sumner, B., Ndebele, K. & Tchounwou, P. B. Arsenic Trioxide Induces Apoptosis via Specific Signaling Pathways in HT-29 Colon Cancer Cells. *Journal of Cancer Science & Therapy* **9**, 298 – 306 (2016).

[269] Bomberger, J. M., Coutermarsh, B. A., Barnaby, R. L. & Stanton, B. A. Arsenic Promotes Ubiquitinylation and Lysosomal Degradation of Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) Chloride Channels in Human Airway Epithelial Cells*. *Journal of Biological Chemistry* **287**, 17130–17139 (2012).

[270] Bustamante-Marin, X. M. & Ostrowski, L. E. Cilia and Mucociliary Clearance. *Cold Spring Harbor Perspectives in Biology* **9**, a028241 (2017).

[271] Parvez, F. *et al.* Arsenic Exposure and Impaired Lung Function. Findings from a Large Population-based Prospective Cohort Study. *American Journal of Respiratory and Critical Care Medicine* **188**, 813–819 (2013).

[272] Selvaraj, V., Armistead, M. Y., Cohenford, M. & Murray, E. Arsenic trioxide (As2O3) induces apoptosis and necrosis mediated cell death through mitochondrial membrane potential damage and elevated production of reactive oxygen species in PLHC-1 fish cell line. *Chemosphere* **90**, 1201–1209 (2013).

[273] Fattorini, D. & Regoli, F. Arsenic speciation in tissues of the mediterranean polychaete Sabella spallanzanii. *Environmental Toxicology and Chemistry* **23**, 1881–1887 (2004).

[274] Shukla, J. P. & Pandey, K. Impaired ovarian functions in arsenic-treated freshwater fish, Colisa fasciatus (BL. and SCH.). *Toxicology Letters* **20**, 1–3 (1984).

[275] Zhang, W. *et al.* Arsenic bioaccumulation and biotransformation in aquatic organisms. *Environment International* **163**, 107221 (2022).

[276] Atlani, M. *et al.* Heavy metal association with chronic kidney disease of unknown cause in central India-results from a case-control study. *BMC Nephrology* **25**, 120 (2024).

[277] Adams, S. V. & Newcomb, P. A. Urinary Cadmium as a Marker of Exposure in Epidemiological Studies. *Environmental Health Perspectives* **121**, A296 (2013).

[278] Mao, W. P. *et al.* Cadmium directly induced mitochondrial dysfunction of human embryonic kidney cells. *Human & Experimental Toxicology* **30**, 920–929 (2011).

[279] Xie, J. & Shaikh, Z. A. Cadmium-Induced Apoptosis in Rat Kidney Epithelial Cells Involves Decrease in Nuclear Factor-Kappa B Activity. *Toxicological Sciences* **91**, 299–308 (2006).

[280] Teeyakasem, W., Nishijo, M., Honda, R., Satarug, S., Swaddiwudhipong, W. & Ruangyuttikarn, W. Monitoring of cadmium toxicity in a Thai population with high-level environmental exposure. *Toxicology Letters* **169**, 185–195 (2007).

[281] Wang, D., Sun, H., Wu, Y., Zhou, Z., Ding, Z., Chen, X. & Xu, Y. Tubular and glomerular kidney effects in the Chinese general population with low environmental cadmium exposure. *Chemosphere* **147**, 3–8 (2016).

[282] Luo, J., Pei, S., Jing, W., Zou, E. & Wang, L. Cadmium inhibits molting of the freshwater crab Sinopotamon henanense by reducing the hemolymph ecdysteroid content and the activities of chitinase and N-acetyl-β-glucosaminidase in the epidermis. *Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology* **169**, 1–6 (2015).

[283] Huang, Q.-S. *et al.* Cloning and tissue expressions of seven chitinase family genes in Litopenaeus vannamei. *Fish & Shellfish Immunology* **29**, 75–81 (2010).

[284] Bodar, C. W. M., Voogt, P. A. & Zandee, D. I. Ecdysteroids in Daphnia magna: their role in moulting and reproduction and their levels upon exposure to cadmium. *Aquatic Toxicology* **17**, 339–350 (1990).

[285] Ortega, P., Vitorino, H. A., Green, S., Zanotto, F. P., Chung, J. S. & Moreira, R. G. Experimental effects of cadmium on physiological response of Callinectes danae (Crustacea, Portunidae) from environments with different levels of Cd contamination. *Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology* **251**, 109210 (2022).

[286] Zhang, C., Jin, Y., Yu, Y., Xiang, J. & Li, F. Cadmium-induced oxidative stress, metabolic dysfunction and metal bioaccumulation in adult palaemonid shrimp Palaemon macrodactylus (Rathbun, 1902). *Ecotoxicology and Environmental Safety* **208**, 111591 (2021).

[287] Karthikeyan, P., Marigoudar, S. R., Mohan, D., Sharma, K. V. & Ramana Murthy, M. V. Prescribing sea water quality criteria for arsenic, cadmium and lead through species sensitivity distribution. *Ecotoxicology and Environmental Safety* **208**, 111612 (2021).

[288] Méndez-Fernández, L., Casado-Martínez, C., Martínez-Madrid, M., Moreno-Ocio, I., Costas, N., Pardo, I. & Rodriguez, P. Derivation of sediment Hg quality standards based on ecological assessment in river basins. *Environmental Pollution* **245**, 1000–1013 (2019).

[289] Yanagihara, M., Hiki, K. & Iwasaki, Y. Can Chemical Toxicity in Saltwater Be Predicted from Toxicity in Freshwater? A Comprehensive Evaluation Using Species Sensitivity Distributions. *Environmental Toxicology and Chemistry* **41**, 2021–2027 (2022).

[290] Malik, A., Khalid, F., Hidait, N., Anjum, K. M., Mahad, S., Razaq, A., Azmat, H. & Majeed, M. B. B. Arsenic toxicity in fish: Sources and impacts. In Huq, S. M. I. (ed.) *Arsenic in the Environment - Sources, Impacts and Remedies*, chap. 5 (IntechOpen, London, 2023).

[291] Smedley, P. L., Edmunds, W. M. & Pelig-Ba, K. B. Mobility of arsenic in groundwater in the Obuasi gold-mining area of Ghana: some implications for human health. *Geological Society, London, Special Publications* **113**, 163–181 (1996).

[292] Merrington, G. *et al.* Water quality guidelines for chemicals: learning lessons to deliver meaningful environmental metrics. *Environmental Science and Pollution Research* **21**, 6–16 (2014).

[293] Mitchell, E. J. & Frisbie, S. H. A comprehensive survey and analysis of international drinking water regulations for inorganic chemicals with comparisons to the World Health Organization's drinking-water guidelines. *PLOS ONE* **18**, e0287937 (2023).

[294] Van Winckel, T. *et al.* Towards harmonization of water quality management: A comparison of chemical drinking water and surface water quality standards around the globe. *Journal of Environmental Management* **298**, 113447 (2021).

[295] Sun, X., Ding, T.-T., Wang, Z.-J., Huang, P. & Liu, S.-S. Optimized Derivation of Predicted No-Effect Concentrations (PNECs) for Eight Polycyclic Aromatic Hydrocarbons (PAHs) Using HC10 Based on Acute Toxicity Data. *Toxics* **11**, 563 (2023).

[296] Pandey, J. & Singh, R. Heavy metals in sediments of Ganga River: up- and downstream urban influences. *Applied Water Science* **7**, 1669–1678 (2017).

[297] Zhu, H., Tang, X., Gu, C., Chen, R., Liu, Y., Chu, H. & Zhang, Z. Assessment of human exposure to cadmium and its nephrotoxicity in the Chinese population. *Science of The Total Environment* **918**, 170488 (2024).

[298] Edwards, I. R. & Aronson, J. K. Adverse drug reactions: definitions, diagnosis, and management. *The Lancet* **356**, 1255–1259 (2000).

[299] Coleman, J. J. & Pontefract, S. K. Adverse drug reactions. *Clinical Medicine* **16**, 481–485 (2016).

[300] Xue, R., Liao, J., Shao, X., Han, K., Long, J., Shao, L., Ai, N. & Fan, X. Prediction of Adverse Drug Reactions by Combining Biomedical Tripartite Network and Graph Representation Model. *Chemical Research in Toxicology* **33**, 202–210 (2020).

[301] Galletti, C., Aguirre-Plans, J., Oliva, B. & Fernandez-Fuentes, N. Prediction of Adverse Drug Reaction Linked to Protein Targets Using Network-Based Information and Machine Learning. *Frontiers in Bioinformatics* **2**, 906644 (2022).

[302] Park, J., Lee, S., Kim, K., Jung, J. & Lee, D. Large-scale prediction of adverse drug reactions-related proteins with network embedding. *Bioinformatics* **39**, btac843 (2023).

[303] Ghosh, S., Pandey, S. K. & Roy, K. Predictive classification-based read-across for diverse functional vitiligo-linked chemical exposomes (ViCE): A new approach for the assessment of chemical safety for the vitiligo disease in humans. *Toxicology in Vitro* **104**, 106018 (2025).

[304] Bell, S. M., Angrish, M. M., Wood, C. E. & Edwards, S. W. Integrating Publicly Available Data to Generate Computationally Predicted Adverse Outcome Pathways for Fatty Liver. *Toxicological Sciences* **150**, 510–520 (2016).

[305] Fan, B. *et al.* Construction of an adverse outcome pathway framework based on integrated data to evaluate arsenic-induced non-alcoholic fatty liver disease. *Environment International* **183**, 108381 (2024).

[306] d'Aquin, M. On the role of knowledge graphs in AI-based scientific discovery. *Journal of Web Semantics* **84**, 100854 (2025).

[307] Zhang, Y. *et al.* A comprehensive large-scale biomedical knowledge graph for AI-powered data-driven biomedical research. *Nature Machine Intelligence* **7**, 602–614 (2025).

[308] Xie, J., Liu, W., Hu, W., Ouyang, M. & Huang, T. Graph Neural Network-Based Toxicity Prediction by Integrating Molecular Fingerprints and Knowledge Graph Features. *Toxics* **13**, 953 (2025).

[309] Alava, J. J., Cisneros-Montemayor, A. M., Sumaila, U. R. & Cheung, W. W. L. Projected amplification of food web bioaccumulation of MeHg and PCBs under climate change in the Northeastern Pacific. *Scientific Reports* **8**, 13460 (2018).

[310] Windsor, F. M., Pereira, M. G., Morrissey, C. A., Tyler, C. R. & Ormerod, S. J. Environment and food web structure interact to alter the trophic magnification of persistent chemicals across river ecosystems. *Science of The Total Environment* **717**, 137271 (2020).