ZerOPM

Zero pollution of Persistent, Mobile substances

Grant Agreement No. 101036756

Deliverable 5.1

September 2024

A PM substance evaluation of as many substances as possible on the global chemical inventory, presented within a database and report

Work Package 5 – Substance Grouping

Deliverable Work Package Leader:
NGI

Revision: 0
Dissemination level: Public

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101036756.

Any dissemination of results must indicate that it reflects only the author's view and that the Agency is not responsible for any use that may be made of the information it contains.

The present document has not yet received final approval from the European Commission and may be subject to changes.

Note about contributors

Lead partner responsible for the deliverable:	NGI
Deliverable prepared by:	Sivani Baskaran, Raoul Wolf, Hans Peter Arp
Partner responsible for quality control:	EMPA, DVGW
Deliverable reviewed by:	Zhanyun Wang (EMPA), Sarah Hale (DVGW)
Other contributors:	University of Luxembourg (uni.lu) Emma Schymanski, Emma Palm

Project information

Project period:	October 2021 – September 2026
Duration (no. of months):	60
Website:	www.zeropm.eu
Project coordinator:	Norwegian Geotechnical Institute

Project partners:

Summary

The aim of Work Package 5 (WP5) Substance Grouping is to address the hypothesis that grouping persistent and mobile (PM) substances based on specific molecular substructures will support prioritization and directed risk assessment procedures of PM substance groups. This deliverable presents key steps towards addressing this hypothesis. Herein we developed a new Global Chemical Inventory (GCI) that contained as many high-quality chemical structure information as possible. In addition, the chosen architecture will allow for future additions to the database, such as a graphical user interface or additional information (e.g., use classes). The high-quality structures in the GCI were used for in silico assessments of persistence and mobility. The assessment is grounded in the classifications for persistent, mobile and toxic (PMT) and very persistent, very mobile (vPvM) substances in the recently amended Chemicals, Labelling and Packaging (CLP) (EC 1272/2008). For the mobility assessments, experimental log K_{OC} values (if available) along with 6 different experimental different prediction techniques were used to obtain a distribution of likely K_{OC} values. For each model, a weakly informed Bayesian analysis, weighted with information about the applicability domain of the model, was used to derive a probability distribution of K_{OC} values. For persistence, a different approach was taken due the diversity of tests and models used to assess persistence, including readily biodegradability tests (indicating non persistence) or simulated half-lives in different environmental media. From this, persistence probabilities from the estimated half-life data were combined with the experimental and regulatory data available (including readily biodegradability tests) using a Bayesian Modelling Average approach. The current version of the persistent and mobility assessments is being made available on Zenodo via this link, and will be ultimately available as part of the searchable ZeroPM database (https://database.zeropm.eu/). Further, to consider the persistent and mobile transformation products of substances on the GCI, an approach to identify dead-end transformation products, and whether these dead-end transformation products should be considered persistent and mobility is also presented. This data set can be used to develop "bottom up" or "top down" strategies towards prioritizing PM substance groups. The "bottom up" approach can be done with data approaches by looking for common substance groups amongst the PMs substances vs non-PM substances. The "top down" considers hypothesized PM substance groups based on stakeholder concerns and expert knowledge, and uses the data to assess if such groups are associated with PM substances identified on the GCI. Future work in this deliverable will be towards utilizing the data generated in this deliverable towards "top down" and "bottom up" approaches for identifying potential priority PM substance groups.

Contents

Contents 5

List of Tables 6

List of Figures 6

1 Introduction 7

2 Global Chemical Inventory 8

3 PM evaluations 9

3.1 M Assessments 10
3.2 P assessments 15

4 Transformation Products 17

5 Next Steps 18

6 Acknowledgements 19

7 References 20

8 Appendix 1 25

List of Tables

Table 1. Summary of PM criteria as reported in the EU legislation[17]. 10
Table 2. Upcoming Deliverables in WP5 19

List of Figures

Figure 1. Workflow of the tasks in WP5 of ZeroPM to prioritize PM substance groups for further action for regulatory authorities and towards safe and sustainable by design. 7
Figure 2. A) Summary of the workflow used for KOC value estimations. The process beings with 199 000 SMILES and ends with probability distributions of the KOC value for each chemical and for each model. B) Experimental, regulatory and estimated data are combined into a single new probability distribution for each chemical. 12
Figure 3. Distribution of chemicals in the GCI if the minimum (left), mean (middle), maximum (right) of the KOC distribution is used to determine the log10 KOC value in a mobility assessment. 14
Figure 4. Percentage of substances considered very mobile (vM, red), mobile (M, yellow), and not mobile (not M, green) depending on which endpoint of the K_{OC} distribution is used for the assessment. 14
Figure 5. A) Summary of the workflow used deriving probabilities of persistency. The process beings with 199 000 SMILES. Ready biodegradability (RBD) data is used to determine the probability of a substance not being persistent while log half-life data (days, DT50) is used to predict the probability of not P, P, or vP. B) Experimental, regulatory and estimated data are combined using a Bayesian Modelling Average (BMA) approach to obtain the probability of not P, P, and vP for each chemical. 16
Figure 6. Distribution of the probabilities that chemicals in the data set are persistent (including P or vP) on the left and the distribution of the probabilities that chemicals in the data set are not persistent (right). 16
Figure 8 Distribution of XlogP values for parent compounds (blue) and their corresponding TPs (red), demonstrating that TPs are generally more polar (and thus more mobile). Density refers to kernel density. Taken from Figure 4b in Palm et al. 2023[48] 18
Figure 2 Screenshot of the graphical user interface (GUI) implemented for the global chemical inventory. The screenshot was taken on 24th August 2024. 25

1 Introduction

ZeroPM, which stands for Zero pollution of Persistent, Mobile substances, is a 5-year European research project funded under the Horizon 2020 research and innovation program. ZeroPM will interlink and synergize three strategies to protect the environment and human health from persistent and mobile (PM) substances: Prevent, Prioritize and Remove. To do this, ZeroPM will develop an evidence-based multilevel framework. The framework will guide policy, technological and market incentives to minimize use, emissions and pollution of entire groups of PM substances.

The aim of Work Package 5 (WP5) Substance Grouping is to Substance Grouping is to address the central hypothesis that grouping persistent and mobile substances based on specific molecular substructures will support persistent and mobile substance prioritization and directed risk assessment procedures.

Grouping based on substructure is not only convenient to link different PM substances together, but also to group transformation products that are PM substances. For instance, melamine (a triazine) has many transformation products that are also triazines[1]. This grouping approach also discourages replacing PM substances with another PM substance with a common substructure as a potential case of problematic "regrettable substitution" . Through 7 tasks, WP5 will support "safe and sustainable by design" approaches, and an "early warning system" for PM substances to be used by regulators, industry, scientists and water resource managers (Figure 1).

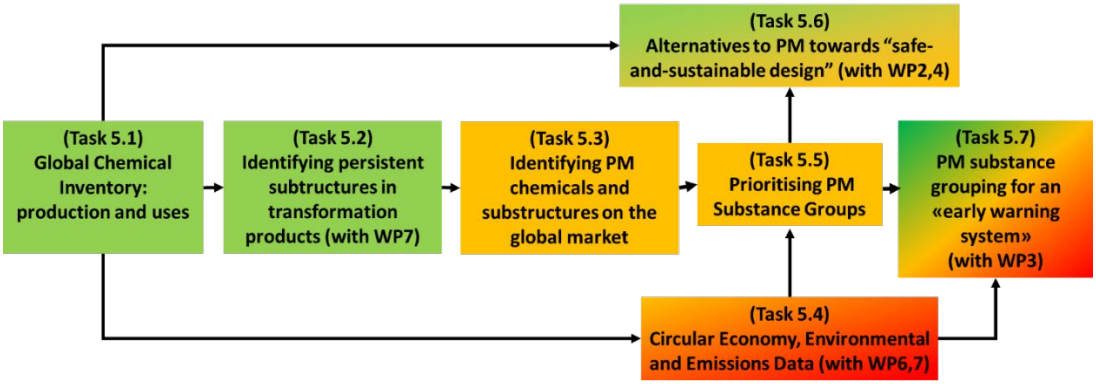

Figure 1. Workflow of the tasks in WP5 of ZeroPM to prioritize PM substance groups for further action for regulatory authorities and towards safe and sustainable by design.

This deliverable, presenting a PM substance evaluation of as many substances as possible on the global chemical inventory, combines work in Tasks 5.1, 5.2 and 5.3. From Task 5.1, an updated "Global Chemical Inventory" (GCI) was established (Section 2). From Task 5.3, an approach for assessing persistence and mobility for as many chemicals on the GCI was developed (Section 3). Finally, to introduce the work towards prioritizing PM substance groups in later tasks, the approach to identifying persistent substructures from transformation reactions from Task 5.2 is presented (Section 4).

2 Global Chemical Inventory

An overview of the landscape of chemical substances that were registered on the global market was previously published by Wang *et al.* in 2020.[2] Their work provided an important document into the diversity and challenges of chemical data on an international scale. In ZeroPM, this pioneering effort has been (1) re-evaluated with updated data sources, (2) using a new methodology to identify/represent the structures of chemicals, (3) combined with a workflow to treat multi-component structures, and (4) showcased in a modern database infrastructure. The new GCI is to provide as much high-quality chemical structure information as possible for *in silico* assessments of persistence and mobility in later tasks. In addition, the chosen architecture will allow for future additions to the database, such as a graphical user interface or additional information (e.g., use classes).

A common basis for the work done within WP5 of ZeroPM is the adherence to the FAIR principles (*F*indable, *A*ccessible, *I*nteroperable, and *R*eproducible). This global chemical inventory has been constructed in a code-first approach, and all the codes produced under this task have been or will be made available on the project's GitHub repository.[3] The associated data sets and dissemination documents have been or will be made available on the project's Zenodo community.[4] In order to follow the FAIR principles from a chemoinformatic perspective, the InChI (*I*nternational *Ch*emical *I*dentifier) was chosen as the line-readable "gold standard" for chemical structures.[5] It offers greater flexibility and ease of use when compared to more complex notations, such as MOL or SDF representations.

For these tasks, multiple functionalities were coded in the programming language R[6] and wrapped as dedicated packages hosted on ZeroPM's GitHub repository.[7] The first step was to re-evaluate all information that was previously used, *i.e.*, re-downloading all previously used data sets, if still possible. Out of the 22 data sources reported by Wang *et al.*[2], 16 could be verified; six data sources could not be verified and may have been removed from public access by the corresponding authorities altogether. Four of the re-evaluated data sources were updated with new entries. For this work, the most up-to-date versions of the data sources were used at the time of assembly; unverifiable data sources were not considered.

A practical challenge of the re-assembly was the diversity in data formats. The file formats included .csv, .xls(x), .pdf, .doc(x), but also non-automatically extractable information from websites that had to be manually extracted. Some individual data sets were only available as scanned pictures, for which an optical character recognition algorithm was applied to extract as much of the information as possible. Before further processing, all extracted data was subject to an extensive cleanup process, in which non-UTF8-compliant characters were removed or appropriately replaced, *e.g.*, transforming the Greek letter "α" to the word "alpha" and removing other non-Latin alphabet characters for full compatibility with the structure identification approach described below.

In principle, all data sources contained two distinct types of information about the chemicals: chemical names (including synonyms and abbreviations) and registry numbers, such as CAS registry numbers (CASRNs). An important quality concern within the re-evaluation of the global chemical inventory has been the discrepancies between CASRNs and the provided chemical structures. This was addressed in a multi-step approach: both the CASRN and the chemical name(s) of an entry were used as input in the following, fully automated API retrieval services: PubChem[8] (both compound and substance domain), CAS Common Chemistry[9,10], CompTox[11], and CACTVS[12]. These services return a ranked list of candidate structures, based on their respective internal determination mechanism. This ranking information was used to assign every candidate structure a weighted average score based on the rank product theory, *i.e.*, a consensus rank[13]. For each inventory entry, the structure with the highest consensus rank was chosen as the most likely chemical structure associated with the respective entry.

In total, 16 data sources with a total of 25 files were analysed, covering a total of 38 countries. This resulted in 543,851 unique inventory entries containing 1,719,361 unique search terms, for which the API queries provided 497,532 results, including 382,542 chemical structures. After the consensus ranking approach, 144,351 chemical structures were identified as the most probable structures for their respective inventory entries (a 3% increase compared to the work by Wang *et al.*[2]), including 42,901 multicomponent entries, *e.g.*, salts. The 144,351 unique structures were assigned unique ZeroPM IDs for the subsequent work on the assessment of persistence and mobility.

3 PM evaluations

Chemical structure information such as InChIs were taken from the GCI. As most models use SMILES notation for predictions, some structural data conversions were necessary and completed using OpenBabel[14] and ChemineR[15]. During this process, multi-constituent chemical substances were split into mono-constituent compounds, and stereo and isotopic information were removed. To ensure that the predictions were representative of the chemical structure commonly found in the environment, tautomers for all SMILES were calculated using OpenBabel and a random sample of up to five tautomers were selected for further analysis. Chemicals that were larger than 500 g/mol were removed for this analysis as their size is likely to reduce their mobility and most quantitative-structure-activity-relationship (QSAR) models are not generally suitable for larger substances. This left approximately 119,000 SMILES used for estimations for approximately 93,000 substances.

The criteria for classification of PM substances was based on Section 4.4 of the Commission Delegated Regulation (EU) No. 2023/707 which amended the EC Regulation No. 1272/2008[16]. A summary of the criteria is provided in Table 1.

Table 1. Summary of PM criteria as reported in the EU legislation[17]. DT_{50} = degradation half-life; K_{OC} = organic-carbon—water partition ratio

Hazard Classification	Relevant Phase or Conditions	Threshold
Persistent (P)	Marine water	$DT_{50} > 60$ days
	Fresh or estuarine water	$DT_{50} > 40$ days
	Marine sediment	$DT_{50} > 180$ days
	Fresh or estuarine water sediment	$DT_{50} > 120$ days
	Soil	$DT_{50} > 120$ days
Very Persistent (vP)	Marine, fresh or estuarine water	$DT_{50} > 60$ days
	Marine, fresh or estuarine water sediment	$DT_{50} > 180$ days
	Soil	$DT_{50} > 180$ days
Mobile	For ionisable substances, use the lowest K_{OC} for pH 4-9	$\log_{10} K_{OC} < 3$
Very Mobile	For ionisable substances, use the lowest K_{OC} for pH 4-9	$\log_{10} K_{OC} < 2$

In the regulatory context, Persistence and Mobility are considered intrinsic substance properties[17]. Persistence is considered intrinsic based on the on the intrinsic recalcitrance of the molecule; however, it is also acknowledged in the CLP and UN-GHS[18] that how persistence is manifested depends on actual conditions in the receiving environmental compartment (e.g. redox potential, pH, presence of suitable micro-organisms, concentrations of the substances and occurrences and concentration of other substances). Similarly, mobility is also an intrinsic property driven by the intrinsic factors related to the molecular structures and interactions (such as molecular size, van der Waals interactions of the molecular surface, ionic interactions), but how this is manifested depends on the actual conditions of the receiving environment.

The challenge when working with degradation half-lives for persistence (DT_{50}) and organic-carbon—water partition coefficients (K_{OC}) for mobility is that experimental data are dependent on both intrinsic substance properties and environmentally dependent variables. Lab and field measurements of both types of parameters are dependent on different parameters that cannot be kept constant, or easily characterised (such as the soil composition, or the microbial activity and species in an environment). The physical-chemical properties of chemicals can also change with temperature or the pH of the system. However, in most cases, relevant meta data about the field or laboratory measurements are not provided consistently, nor are there vast amounts of data from different systems available. Thus, results from different labs, dossiers or papers cannot be easily compared or combined. We have therefore taken a broader approach using a Bayesian analysis to predict the likely distribution of the K_{OC} value for each chemical to use in the mobility (M) assessments, and a Bayesian approach to predicting the probability of a chemical being persistent for the persistence (P) assessments.

3.1 M Assessments

Experimental data were extracted from known datasets such as OChemData and from the training data of specific existing QSAR models for predicting the K_{OC}[19–30]. Different models for the same parameter are expected to have similar or overlapping training data sets, and therefore, all values were rounded to two decimal places and duplicate values were removed. Due to the size and scope of the chemicals being assessed, manual

literature searches for additional data were not possible. Data reported in the REACH registries were obtained using eChemPortal[31] and QSAR Toolbox[32]; we define this set of data as the regulatory data. We keep this dataset separate from the other experimental data, as meta information for the regulatory data is limited and includes experimental or estimated values of varying quality. The reported data were processed to standardize units, identify censored data, and remove data related to mixtures and UVCBs (Substances of Unknown or Variable Composition, Complex Reaction Products, or Biological Materials). The regulatory data are also expected to overlap with the reported experimental values, so those values that were identical up to two decimal places were considered duplicates and removed.

Six different prediction techniques were used to obtain estimated K_{OC} values. For each model, we ran a weakly informed Bayesian analysis which includes the predicted K_{OC} value of each model and, if available, the reported error for the prediction. Information about the applicability domain of the model was considered when weighting the model result in the analysis. Each model can result in a single K_{OC} value prediction, or up to three predictions obtained using different "settings" or parameterizations of the same model. For example, there are 6 different poly-parameter linear free energy relationship (ppLFER) equations for predicting K_{OC}. Additionally, two of the models, multi-linear regression models by Trapp and Franco[33] and ACD/Labs[34] can predict the K_{OC} at different pH conditions. For each prediction method, the predicted values, reported errors and weights are used in the Bayesian analysis approach to get a probability distribution for the predicted K_{OC} from each model for each substance. Figure 2 shows a summary of the workflow for estimated K_{OC}, with each model described as follows.

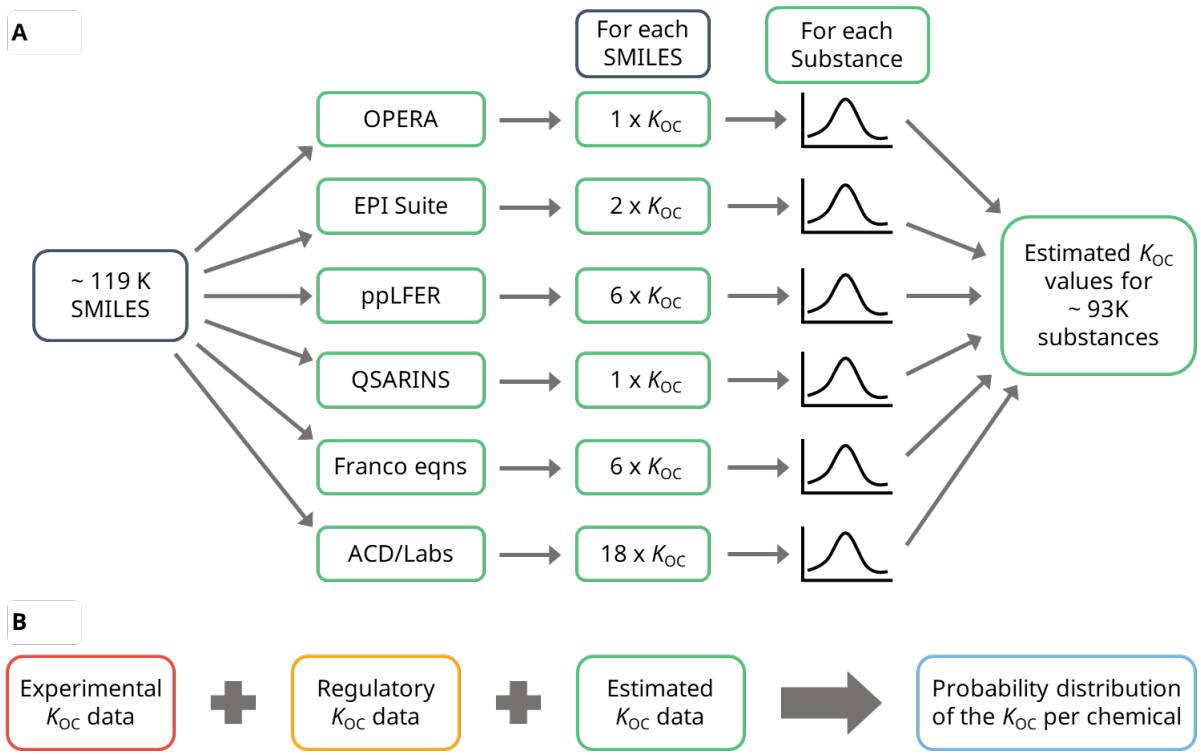

Figure 2. A) Summary of the workflow used for K_{OC} value estimations. The process beings with 119 000 SMILES and ends with probability distributions of the K_{OC} value for each chemical and for each model. B) Experimental, regulatory and estimated data are combined into a single new probability distribution for each chemical.

The first model was OPERA,[29] which predicted the K_{OC} of a chemical and provided information about the applicability domain (AD) of the chemical given the training data and the prediction error. OPERA uses a *k*-nearest neighbour model to assess the applicability domain, which uses the PaDEL descriptors to assess the similarity of a chemical to the other chemicals in its training set. Although OPERA's process should standardize the input SMILES, we found that the same chemical with different tautomer SMILES were standardized differently and resulted in different predicted values.

The second model was KOCWIN from EPI Suite™, which includes two models for predicting K_{OC}, one based on the octanol-water partition coefficients (K_{OW}), and another based on the molecular connectivity index (MCI). Both models are trained on the same data set, of a limited K_{OC} range, while predictions from the models could vary by orders of magnitude. This was likely because of the way fragments and functional groups are defined and parameterized by the MCI-based model. Identifying the functional groups and/or fragments that are not parameterized by the model is outside the scope of such an assessment, and so the range of the molecular weight and K_{OC} values in the training dataset is used to classify chemicals as either inside or outside the AD. For quaternary ammonium compounds and some ionizable chemicals, KOCWIN provided predictions but flagged the chemical as being outside the scope of the model in the notes or warning section of the full output. This information was included in our weighting of the data and as part of the AD assessment.

The third estimation tool was ppLFERs, which are linear regression equations that use the Abraham solute descriptors to quantify specific solute-solvent interactions[35]. The solute descriptors were calculated using IFSQSAR[36,37]. There are multiple ppLFERs for predicting K_{OC}; the data used to train the models are expected to overlap because of the limited number of well-reported K_{OC} values, while some models have been parameterized for specific environmental conditions or soil types. For example, Endo *et al.*[23] reports three ppLFER models that have been parameterized to lignite and 2 different concentrations of peat. We used the different ppLFER equations reported by Bronner and Goss [26], Nguyen *et al.*[21], Poole and Poole[38], and Endo *et al.*[23]. Bronner and Goss[26] and Ngyuen *et al.*[21] both reported multiple ppLFER equations and the equation that was considered most reliable or produced the lowest propagated error were selected. Error propagation using the reported errors for the solvent descriptors and for the solute descriptors was used to calculate the error on the K_{OC} prediction. Chemicals were assessed to be inside or outside the applicability domain of the specific ppLFER equation by calculating the leverage based on the solute descriptors of the training data and of the chemicals being assessed[39]. Due to the nature and capacity of our Bayesian analysis method, we obtain the K_{OC} using 6 different ppLFER equations.

The fourth K_{OC} model was QSARINS-Chem, which also uses linear regression equations, this time with the PaDEL descriptors to predict the $\log_{10} K_{OC}$[40]. The PaDEL descriptors were calculated as part of the OPERA model and were used in the calculation of K_{OC} using the equation reported in an earlier version of QSARINS-chem[41]. The prediction error and applicability domain of the model was assessed using the same leverage-based approach described above for ppLFERs.

The final two models are more suitable for ionizable substances, the equations parameterized by Franco and Trapp[33] and the ACD/Labs software[34]. Such models to consider ionizable substances is important, as a large number of previously identified persistent and mobile substances were ionizable.[42] The ACD/Labs software is the only commercial software used. Franco and Trapp[33] provide three equations for predicting pH-specific K_{OC} values using the pKa and $\log_{10} K_{OW}$. The equation used for each chemical was determined by the nature of a chemical (an acid, base, or amphoter) and whether it was within the applicability domain of the pK_a thresholds reported in the paper for categorizing chemicals. For example, amphoters where the difference between the pK_a and pK_b was less than two, were specifically reported to be outside the domain of the model[33]. pK_a, pK_b and log K_{OW} for this prediction were obtained from the OPERA model. As the parameterization of the OPERA pK_a/pK_b and K_{OW} model are different to the parameterization of the OPERA K_{OC} model, the results of both models are not based on the same data. Error propagation and a leverage-based approach for the chemical similarity was done using the method described above. As both Franco and Trapp equations and the ACD/Labs predictions are dependent on pK_a, estimates were calculated at six different pHs between 4 and 9.

The probability distribution of the estimated K_{OC} for each chemical from each model is combined with the regulatory and experimental data using the Bayesian model. During the analysis, the data were weighted based on the reliability of the source, where the most reliable values came from experimental data and the least reliable was considered

to be the distribution generated using EPISuite data. The analysis is completed on a per chemical basis, and we are left with a probability distribution for the K_{OC} that is based on all available information.

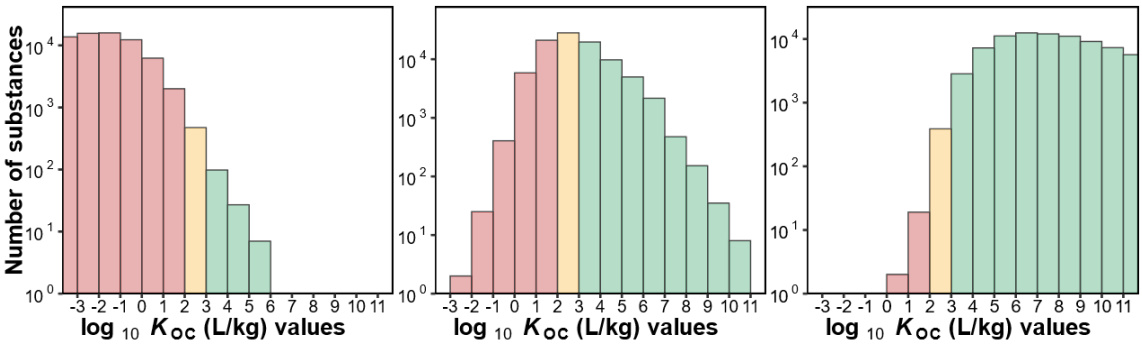

Figure 3. Distribution of chemicals in the GCI if the minimum (left), mean (middle), maximum (right) of the K_{OC} distribution is used to determine the $\log_{10} K_{OC}$ value in a mobility assessment. Redi indicates if the vM criteria is met, yellow the M criteria, and green for "not mobility"

As we are working with probability distributions, there are multiple aggregate endpoints that can be used to assess the mobility, such as the mean or average value of the distribution, the median or midpoint, minimum, maximum, or even a percentile of the distribution. In Figure 3, we show the distribution of chemicals within the GCI along a $\log_{10} K_{OC}$ range between -4 and 12 using three different aggregate endpoints. Figure 4 further shows how the results of the M assessment change with the selection of different distribution endpoints. Of those displayed, the endpoint that uses the central tendency (*e.g.*, the mean or median) is the best for current M assessments. To be more risk adverse during the M assessment, the 50th percentile or lower of the distribution could be used; however, further justification of using a particular percentile in such an approach would be necessary. For example, it may be reasonable, for instance, to use a lower percentile value (*e.g.*, 10th or 25th percentile) to classify ionizable chemicals that are likely to be influenced by pH, to err on the side of precaution, as the CLP criteria refer to the minimum $\log_{10} K_{OC}$ within a pH range of 4-9.

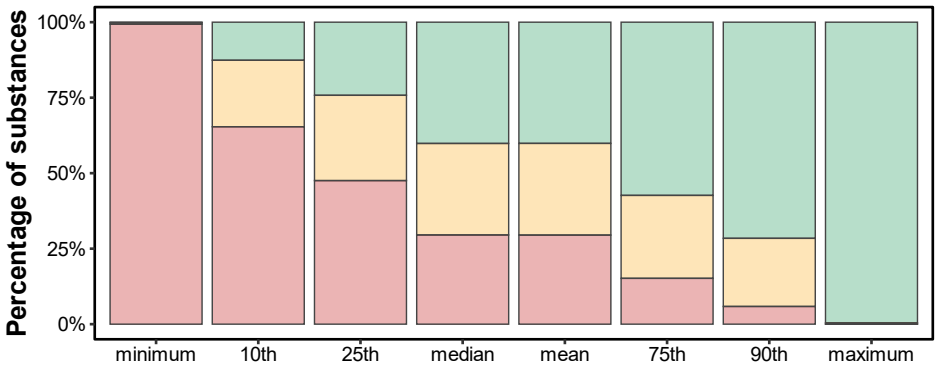

Figure 4. Percentage of substances considered very mobile (vM, red), mobile (M, yellow), and not mobile (not M, green) depending on which endpoint of the K_{OC} distribution is used for the assessment.

3.2 P assessments

Due to the nature of the P assessments, a slightly different approach was undertaken. Data used for per P assessments are available in two forms, a half-life of a chemical in a system, or the result of a ready biodegradability test. The former is more easily compared to the thresholds reported in Table 1; however, the latter, a binary result that indicates if something is readily biodegradable, is also an acceptable method for identifying non-persistent chemicals according to the REACH Guidance documents[17].

Data availability and QSARs for predicting the half-life or ready biodegradability of a chemical are limited. Experimental data for ready biodegradability and half-life data were obtained from the OChem Database[30] and from the training data for OPERA's ready biodegradability model[29]. Regulatory data were extracted using the OECD QSAR Toolbox and eChemPortal[31,32]. Duplicates data points were removed in the same method used for experimental and regulatory K_{OC} data.

For persistence, data from EPI Suite's BIOWIN and OPERA's ready biodegradability and half-life models were used. From BIOWIN, the results of BIOWIN3 and BIOWIN5 model were combined using the recommended approach by EPI Suite™ [43] and used to identify which chemicals were readily biodegradable. Ready biodegradability data were then converted to a probability of the chemical being readily biodegradable using the Bayesian analysis. A similar calculation was done using the OPERA ready biodegradability model. Using the Bayesian analysis, the half-life data were also converted from a categorical value based on the persistence thresholds for fresh water (Table 1) to a probability that the chemical is not persistent (not P), persistent (P) or very persistent (vP).

The persistence probabilities from the estimated half-life data were then combined with the experimental and regulatory data available using a Bayesian Modelling Average approach. A summary of the analysis approach is depicted in Figure 5. However, there is a large limit to the availability of directly and reliably measured half-life data and subsequently thorough persistence assessments. By assigning a probability score to each chemical for being not P or P, we then more easily recognize the uncertainty associated with the unreliability of the data.

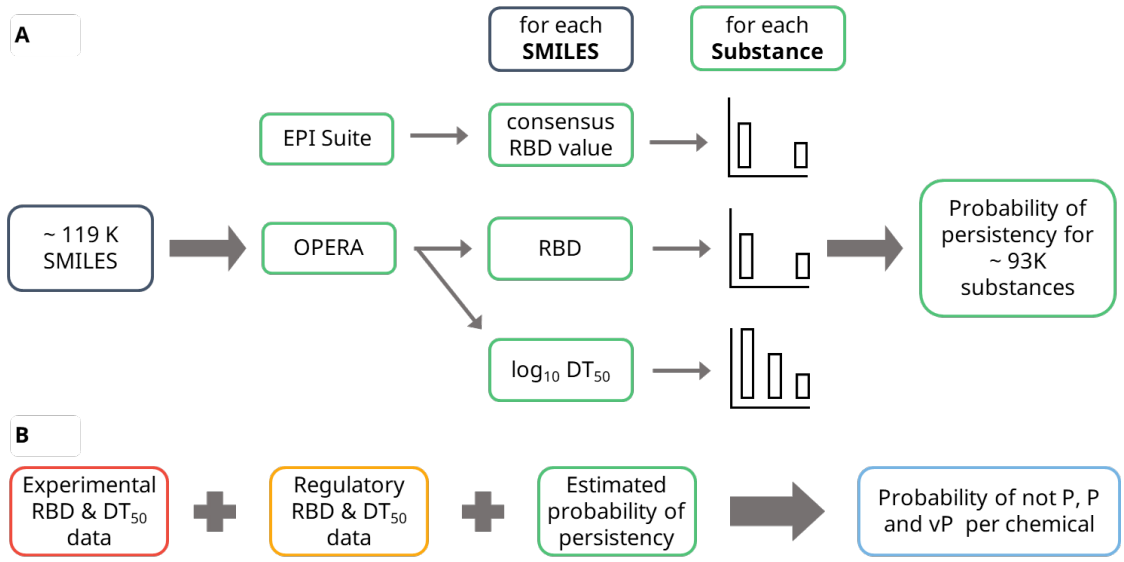

Figure 5. A) Summary of the workflow used to derive the probabilities of persistence. The process beings with 119 000 SMILES. Ready biodegradability (RBD) data are used to determine the probability of a substance not being persistent while log half-life data (days, DT50) is used to predict the probability of a chemical being not P, P, or vP. B) Experimental, regulatory and estimated data are combined using a Bayesian Modelling Average (BMA) approach to obtain the probability of being not P, P, and vP for each chemical.

In Figure 6, we show the distribution of the probabilities that a chemical is either persistent or not persistent. The probabilities that a given substance is in either category is rarely above 75%. This is due to two factors, the assumption that the probability of a substance being persistent and not persistent is equal to one, and within a Bayesian prior framework, there is equal probability of any chemical being P or not P before including any additional information from experimental, regulatory or estimated sources. The data shows a distinct lack of available information and the need to increase our confidence in persistence assessments with more information through experimental work that is openly and easily accessible and through improvement of QSAR models for half-life and ready biodegradability predictions.

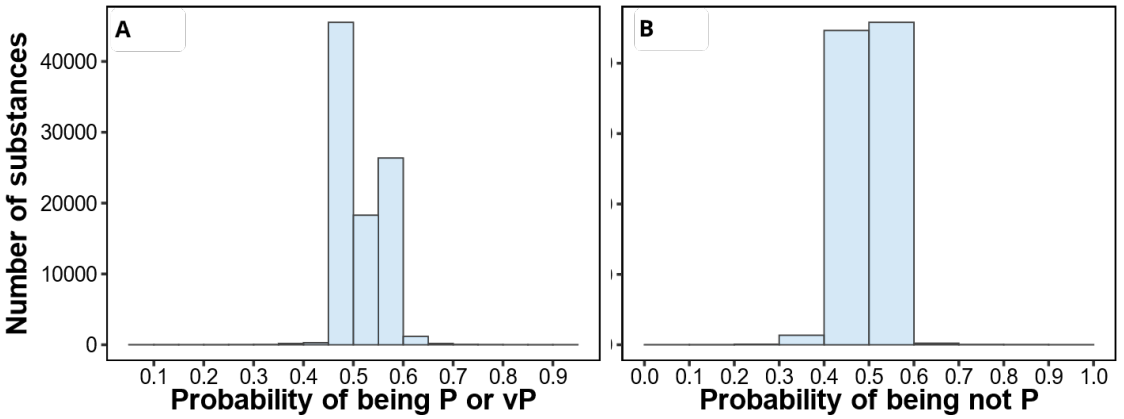

Figure 6. Distribution of the probabilities that chemicals in the data set are persistent (including P or vP) on the left and the distribution of the probabilities that chemicals in the data set are not persistent (right).

The resulting PM assessments from this deliverable will be the them of to two scientific papers in preperation. The current version of the persistent and mobility assessments is being made accessible on Zenodo via this link. Later, after this Deliverable is approved and further QA, such as via publication in the peer-reviewed literature, it will be also available in the searchable ZeroPM database (https://database.zeropm.eu/) (anticipated in early 2025).

4 Transformation Products

Transformation products (TPs) can form from chemicals entering the environment due to a combination of abiotic and biological processes. While TPs are often less harmful than their parent compounds, there are cases of TPs that are more harmful and/or persistent than their parents, causing concern due to the lack of attention often given to TPs in chemicals assessments[44]. However, there is also a scarcity of public data available on TPs, with open information readily available for only 1.3 % of the 140,000 chemicals in the Global Chemical Inventory (GCI) (excluding substances without exact structural information), according to our analysis. The following outlines the activities within Task 5.2 of WP5 to alleviate and improve this situation.

Firstly, the updateable collection of all open TP reactions in literature compiled in a collaboration between the Environmental Cheminformatics group (ECI) with PubChem[45] has been maintained throughout ZeroPM so far. This dataset includes the ChEMBL dataset (integrated within PubChem) and eight lists from the NORMAN Suspect List Exchange (NORMAN-SLE)[46]. The latest update (2 Apr. 2024) includes 7504 entries, corresponding to 6122 unique reactions related to 6806 unique compounds, including 2497 predecessors (parents) and 5119 successors (TPs)[47]. This dataset is integrated within PubChem and patRoon,[48] and has been supplied to the BioTransformer developers for use in the development of BioTransformer 4.0. Guidance has been given on creating Findable, Accessible, Interoperable and Reusable (FAIR) transformation information, including templates[49].

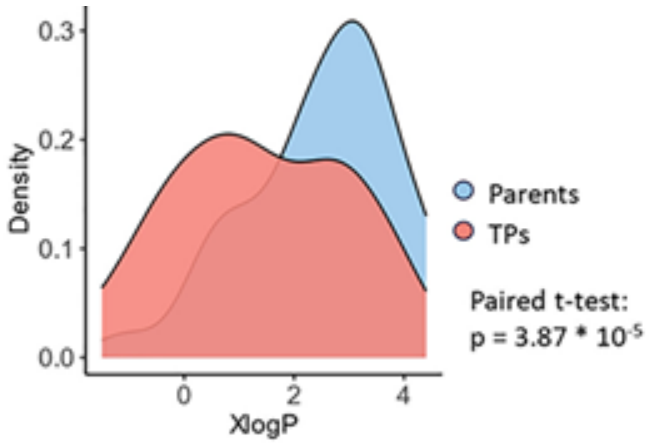

Figure 7 Distribution of XlogP values for parent compounds (blue) and their corresponding TPs (red), demonstrating that TPs are generally more polar (and thus more mobile). Density refers to kernel density. Taken from Figure 4b in Palm et al. 2023[50]

Finally, predictive systems such as BioTransformer will be used to cover the gap of TP information in the GCI. Although previous versions of BioTransformer had primarily human and microbial metabolism, with some environmental reactions integrated through enviPath[51], the current version (under development in a collaboration with the WishartLab) also has abiotic reactions, increasing the relevance of these predictions to ZeroPM. Testing with the latest BioTransformer 4.0 is underway, with promising results despite some issues with combinatorial explosion and over-predicted reaction pathways. This process is helping prioritize further reactions to extract and improve predictions, for use in the full GCI in the coming months. Preliminary calculations have already revealed several "dead-end" TPs of concern, which in turn will reveal substructural moieties for potential grouping approaches. Calculations in the recent grouping effort have revealed that huge numbers of chemicals (hundreds of thousands to millions) may be included in such moiety-based grouping approaches[52]; how to prioritize chemicals in this case will be a key discussion at the upcoming work in ZeroPM. The scoping of how to include this in the GCI, or as a separate resource, has already started and will continue over the coming months.

5 Next Steps

Within ZeroPM, we are developing methodologies for the prevention, prioritization and removal of PM substances and PM substance groups. Key steps that have been presented towards this in this deliverable are i) identifying, to the best of our ability, the actual structural information of substances on the GCI; ii) assessing their persistence and mobility using a diverse array of experimental data and models that were integrated through Bayesian analysis, and iii) developing methodologies to identify dead-end transformation products of substances on the GCI and determine whether they would be persistence and mobility. This works serves as the foundation to future prioritization of PM substance groups going forward, as will be presented in the subsequent deliverables (Table 2).

Table 2. Upcoming Deliverables in WP5

Deliverables	Description
D5.2 (August 2025)	Circular Economy, Environmental, and Emissions database for the global inventory
D5.3 (July 2026)	Prioritized PM Substance and Substance Group suspect list, and non-PM alternatives for action by policy makers and researchers
D5.4 (July 2026)	Policy oriented technical and practical guideline for an "early-warning system" for priority PM substances groups

As a starting point, the occurrence of common moieties/substructures in PM substances and dead-end transformation products (Task 5.2) will be used to identify a shortlist of PM substance groups (Task 5.3). These substance groups can be identified "bottom-up" by statistical comparison how such common substructures occur within the identified PM substances, to establish a new PM substance group. Further, they can be identified by "top-down" strategies, by testing if hypothesized PM substance groups are associated with dead-end transformation products or likelihood of being PM substance in the GCI. Based on the size of the data, both "bottom up" and "top down" approaches appear to lead to extremely large substance groups without further filters for prioritization,

Within future work to refine such PM substance groups from "bottom up" or "top down" approaches data related to the production, emissions, presence and exposure of PM substances will be considered, At events such as the ZeroPM prioritization workshop (https://zeropm.eu/prioritization-workshop/), stakeholders to present aspects they were concerned about related to hard-to-treat drinking water contamination, The stakeholders indicated that production volumes, exposure, lack of water treatment infrastructure, and hazards are important considerations to prioritize PM substances groups for different purposes, particularly considering the need for prioritization in a context of limited resources. Key points towards prioritization are that substances clearly fall within the regulatory PMT/vPvM hazard criteria in the CLP, and that there is immediate concern about that substance group due to high exposure (e.g. via monitoring data, production data) or additional hazards (e.g. toxicological data).

The output of this deliverable is presented on Zenodo via this link. After approval of this deliverable and peer-review via scientific manuscripts, it will also be available on the searchable ZeroPM database (https://database.zeropm.eu/). Publication of this work on the ZeroPM database will allow users to identify which PM substances have been registered on the GCI and within a specific country or region. This database, building on the FAIR principles, will enable the broader regulatory and research community to develop their own approaches to prioritize PM substance groups in their specific country

6 Acknowledgements

Thanks to Dr. Nezam Bozorgzadeh at NGI for feedback on the Bayesian Analysis approach and facilitating access to computational resources.

7 References

(1) Zheng, Z.; Arp, H. P. H.; Peters, G.; Andersson, P. L. Combining In Silico Tools with Multicriteria Analysis for Alternatives Assessment of Hazardous Chemicals: Accounting for the Transformation Products of decaBDE and Its Alternatives. *Environ. Sci. Technol.* **2021**, *55* (2), 1088–1098. https://doi.org/10.1021/acs.est.0c02593.

(2) 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. *Environ. Sci. Technol.* **2020**, *54* (5), 2575–2584. https://doi.org/10.1021/acs.est.9b06379.

(3) *ZeroPM on GitHub*. https://github.com/ZeroPM-H2020.

(4) *ZeroPM on Zenodo*. Zero pollution of Persistent, Mobile substances on Zenodo. https://zenodo.org/communities/zeropm-h2020/.

(5) Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. InChI - the Worldwide Chemical Structure Identifier Standard. *Journal of Cheminformatics* **2013**, *5* (1), 7. https://doi.org/10.1186/1758-2946-5-7.

(6) R Core Team. R: A Language and Environment for Statistical Computing. https://www.R-project.org/.

(7) Wolf, R. Cleanventory, 2024. https://github.com/ZeroPM-H2020/cleanventory.

(8) PubChem. *PubChem*. https://pubchem.ncbi.nlm.nih.gov/ (accessed 2023-01-13).

(9) Jacobs, A.; Williams, D.; Hickey, K.; Patrick, N.; Williams, A. J.; Chalk, S.; McEwen, L.; Willighagen, E.; Walker, M.; Bolton, E.; Sinclair, G.; Sanford, A. CAS Common Chemistry in 2021: Expanding Access to Trusted Chemical Information for the Scientific Community. *J. Chem. Inf. Model.* **2022**, *62* (11), 2737–2743. https://doi.org/10/grnnv9.

(10) *CAS Common Chemistry*. https://commonchemistry.cas.org/ (accessed 2023-01-13).

(11) Williams, A. J.; Grulke, C. M.; Edwards, J.; McEachran, A. D.; Mansouri, K.; Baker, N. C.; Patlewicz, G.; Shah, I.; Wambaugh, J. F.; Judson, R. S.; Richard, A. M. The CompTox Chemistry Dashboard: A Community Data Resource for Environmental Chemistry. *Journal of Cheminformatics* **2017**, *9* (1). https://doi.org/10.1186/s13321-017-0247-6.

(12) National Cancer Institute. *Chemical Identifier Resolver*. https://cactus.nci.nih.gov/chemical/structure.

(13) Breitling, R.; Armengaud, P.; Amtmann, A.; Herzyk, P. Rank Products: A Simple, yet Powerful, New Method to Detect Differentially Regulated Genes in Replicated Microarray Experiments. *FEBS Letters* **2004**, *573* (1), 83–92. https://doi.org/10.1016/j.febslet.2004.07.055.

(14) 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* **2011**, *3* (1), 33. https://doi.org/10.1186/1758-2946-3-33.

(15) Backman, T. W. H.; Cao, Y.; Girke, T. ChemMine Tools: An Online Service for Analyzing and Clustering Small Molecules. *Nucleic Acids Res* **2011**, *39* (Web Server issue), W486–W491. https://doi.org/10.1093/nar/gkr320.

(16) European Union. *Commission Delegated Regulation (EU) 2023/707 of 19 December 2022 Amending Regulation (EC) No 1272/2008 as Regards Hazard*

Classes and Criteria for the Classification, Labelling and Packaging of Substances and Mixtures (Text with EEA Relevance); 2022; Vol. 093. http://data.europa.eu/eli/reg_del/2023/707/oj/eng (accessed 2024-09-09).

(17) European Commission. *COMMISSION DELEGATED REGULATION (EU) .../... of XXX Amending Regulation (EC) No 1272/2008 as Regards Hazard Classes and Criteria for the Classification, Labelling and Packaging of Substances and Mixtures*; Classification, Labelling and Packaging; Annex Ares(2022)6485391-20/09/2022; European Commission, 2022; p 30. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13578-Hazardous-chemicals-updated-rules-on-classification-labelling-and-packaging_en.

(18) UN. Globally Harmonized System of the Classification and Labelling of Chemicals (GHS). Tenth Edition., 2023. https://unece.org/sites/default/files/2023-07/GHS%20Rev10e.pdf.

(19) Lohninger, H. Estimation of Soil Partition Coefficients of Pesticides from Their Chemical Structure. *Chemosphere* **1994**, *29* (8), 1611–1626. https://doi.org/10/dtzpkn.

(20) Sabljić, A.; Güsten, H.; Verhaar, H.; Hermens, J. QSAR Modelling of Soil Sorption. Improvements and Systematics of Log KOC vs. Log KOW Correlations. *Chemosphere* **1995**, *31* (11), 4489–4514. https://doi.org/10/b92q3r.

(21) Nguyen, T. H.; Goss, K.-U.; Ball, W. P. Polyparameter Linear Free Energy Relationships for Estimating the Equilibrium Partition of Organic Compounds between Water and the Natural Organic Matter in Soils and Sediments. *Environ. Sci. Technol.* **2005**, *39* (4), 913–924. https://doi.org/10.1021/es048839s.

(22) Niederer, C.; Goss, K.-U.; Schwarzenbach, R. P. Sorption Equilibrium of a Wide Spectrum of Organic Vapors in Leonardite Humic Acid: Modeling of Experimental Data. *Environ. Sci. Technol.* **2006**, *40* (17), 5374–5379. https://doi.org/10/bwzx53.

(23) Endo, S.; Grathwohl, P.; Haderlein, S. B.; Schmidt, T. C. LFERs for Soil Organic Carbon−Water Distribution Coefficients (KOC) at Environmentally Relevant Sorbate Concentrations. *Environ. Sci. Technol.* **2009**, *43* (9), 3094–3100. https://doi.org/10/fhq8cg.

(24) Franco, A.; Fu, W.; Trapp, S. Influence of Soil pH on the Sorption of Ionizable Chemicals: Modeling Advances. *Environmental Toxicology and Chemistry* **2009**, *28* (3), 458–464. https://doi.org/10/c4mbqs.

(25) Poole, C. F.; Atapattu, S. N.; Poole, S. K.; Bell, A. K. Determination of Solute Descriptors by Chromatographic Methods. *Analytica Chimica Acta* **2009**, *652* (1), 32–53. https://doi.org/10.1016/j.aca.2009.04.038.

(26) Bronner, G.; Goss, K.-U. Predicting Sorption of Pesticides and Other Multifunctional Organic Chemicals to Soil Organic Carbon. *Environ. Sci. Technol.* **2011**, *45* (4), 1313–1319. https://doi.org/10/dqsv37.

(27) Jagiello, K.; Sosnowska, A.; Walker, S.; Haranczyk, M.; Gajewicz, A.; Kawai, T.; Suzuki, N.; Leszczynski, J.; Puzyn, T. Direct QSPR: The Most Efficient Way of Predicting Organic Carbon/Water Partition Coefficient (Log KOC) for Polyhalogenated POPs. *Struct Chem* **2014**, *25* (3), 997–1004. https://doi.org/10/f5xwvs.

(28) Mansouri, K. Kmansouri/OPERA, 2018. https://github.com/kmansouri/OPERA (accessed 2020-02-05).

(29) Mansouri, K.; Grulke, C. M.; Judson, R. S.; Williams, A. J. OPERA Models for Predicting Physicochemical Properties and Environmental Fate Endpoints. *Journal of Cheminformatics* **2018**, *10* (1). https://doi.org/10.1186/s13321-018-0263-1.

(30) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V. V.; Tanchuk, V. Y.; Todeschini, R.; Varnek, A.; Marcou, G.; Ertl, P.; Potemkin, V.; Grishina, M.; Gasteiger, J.; Schwab, C.; Baskin, I. I.; Palyulin, V. A.; Radchenko, E. V.; Welsh, W. J.; Kholodovych, V.; Chekmarev, D.; Cherkasov, A.; Aires-de-Sousa, J.; Zhang, Q.-Y.; Bender, A.; Nigsch, F.; Patiny, L.; Williams, A.; Tkachenko, V.; Tetko, I. V. Online Chemical Modeling Environment (OCHEM): Web Platform for Data Storage, Model Development and Publishing of Chemical Information. *J Comput Aided Mol Des* **2011**, *25* (6), 533–554. https://doi.org/10.1007/s10822-011-9440-2.

(31) OECD. *eChemPortal*. The Global Portal to Information on Chemical Substances. https://www.echemportal.org/echemportal/ (accessed 2022-10-08).

(32) The OECD QSAR Toolbox. https://www.oecd.org/chemicalsafety/risk-assessment/oecd-qsar-toolbox.htm (accessed 2022-10-16).

(33) Franco, A.; Trapp, S. Estimation of the Soil–Water Partition Coefficient Normalized to Organic Carbon for Ionizable Organic Chemicals. *Environmental Toxicology and Chemistry* **2008**, *27* (10), 1995–2004. https://doi.org/10/cqsd2f.

(34) PhysChem Suite.

(35) Abraham, M. H.; Gola, J. M. R.; Cometto-Muñiz, J. E.; Cain, W. S. Solvation Properties of Refrigerants, and the Estimation of Their Water–Solvent and Gas–Solvent Partitions. *Fluid Phase Equilibria* **2001**, *180* (1), 41–58. https://doi.org/10.1016/S0378-3812(00)00511-2.

(36) Brown, T. N. IFSQSAR - A Python Package for Applying QSARs, 2022. https://github.com/tnbrowncontam/ifsqsar (accessed 2020-05-20).

(37) Brown, T. N. QSPRs for Predicting Equilibrium Partitioning in Solvent–Air Systems from the Chemical Structures of Solutes and Solvents. *J Solution Chem* **2022**, *51* (9), 1101–1132. https://doi.org/10.1007/s10953-022-01162-2.

(38) Poole, S. K.; Poole, C. F. Chromatographic Models for the Sorption of Neutral Organic Compounds by Soil from Water and Air. *Journal of Chromatography A* **1999**, *845* (1), 381–400. https://doi.org/10/bjd7gm.

(39) Zhang, Z.; Sangion, A.; Wang, S.; Gouin, T.; Brown, T.; Arnot, J. A.; Li, L. Chemical Space Covered by Applicability Domains of Quantitative Structure–Property Relationships and Semiempirical Relationships in Chemical Assessments. *Environ. Sci. Technol.* **2024**. https://doi.org/10.1021/acs.est.3c05643.

(40) Chirico, N.; Sangion, A.; Gramatica, P.; Bertato, L.; Casartelli, I.; Papa, E. QSARINS-Chem Standalone Version: A New Platform-Independent Software to Profile Chemicals for Physico-Chemical Properties, Fate, and Toxicity. *Journal of Computational Chemistry* **2021**, *42* (20), 1452–1460. https://doi.org/10/gjzn79.

(41) Gramatica, P.; Cassani, S.; Chirico, N. QSARINS-Chem: Insubria Datasets and New QSAR/QSPR Models for Environmental Pollutants in QSARINS. *Journal of Computational Chemistry* **2014**, *35* (13), 1036–1044. https://doi.org/10.1002/jcc.23576.

(42) Arp, H. P. H.; Hale, S. E. Assessing the Persistence and Mobility of Organic Substances to Protect Freshwater Resources. *ACS Environmental Au* **2022**, *2* (6), 482–509.

(43) United States Environmental Protection Agency. Estimation Programs Interface Suite™, 2012. https://www.epa.gov/tsca-screening-tools/download-epi-suitetm-estimation-program-interface-v411.
(44) *Should Transformation Products Change the Way We Manage Chemicals? | Environmental Science & Technology*. https://pubs.acs.org/doi/10.1021/acs.est.4c00125 (accessed 2024-09-24).
(45) Schymanski, E. L.; Kondić, T.; Neumann, S.; Thiessen, P. A.; Zhang, J.; Bolton, E. E. Empowering Large Chemical Knowledge Bases for Exposomics: PubChemLite Meets MetFrag. *J Cheminform* **2021**, *13* (1), 19. https://doi.org/10.1186/s13321-021-00489-0.
(46) Mohammed Taha, H.; Aalizadeh, R.; Alygizakis, N.; Antignac, J.-P.; Arp, H. P. H.; Bade, R.; Baker, N.; Belova, L.; Bijlsma, L.; Bolton, E. E.; Brack, W.; Celma, A.; Chen, W.-L.; Cheng, T.; Chirsir, P.; Čirka, Ľ.; D'Agostino, L. A.; Djoumbou Feunang, Y.; Dulio, V.; Fischer, S.; Gago-Ferrero, P.; Galani, A.; Geueke, B.; Głowacka, N.; Glüge, J.; Groh, K.; Grosse, S.; Haglund, P.; Hakkinen, P. J.; Hale, S. E.; Hernandez, F.; Janssen, E. M.-L.; Jonkers, T.; Kiefer, K.; Kirchner, M.; Koschorreck, J.; Krauss, M.; Krier, J.; Lamoree, M. H.; Letzel, M.; Letzel, T.; Li, Q.; Little, J.; Liu, Y.; Lunderberg, D. M.; Martin, J. W.; McEachran, A. D.; McLean, J. A.; Meier, C.; Meijer, J.; Menger, F.; Merino, C.; Muncke, J.; Muschket, M.; Neumann, M.; Neveu, V.; Ng, K.; Oberacher, H.; O'Brien, J.; Oswald, P.; Oswaldova, M.; Picache, J. A.; Postigo, C.; Ramirez, N.; Reemtsma, T.; Renaud, J.; Rostkowski, P.; Rüdel, H.; Salek, R. M.; Samanipour, S.; Scheringer, M.; Schliebner, I.; Schulz, W.; Schulze, T.; Sengl, M.; Shoemaker, B. A.; Sims, K.; Singer, H.; Singh, R. R.; Sumarah, M.; Thiessen, P. A.; Thomas, K. V.; Torres, S.; Trier, X.; Van Wezel, A. P.; Vermeulen, R. C. H.; Vlaanderen, J. J.; Von Der Ohe, P. C.; Wang, Z.; Williams, A. J.; Willighagen, E. L.; Wishart, D. S.; Zhang, J.; Thomaidis, N. S.; Hollender, J.; Slobodnik, J.; Schymanski, E. L. The NORMAN Suspect List Exchange (NORMAN-SLE): Facilitating European and Worldwide Collaboration on Suspect Screening in High Resolution Mass Spectrometry. *Environ Sci Eur* **2022**, *34* (1), 104. https://doi.org/10.1186/s12302-022-00680-6.
(47) Schymanski, E.; Bolton, E.; Cheng, T.; Thiessen, P.; Zhang, J. (Jeff); Helmus, R.; Blanke, G. Transformations in PubChem - Full Dataset, 2024. https://doi.org/10.5281/ZENODO.10998412.
(48) Helmus, R.; Van De Velde, B.; Brunner, A. M.; Ter Laak, T. L.; Van Wezel, A. P.; Schymanski, E. L. patRoon 2.0: Improved Non-Target Analysis Workflowsincluding Automated Transformation Product Screening. *JOSS* **2022**, *7* (71), 4029. https://doi.org/10.21105/joss.04029.
(49) Schymanski, E. L.; Bolton, E. E. FAIRifying the Exposome Journal: Templates for Chemical Structures and Transformations. *Exposome* **2022**, *2* (1), osab006. https://doi.org/10.1093/exposome/osab006.
(50) *ShinyTPs: Curating Transformation Products from Text Mining Results | Environmental Science & Technology Letters*. https://pubs.acs.org/doi/10.1021/acs.estlett.3c00537 (accessed 2024-09-24).
(51) Djoumbou-Feunang, Y.; Fiamoncini, J.; Gil-de-la-Fuente, A.; Greiner, R.; Manach, C.; Wishart, D. S. BioTransformer: A Comprehensive Computational Tool for Small Molecule Metabolism Prediction and Metabolite Identification. *J Cheminform* **2019**, *11* (1), 2. https://doi.org/10.1186/s13321-018-0324-5.

(52) Chirsir, P.; Palm, E. H.; Baskaran, S.; Schymanski, E. L.; Wang, Z.; Wolf, R.; Hale, S. E.; Arp, H. P. H. Grouping Strategies for Assessing and Managing Persistent and Mobile Substances. *Environ Sci Eur* **2024**, *36* (1), 102. https://doi.org/10.1186/s12302-024-00919-4.

8 Appendix 1

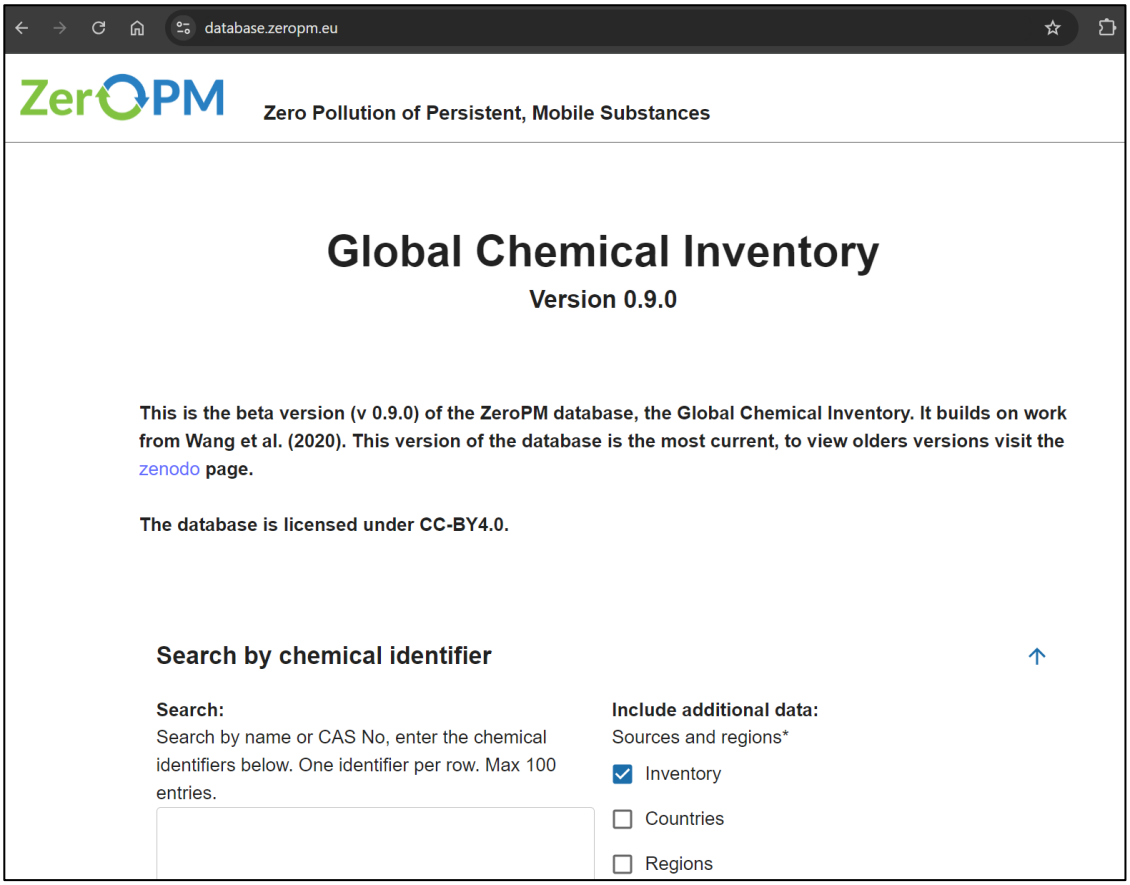

Figure 8 Screenshot of the graphical user interface (GUI) implemented for the global chemical inventory. The screenshot was taken on 24th August 2024., and is accessible at https://database.zeropm.eu/

H2020 project Zero pollution of Persistent, Mobile substances
EU Grant Agreement No. 101036756