

README

Dataset title: Dataset connected to the project "*Cross-species bioanalytical assessment of the oxidative stress and xenobiotic metabolism modes of action for environmental water samples – a multi-layered approach to testing and modelling*"

Recommended citation: Lungu-Mitea, S., Mandava, G., Horáčková, J., Golovko, O., Toušová, Z., Frieberg, K., Alsayfi, S., Örn, S., Bednář, D., Ahrens, L., Hilscherová, K., & Lundqvist, J. (2026). Dataset connected to the project "*Cross-species bioanalytical assessment of the oxidative stress and xenobiotic metabolism modes of action for environmental water samples – a multi-layered approach to testing and modelling*" (Version 1) [Data set]. Swedish University of Agricultural Sciences. <https://doi.org/10.5878/zwnq-qb88>

See the Researchdata.se catalogue entry for this dataset for complete metadata, including the dataset description, recommended citation, contact information, and details on how to obtain the dataset.

Creator: Sebastian Lungu-Mitea

Responsible organisation: Swedish University of Agricultural Sciences

Description: The dataset comprises raw data, metadata, additional analyses, and a supplementary manuscript that links to three publications: the main article and two *Data in Brief* co-submissions.

Main article: *Cross-species bioanalytical assessment of the oxidative stress and xenobiotic metabolism modes of action for environmental water samples – a multi-layered approach to testing and modelling*

Sebastian Lungu-Mitea (a,b,c,d), Geeta Mandava (a), Jana Horackova (f), Oksana Golovko (e), Zuzana Tousova (b), David Bednar (f,g), Lutz Ahrens (e), Klara Hilscherova (b), Johan Lundqvist (a)

Data in Brief: *A dataset on the multi-species and multi-endpoint bioanalytical assessment of environmental samples derived from Swedish wastewater treatment plants*

Sebastian Lungu-Mitea (a,b,c,d), Geeta Mandava (a), Oksana Golovko (e), Kim Frieberg (a), Sara Alsayfi (a), Stefan Örn (a), Lutz Ahrens (e), Johan Lundqvist (a)

Data in Brief: *Human and zebrafish cross-species comparison on ligand affinity towards the aryl hydrocarbon receptor and Keap1 oxidative stress sensor: a molecular docking dataset*

Sebastian Lungu-Mitea (a,b,c,d), Jana Horackova (f), David Bednar (f,g), Klara Hilscherova (b)

Affiliations

(a) Department of Animal Biosciences, Molecular Toxicology, Swedish University of Agricultural Sciences (SLU), SE-75007 Uppsala, Sweden

(b) RECETOX, Faculty of Science, Masaryk University (MUNI), Brno, Czech Republic

(c) Department of Pharmaceutical Biosciences, Drug Safety and Toxicology, Uppsala University (UU), SE-75124 Uppsala, Sweden

(d) Department of Biology and Environmental Science, Linnaeus University (LNU), SE-39182 Kalmar, Sweden

(e) Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences (SLU), SE-75007 Uppsala, Sweden

(f) Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University (MUNI), Brno, Czech Republic

(g) International Clinical Research Centre, St. Anne's University Hospital Brno, Brno, Czech Republic

License: CC0 (Creative Commons Public Domain Dedication)

Access restrictions: N/A

Essential information for dataset reuse

Summary: The project comprises a main article, two *Data in Brief* co-submissions, and a supplementary manuscript (SM). The two *Data in Brief* co-submissions (“DIB_A”: bioanalytics and “DIB_B”: molecular docking) present the raw data and first-tier analyses in a FAIR principles manner. According to the *Data in Brief* guidelines, the SM serves as a connector between all three manuscripts. Moreover, the SM contains more detailed material & methods sections, additional results, and discussion. All contributions are also linked via the shared data repository (at hand), which comprises all raw data and metadata (Supplementary Information (SI) 1-16).

The first *Data in Brief* co-submission (DIB_A) encompasses raw data and preliminary analysis (LOEC/NOEC) pertinent to the bioanalytical assessment of the investigated environmental samples. This includes additional samples beyond the scope of the main article and employs further test frameworks, specifically the fish embryo acute toxicity test (FET, OECD TG 236). The data in DIB_A are associated with SI files ranging from SI1 to SI8.

The second *Data in Brief* co-submission (DIB_B) incorporates comprehensive molecular docking models along with protein sequence alignment studies. It also features confirmatory analyses that compare AlphaFold-predicted structures with those derived from cryo-EM and X-ray crystallography. DIB_B addresses the consequential implications and is associated with SI files SI9 through SI11.

All raw and computed metadata across the manuscripts are consolidated in this shared dataset, which includes 16 SI files and the SM. These files are numbered chronologically. SIs beyond SI12 contain data explicitly related to the main article (see table below). Every SI file has a dedicated index tab to help readers orient themselves. Legends explaining abbreviations are either provided on the index tab for the entire SI file or within the lower hierarchy tabs of the raw and metadata listings.

Subject: Earth and Environmental Sciences; Interdisciplinary: Environmental Chemistry; Health, Toxicology, and Mutagenesis; Water Science and Technology; Computational Biology

Specific subject area: Environmental Toxicology, Bioanalytical assessment of environmental samples, Molecular docking

Type of data: Table, Chart, Figure, Image, Graph, Molecular docking illustrations, Gene and Protein sequence files, Text

Data format: Raw – within supplementary files; Analysed – within the articles and supplementary files; Illustrative – within the articles and supplementary files

Data source location: Swedish University of Agricultural Sciences (SLU), Uppsala, Uppland, Sweden.

Background and Aim: “New approach methods” (NAMs) are reaching the threshold of regulatory acceptance within retrospective toxicological risk assessment and monitoring. Within the field often termed as “effect-based methods (EBMs), especially in vitro embryo tests and cellular test models, such as reporter gene assays, are heralded given their potential for quick and economic high-throughput assessment of manifold environmental samples. Notably, most cellular models originate from mammals (predominantly humans) or bacteria. Thus, actual toxicological environmental protection goals, such as aquatic vertebrates and invertebrates, seem underrepresented.

This project and the data presented herein aim to investigate the representativeness of mammalian cellular reporter assays as surrogates for aquatic vertebrates in effect-based assessments of the toxicological burden of environmental water samples. Therefore, a thorough investigation was conducted, covering biological, chemical, and computational analyses.

Data collection: Samples from wastewater treatment plants (WWTPs; influent and effluent) and their recipients (upstream and downstream) were collected using solid-phase extraction (SPE-DEX, Horizon Technology). All samples were bioanalytically assessed using a modified zebrafish embryo test, which scored lethal and sublethal endpoints via stereo microscopy (Leica EZ4D) and time-lapse camera recording (Canon EOS 500D). Influent and effluent samples were also analysed using reporter gene assays across three species (zebrafish, human, and mouse) and two modes of action: oxidative stress and xenobiotic metabolism. Cellular viability was assessed via the MTS assay. All endpoints were recorded on a plate reader (Spark, Tecan).

Ligand 3D structures were downloaded from PubChem and optimised in Avogadro 1.2.0 and RESP ESP charge Derive (R.E.D.) Server Development 2.0. PDBQT working files were generated via MGLTools. Complete AlphaFold-predicted protein structures were acquired via UniProt. Tertiary structures of the PAS-B domain were predicted with AlphaFold2 using ColabFold. Partial crystal and cryo-EM structures were obtained from RCSB-PDB. AutoDock Vina was utilised for molecular docking. The docking predictions were analysed in PyMOL. For sequence alignment, FASTA files were retrieved from NCBI. Alignment was performed in Unipro Ugene using the ClustalW algorithm.

Description of data collection: Recipient samples were collected as grab samples. Influent and effluent samples as 1-7 day composites. Zebrafish wild-type strains were kept in a state-of-the-art aquatic facility. Embryo tests were conducted in accordance with international guidelines and published literature. Reporter lines were cultured in a state-of-the-art cell culture laboratory. Mode-of-action-specific reporter assays were conducted in accordance with international guidelines and published literature.

The molecular docking dataset was entirely generated *in silico*. All generated and utilised PDBQT and FASTA-type files are given in the supplementary information (SI10). The following data sources were utilised: PubChem (<https://pubchem.ncbi.nlm.nih.gov>); UniProt (<https://www.uniprot.org>); RCSB Protein Data Bank (<https://www.rcsb.org>); National Center for Biotechnology Information (NCBI) (<https://www.ncbi.nlm.nih.gov/>). All acquired crystal and cryo-EM partial structures are mentioned in DIB_B, section 4, and the SM.

Location: Water sampling was conducted at three southern and western Swedish WWTPs (precise location disclosed due to confidentiality) and their respective catchments. All WWTPs are located in close proximity to waterbodies of primary anthropogenic and ecological concern. Chemical analyses and bioanalytics were conducted at SLU, Uppsala, Sweden. Molecular docking modelling and in silico-mediated effect-directed analysis were conducted at RECETOX, Masaryk University, Brno, Czech Republic.

Time period: Chemical analyses: 2018-2019. Bioanalytics: 2019-2022. Computational analyses: 2022-2024

Roles: Project coordination and responsibility: SLM. Chemical analyses and evaluations: OG and LA. Bioanalytics: SLM, GM, KF, SA, SÖ. Computational analyses: JH, DB, SLM, ZT. Financing: SLM, JH, LA, KH. Supervision: SLM, JH, LA, KH.

Dataset organisation and description

All files in the dataset for this project are identified, listed, and briefly described in the table below. The respective files are cross-referenced within all three articles and the SM.

Description of supplementary information (SI) uploaded to the data repository.			
Name	Content	Associated Article	File type
SI1	Bioanalytical data: raw and analysed data of all experiments conducted with the ZF4 assay	DIB_A	xlsx
SI2	Bioanalytical data: raw and analysed data of all experiments conducted with the MCF7/AREc32 assay	DIB_A	xlsx
SI3	Bioanalytical data: raw and analysed data of all experiments conducted with the ZFL assay	DIB_A	xlsx
SI4	Bioanalytical data: raw and analysed data of all experiments conducted with the HepG2 assay	DIB_A	xlsx
SI5	Bioanalytical data: raw and analysed data of all experiments conducted with the DREcoScreen assay	DIB_A	xlsx
SI6	Bioanalytical data: raw, analysed data, and statistical analyses of the mFET assay	DIB_A	xlsx
SI7	Analysis & statistics: transformation and statistical analyses conducted in all cellular bioassays	DIB_A	xlsx
SI8	Analysis & statistics: transformation and statistical analyses conducted for all pairwise comparisons between assays/species	DIB_A	xlsx
SI9	Molecular docking: summarised binding energies	DIB_B	xlsx
SI10	Molecular docking: PDBQT and FASTA-type files of ligands and receptors	DIB_B	A zip folder containing PDB and FASTA files
SI11	Analysis & statistics: TCDD-energy-equivalent calculations	DIB_B	xlsx

SI12	Analysis & statistics: Physico-chemical properties of all targeted compounds detected in chemical analysis; measured chemical concentration of targeted compounds per environmental sample; calculated REP values, extracted ToxCast data; OCHEM AhR QSAR outputs; bioanalytics for AhR-QSAR positive compounds; BEQ-Chem calculations.	Main article SM	xlsx
SI13	Analysis & statistics: bioavailability (C _{free} concentrations) data generated via linear solvation energy relationships QSPRs from UFZ-LSER	Main article SM	xlsx
SI14	Bioanalytical data: Metazachlor reference compound standard curves, cytotoxicity data, and all raw data generated	Main article SM	xlsx
SI15	Bioanalytical data: raw and metadata from the bioanalytical assessment of ISEDA-identified chemical compounds	Main article SM	xlsx
SI16	Analysis & statistics: BEQ-Bio calculations relating to estimates derived via 3PNL regression analysis (main article) and all other applied data-fitting scenarios (supplementary manuscript)	Main article SM	xlsx
SM	Supplementary manuscript: containing material & methods in detail for the entire project, additional results, and discussion. The SM cross-links all three articles.	All	docx, PDF

Reference information

Abbreviations: All abbreviations used in this project are given in the SM, section 1, table S2.

Data sources: For molecular docking, PDBQT and FASTA-type files were acquired via the public domains of PubChem (<https://pubchem.ncbi.nlm.nih.gov>); UniProt (<https://www.uniprot.org>); RCSB Protein Data Bank (<https://www.rcsb.org>); and the National Center for Biotechnology Information (NCBI) (<https://www.ncbi.nlm.nih.gov/>).

Version control: N/A

Source code information: N/A

Data Quality: N/A