Toxikologiska rådet
– expertorgan för rådgivning och samråd i toxikologiska frågor
The Toxicological Council
– body of experts for advice and consultation on toxicological issues

REPORT 1/21

Research report 2021

Methods for early identification of chemicals that have the potential to harm human health or the environment

Preface

The Toxicological Council is an expert organisation established to facilitate the rapid identification of chemical substances that can be harmful to human health or the environment. The Council includes representatives from governmental authorities and academic institutions. The Toxicological Council identifies and evaluates signals of new, potential and emerging chemical risks and reports its findings to SamTox. This project was conducted as a consultancy commission in order to develop methodologies for identification of new or emerging chemical risks.

The report was written by Suzanne Bruks and Prof. Patrik Andersson from the Department of Chemistry at Umeå University and Ph.D. Vera Franke and Prof. Karin Wiberg from the Department of Aquatic Sciences and Assessment at the Swedish University of Agricultural Sciences (SLU).

The project reference group consisted of Lina Wendt-Rasch, Emma Westerholm, Olof Johansson, Erik Gravenfors and Carl-Henrik Eriksson from the Swedish Chemicals Agency (KemI) and Michael Pettersson from Swedish Geotechnical Institute (SGI). Lina Wendt-Rasch was the project leader and contact person at KemI, which financed the project.

The Swedish Chemicals Agency financed the project in order to support the Toxicological council's assignment to provide updated and relevant information to SamTox. The conclusions presented in the report represent the views of authors and do not necessarily reflect the opinions of individual authorities and academic institutions in the Toxicological council.

Contents

Abbreviations and Glossary

Word	Description
Arnot-Gobas	The Arnot & Gobas (2003) bioconcentration model under VEGA
B	Bioaccumulative
BAF	Bioaccumulation factor
BCF	Bioconcentration factor
BCFBAF	Model under EPI Suite to predict bioconcentration and bioaccumulation
BIOWIN-3	Model under EPI Suite to predict biodegradation
BPA	Phenol, 4,4'-(1-methylethylidene)bis-, also referred to as Bisphenol A
C	Carcinogenic
Caesar	Platform of QSAR models
CAS Number	Unique numerical identifier assigned by the Chemical Abstracts Service (CAS), also referred to as CAS Registry Number or CASRN
ChemIDPlus	Freely available chemical database run by the United States National Library of Medicine
CIR	Chemical identifier resolver – platform for converting a given chemical structure identifier into another representation or structure identifier
CLH	Harmonised classification and labelling – harmonised classification system for chemical hazards throughout the EU to ensure an adequate risk management
CLP	Classification, Labelling and Packaging – a European Union regulation for chemical substances and mixtures based on the Globally Harmonized System (GHS)
CompTox	Information platform developed by the US Environmental Protection Agency
Data curation	Method to prepare an initial chemical inventory or database for computational approaches
Data mining	Method of information processing to collect data relevant for further analysis (in the case of this report: extracting chemical structures from initial datasets for the subsequent application of computational approaches)
DDT	Benzene, 1,1'-(2,2,2-trichloroethylidene)bis[4-chloro-
DEHP	1,2-Benzenedicarboxylic acid, 1,2-bis(2-ethylhexyl) ester
ED	Endocrine disrupting
EPI Suite	Estimation Program Interface – suite of estimation programs for physicochemical property and environmental fate provided by the US Environmental Protection Agency
Endpoint	As used in this report: certain physicochemical, biological and environmental fate properties of a chemical compound
ER	Estrogen receptor
GC-MS	Gas chromatography coupled to mass spectrometry
Hazard screening	Application of models for risk rating of chemical compounds towards a number of endpoints
InChI	International Chemical Identifier – textual identifier for chemical substances initially developed by IUPAC (International Union of Pure and Applied Chemistry)

IRFMN	QSAR models developed by Istituto di Ricerche Farmacologiche Mario Negri available under the platform VEGA
KNIME	Konstanz Information Miner – open source software for data curation
K_{oa}	Octanol-air partition coefficient
K_{oc}	Organic carbon-water partition coefficient
K_{ow}	Octanol-water partition coefficient
LC50	Lethal concentration, 50% – the concentration of a substance required to kill half of a tested population
LC-MS	Liquid chromatography coupled to mass spectrometry
M	Mobility
M	Mutagenic
MassBank Europe	Database for spectral data from high-resolution mass spectrometry studies
MCI	Molecular connectivity indices – a class of molecular descriptors
NOEC	No Observed Effect Concentration
Non-target screening	Non-targeted chemical analysis of a sample using high-resolution mass spectrometry
NORMAN	Network of stakeholders dealing with emerging environmental substances
nP	Not persistent
HBCDD	1,2,5,6,9,10-Hexabromo-cyclododecane
OECD	Organisation for Economic Co-operation and Development
OpenMolecule	Platform for cheminformatics tools
OPERA	Open source QSAR model platform
P	Persistent
PACT	Public activities coordination tool – overview of substance-specific activities that authorities are working on under REACH and the CLP regulation, provided by ECHA
PBT substances	Chemical compounds classified to be persistent, bioaccumulative and toxic
PFAS	Perfluoroalkyl and polyfluoroalkyl substances, also referred to as PFC (perfluorinated compounds)
PFOS	1-Octanesulfonic acid, 1,1,2,2,3,3,4,4,5,5,6,6,7,7,8,8,8-heptadecafluoro-
POPs	Persistent Organic Pollutants
PubChem	Chemical database maintained by the National Library of Medicine (part of the United States National Institutes of Health)
QSAR	Quantitative structure–activity relationship
R	Reproductive toxicity
REACH	Registration, Evaluation, Authorisation and Restriction of Chemicals
SamTox	Coordination group for new potential chemical threats consisting of heads of Swedish authorities working with chemical safety.

SciFinder	Chemical database developed by the Chemical Abstracts Service (CAS)
SMILES	Simplified molecular-input line-entry system – line notation describing a chemical structure
Ss	Skin sensitisation
SVHC	Substance of Very High Concern
TBBPA	Phenol, 4,4'-(1-methylethylidene)bis[2,6-dibromo-
TCEP	Ethanol, 2-chloro-, 1,1',1''-phosphate
T	Toxic
TPhP	Phosphoric acid, triphenyl ester
VEGA	Platform providing various QSAR models, also referred to as VEGA-QSAR
vB substances	Chemical substances classified as very bioaccumulative
vP substances	Chemical substances classified as very persistent
Wskow	Model for predicting water solubility under EPI Suite

Sammanfattning

För att på ett tidigt stadium identifiera kemikalier som potentiellt skadar människors hälsa eller miljön, definierade som NERCs (New or Emerging Risk Chemicals), krävs systematiskt arbete och metodik. Toxikologiska rådet har i uppdrag att hitta och utvärdera potentiella NERCs. NERCs kan identifieras genom expertbedömningar och data från vetenskapliga rapporter men Toxikologiska rådet ser behov av att komplettera detta tillvägagångsätt med en mer systematisk metod för tidig identifiering av NERCs som möjliggör proaktiva åtgärder. Denna rapport presenterar en första utveckling av ett tidigt varningssystem baserat på en metod för datakurering följt av faroscreening med befintliga beräkningsmodeller.

Metoden bygger på att NERCs identifieras i befintliga kemiska kartläggningar (kemikalieinventeringar) för olika produkttyper på den europeiska marknaden. Dessa kartläggningar sammanställs regelbundet av till exempel nationella och internationella myndigheter och kännetecknas vanligtvis av stor variation i uppgivna kemikalienamn och strukturer. Datakurering är därför ett viktigt första delsteg som resulterar i väldefinierade kemiska strukturer och som lämpar sig för användning i beräkningsmodeller. I denna process avlägsnas till exempel oorganiska ämnen, organometaller, polymerer och komplexa blandningar.

En nyligen publicerad arbetsmodell för kurering av data testades på fyra olika kemikalieinventeringar, inklusive kemikalier som används i pappers- och pappindustrin, plasttillsatser, poly- och perfluorerade alkylsubstanser (PFAS) och kemikalier som används i vardagsprodukter. Sammanfattningsvis visade modellen lovande resultat för att tidseffektivt kurera stora datamängder. Resultaten detaljanalyserades för två av de undersökta inventeringarna med endast 4 % falskt positiva som resultat. Totalt identifierades det att 35–80 % av ämnena är lämpliga för beräkningsmodeller och att förkastade substanser främst består av polymerer eller data med felaktiga eller tvetydiga kemiska strukturer. De förkastade ämnena ansågs generellt som korrekt bortvalda. Av kemikalier från pappers- och pappindustrin och plastadditiv kvalificerade 134 respektive 138 som NERCs efter screening av egenskaper för persistens och bioackumulation.

En närmare undersökning av de identifierade potentiella NERCs genomfördes genom en omfattande litteratur- och databasstudie. Förekomst av en stor andel av kemikalierna kunde verifieras i livsmedel, plaster och livsmedelsförpackningar samt i biota, sediment, ytvatten och avloppsvatten. Generellt sett är sökandet efter förekomstdata för NERCs i olika matriser ett krävande manuellt steg, vilket dock kan leda till värdefull ny information som kan användas för utformningen av framtida screeningstudier eller övervakningsprogram. Detta steg ger värdefull information om relevanta matriser och vilken typ av kemisk analys som är mest lämplig för screening eller övervakning. Resultaten från litteraturstudien tyder på att flera identifierade ämnen är NERCs, varav många finns i matriser som är relevanta för exponering av biota och människor. Det utvecklade verktyget kan således rekommenderas som ett tidigt varningssystem. Dock identifierades ett antal utmaningar i datakureringen, tillämpningen av beräkningsmodeller och i den efterföljande sökningen av förekomstdata, vilket kräver fortsatt utveckling av metodiken.

Summary

It is important to have access to methods for early identification of chemicals that have the potential to harm human health or the environment, defined as New or Emerging Risk Chemicals (NERCs). The Toxicological Council of Sweden has been given the assignment to find and evaluate potential NERCs. NERCs are typically identified by expert judgement or data from scientific reports; however, a systematic methodology for early identification of NERCs would enable proactive measures. This report presents an initial development of an early warning system based on a computational data curation methodology followed by hazard screening using existing estimation platforms. Selected case studies were followed by an in-depth search for evidence of occurrence of candidate NERCs in various environmental and human matrices.

NERCs can be identified from existing chemical inventories for different product types on the European market. Chemical inventories are typically characterised by large variation in input information including chemical names and structures. The data-curation step should be suitable for various datasets and chemical inventories compiled by national and international authorities. Data curation is a critical phase enabling the compilation of a dataset with well-defined chemical structures prepared for computational approaches. This phase includes several steps, such as removal of inorganics, organometallics, counter ions, macromolecules, and complex mixtures.

A publicly available data curation method was tested on four different datasets including chemicals used in the paper and paperboard industry, plastic additives, per- and polyfluoroalkyl substances (PFASs), and chemicals used in everyday products. In summary, the workflow showed promising results yielding a fast method to derive curated data for large datasets. A detailed analysis based on random manual checks revealed a false positive rate of 4% in two curated databases. The investigated databases included 35-80% entries applicable for available hazard screening tools. A large fraction of the entries was lost during data curation, either because compounds were not suited for hazard screening tools (e.g. polymers) or due to erroneous/ambiguous data. Chemicals classified as ‘rejected’ were generally found to be correctly assigned. Curated databases of paper and paperboard chemicals and plastic additives were screened for hazard properties, and 134 and 138 of the entries, respectively, qualified for set NERC criteria of persistence and bioconcentration potential.

Mapping of occurrence data can aid decisions on experimental design of future screening studies as it provides information on relevant matrices and the type of chemical analysis most appropriate. In this project, an investigation of the identified NERC candidates was conducted through an extensive literature and database study. Occurrence of a large fraction of the compounds could be confirmed in food, consumer products in polymeric materials and food packaging, as well as in biota, sediment, surface waters and wastewater. Generally, the search for evidence of occurrence of NERCs in various matrices is a labour-intensive step, which, however, can lead to valuable new information for the design of future screening studies or monitoring programs.

Results suggest that several compounds selected through the workflow are potential NERCs with undesirable properties, many of which can be found in matrices relevant for exposure of biota and humans. The developed tool for identification of NERCs can thus be recommended as an early warning system. A number of challenges in data curation, the application of computational approaches as well as for the subsequent search of occurrence data were identified and emphasise the need for further development.

1 Aims and Objectives

The aim of this project was to develop and evaluate a systematic computational method to identify and prioritise New or Emerging Risk Chemicals (NERCs). Specific objectives were to (i) screen relevant data bases, (ii) identify or develop a standardized and automatic data curation step consisting of a generic model suitable for different kinds of data, (iii) generate lists of potential risk chemicals that should be further assessed and examined, and (iv) screen for the occurrence of the identified NERCs in human and environmental matrices using existing data. The compiled lists of potential NERCs and the overall methodology will be investigated and developed further by the Toxicological Council in order to identify NERCs.

2 Introduction

Early warning systems for identification of emerging chemical risks are critical towards minimising the production and usage of chemicals that may cause damage to human health and the environment1. Legislation is generally not proactive and thus complementary tools are needed for identification of unknown and unexpected risk chemicals. It is therefore warranted to develop early warning systems to proactively identify chemicals of concern for early initiation of risk management measures[2–4].

New or Emerging Risk Chemicals (NERCs) can be defined as chemicals that, for various reasons, cause a new or increasing risk to the environment or human health. The Toxicological Council has been given the assignment to find and evaluate potential NERCs, and identification can be made by different methods, such as expert judgements, monitoring chemical substitution processes, reviews of scientific literature, and systematic chemical analytical or biomonitoring screening. Ideally, however, potentially harmful substances should be identified before monitoring data is available. Thus, the aim of this project was to develop and evaluate a systematic computational method to identify and prioritise NERCs.

One of the requirements of this project was to develop a methodology applicable on different kind of input data, such as datasets and chemical inventories compiled by authorities like the Swedish Chemicals Agency (KemI), ECHA and the Ministry of Environment of Denmark. Chemical inventories are typically characterised by large variation in input information including chemical names, structures and other substance identifiers. Input may include unique individual chemicals but also complex chemical mixtures, unusual elements, and polymers, compounds estimated to represent > 30% of industrial chemicals[5].

A major challenge for the identification of NERCs is therefore the inhomogeneous use of chemical identifiers in different datasets and chemical inventories hampering the effective compilation of homogenised lists for hazard screening. Consequently, data curation is a critical phase that warrants development to construct a database with well-defined identifiers for chemical structures, such as SMILES, prepared for various computational approaches. The curation step is also important when handling large datasets, as these may contain errors that would impact the final result. The usage of a systematic workflow based on data mining have been suggested as a good approach to perform data curation[6]. Data curation includes several individual steps, such as removal of inorganics, organometallics, counter ions, macromolecules, mixtures, tautomeric forms etc., to optimise the raw data into chemical structures suitable for computational analysis[7–9].

To collect structural information from several sources simultaneously increases the reliability of the correctness of the structure. By ignoring the step of data curation and just use bulk data of various kind into hazard screening modelling, a large number or unreliable results will be obtained. Therefore, one of the major objectives of this study was to identify or develop a standardised and automatic data curation step consisting of a generic model suitable for different sorts of data (e.g., pharmaceuticals, industrial chemicals, compound specific lists and screening results) for generation of standardised molecular structures.

3 Material and Methods

3.1 General methodology

The developed strategy consists of several steps (Figure 1), beginning with a semi-automated data curation step ("Data curation"), followed by computational approaches (e.g. QSARs) to calculate environmental and human health hazard endpoints ("Hazard screening"). The combined results from several hazard screening approaches were then used for risk rating to produce lists of potential NERCs, which subsequently were investigated further in a literature and database study focusing on chemical occurrence in various matrices ("Occurrence study").

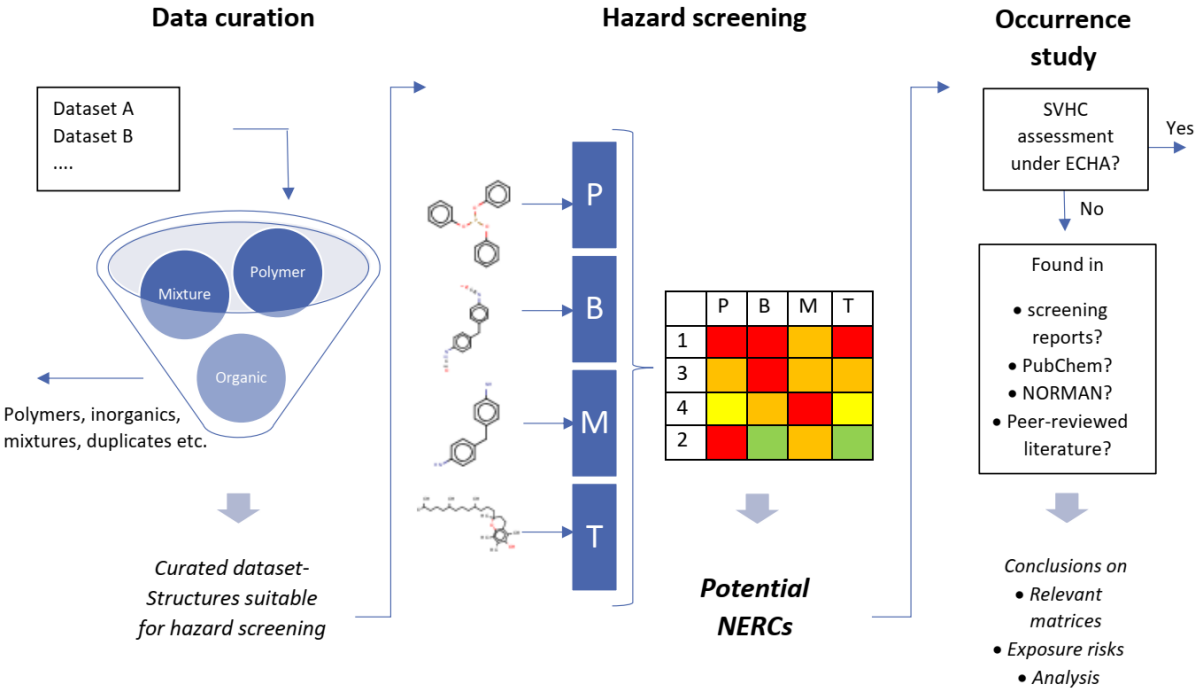


Figure 1. Overview of the methodology applied in this project.

The developed data curation method was tested on four different datasets, whereof two of the resulting output data were applied and analysed in the hazard screening and occurrence study steps. The datasets used were KemI's subset of substances with relevance to paper and paperboard products manufacturing (high and medium high risk of being found on the Swedish market)[10], KemI's subset of substances with relevance to plastic product manufacturing based on the Swedish product register[11], the data inventory of per- and polyfluoroalkyl substances (PFASs) published by the OECD/UNEP Global PFC Group NS[12], and the Ministry of Environment of Denmark's dataset over chemicals in everyday products[13] (see also Section

3.2). To assess the quality of our predictions, positive controls were used (n=10, Figure S1 and Table S1 in the Supplementary information). They were chosen as representatives of well-known hazardous substances for the environment and human health. Properties established by ECHA for these substances were compiled and compared to the predictions made from the hazard screening models.

Hazard screening was performed by applying open source models to cover a broad spectrum of endpoints focused on, in particular, persistence, bioaccumulation and mobility (see Section 3.4). The mobility endpoint was chosen to also include polar organic compound rather than just very hydrophobic (nonpolar) substances, which often score high for persistence and bioaccumulation[14]. Data was evaluated based on mutual hazard scoring of the same endpoint from several different models. To capture the objective of identifying NERCs, hazard screening of chemicals was based on a conservative approach, thus applying a safety margin on threshold values currently applied in chemical regulation. The generated lists were further filtered based on ongoing risk evaluations by ECHA[15–17].

3.2 Studied datasets

3.2.1 Paper and paperboard

As part of the Swedish Government's commission to KemI to identify hazardous substances in chemical products and articles, a report on substances identified in the paper industry was presented 2019[10]. The aim of the report was to identify substances that can be present on the Swedish market in consumer products of paper and paperboard. In total, a list of over 17,000 substances was compiled.

Due to the large amount of data presented in the study[10], the substances were prioritised based on probable occurrence on the Swedish market for the purpose of the current report. Lists of 571 substances with high probability and 1,333 substances with medium probability to be found on the Swedish market were merged for data curation. The original data had some missing CAS Numbers and these were localised and all empty cells renamed NA, as the data curation workflow does not accept empty cells.

3.2.2 Plastic additives

Within the assignment to map hazardous substances in chemical products and articles, KemI has also performed mapping of hazardous substances in plastic materials[11]. Within that project, a vast number of process chemicals and functional additives in plastics (many identified as substances of concern for human health) were targeted and prioritised. The aim of the study was to contribute to the general knowledge of chemical substances in plastics, using a dataset from the Swedish product register related to plastics manufacturing. The dataset contained roughly 2,500 substances registered in the product register between 2010 and 2016, and was recently investigated with the aim to identify previously unknown PBT/vPvB substances based on modelling using derived SMILES[18]. That project did not include any automatic and systematic data curation step. Here, we analysed the importance of the data curation step by comparing the outcome of the two studies, see comparison under Section 4.1.2.

3.2.3 PFASs

The compiled data inventory of 4730 PFASs, based on the definition to contain at least one perfluoroalkyl moiety, summoned by OECD[12] and published 2018, was used within this project as an example for a single group of chemicals. The results from the data curation were compared with a recently published study on the same dataset, see Section 4.1.2[19]. The PFASs list was downloaded from the OECD webpage in November 2020.

3.2.4 Everyday products

The Ministry of Environment of Denmark has performed mapping of chemicals in consumer products since 2003, which is compiled in 182 reports to date[13]. The findings are summarised in a database covering chemicals found in everyday products. The dataset is publicly available from their webpage. The list, however, contained several duplicates, as the same compound has been detected in several studies, and also erroneous CAS Numbers. Thus, the dataset (Excel files) was pre-treated before use. Notably, many of the substances in the list were spelled in Danish. The dataset was downloaded in November 2020.

3.2.5 Positive controls

Ten positive controls were selected to represent well-studied and known hazardous chemicals, displaying varying hazard profiles, see Table 1 and Figure S1 in the Supplementary information). These represent known persistent organic pollutants (POPs), e.g. DDT and HBCDD, mobile chemicals such as PFOS and BPA, but also chemicals of concern for human health and reproductive toxicity, for instance DEHP. Molecular structures of the positive controls are given in Figure S1 in the Supplementary information.

Table 1. Positive controls used in this study with abbreviation, chemical name, CAS Number, and hazard labelling according to ECHA[20–29]. For structures, see Figure S1 in the Supplementary information.

No	Abbreviation	Chemical name	CAS Number	Hazard labelling
1	HBCDD	Cyclododecane, 1,2,5,6,9,10-hexabromo-	3194-55-6	PBT, POP, R, Ss
2	Triclosan	Phenol, 5-chloro-2-(2,4-dichlorophenoxy)-	3380-34-5	PBT, ED
3	DDT	Benzene, 1,1'-(2,2,2-trichloroethylidene)bis[4-chloro-	50-29-3	POP, C
4	TBBPA	Phenol, 4,4'-(1-methylethylidene)bis[2,6-dibromo-	79-94-7	PBT, ED
5	BPA	Phenol, 4,4'-(1-methylethylidene)bis-	80-05-7	ED, Ss, R
6	TPhP	Phosphoric acid, triphenyl ester	115-86-6	ED
7	Benzo[a]pyrene	Benzo[a]pyrene	50-32-8	PBT, POP, R, M, C, Ss
8	DEHP	1,2-Benzenedicarboxylic acid, 1,2-bis(2-ethylhexyl) ester	117-81-7	ED, R
9	TCEP	Ethanol, 2-chloro-, 1,1',1''-phosphate	115-96-8	R, C
10	PFOS	1-Octanesulfonic acid, 1,1,2,2,3,3,4,4,5,5,6,6,7,7,8,8,8-heptadecafluoro-(	1763-23-1	POP, R, C

PBT: Persistent, Bioaccumulative and Toxic; POP: Persistent organic pollutant; ED: Endocrine disrupting; CMR: Carcinogenic, Mutagenic and Reprotoxic; Ss: Skin sensitising.

3.3 Data curation using KNIME

To ascertain the tool for data curation would be publicly available, the open-source software Konstanz Information Miner (KNIME)[30] was chosen as platform for the procedure. A recently published KNIME workflow was applied consisting of three major parts (*Figure 2*)[31]. The first two parts are automated, and the third requires manual inspection. The major data curation step takes place under the second part, where retrieval of SMILES from two different databases (the chemical identifier resolver (CIR)[32] and CompTox[33]) followed by several data clean-up steps are performed to generate a preliminary output file for manual inspection. This file also contains information on removed counter ions. The final output consists of a curated dataset of canonical SMILES suitable for hazard screening such as QSAR analysis. The three different parts of the data curation workflow are described in the following sections.

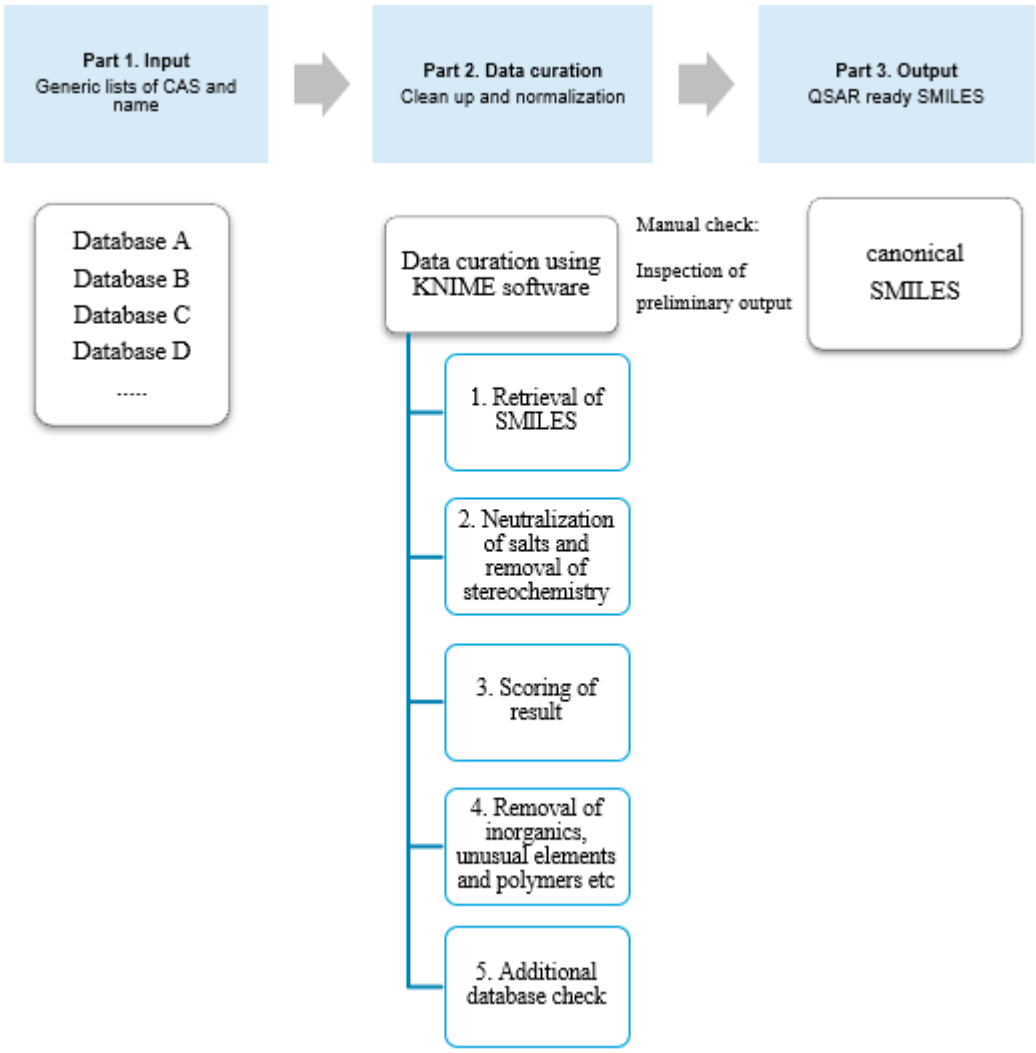


Figure 2. Overview of the workflow for the data curation method developed by Gadaleta et al[31].

3.3.1 Part 1. Input

The data curation workflow can be run on any list of chemicals that contain CAS Registry Numbers and chemical names. An Excel file with three columns needs to be prepared before running the workflow; the first column should include CAS Number, the second chemical

name and the third column should include a compound id number (1, 2, 3…). If any of the compounds have missing CAS Number or name, the cell should not be left empty, but named NA. A manual inspection of the initial data needs to be performed before loading the input file, to make sure any single cell does not contain several CAS Numbers, special symbols or blank spaces. The Excel file is loaded into the KNIME workflow. Instructions on how to run the workflow are included in the Supplementary information.

3.3.2 Part 2. Data curation

The process of the data curation workflow under KNIME is illustrated in Figure 3 (scoring system). For full details, see Gadaleta et al.[31]

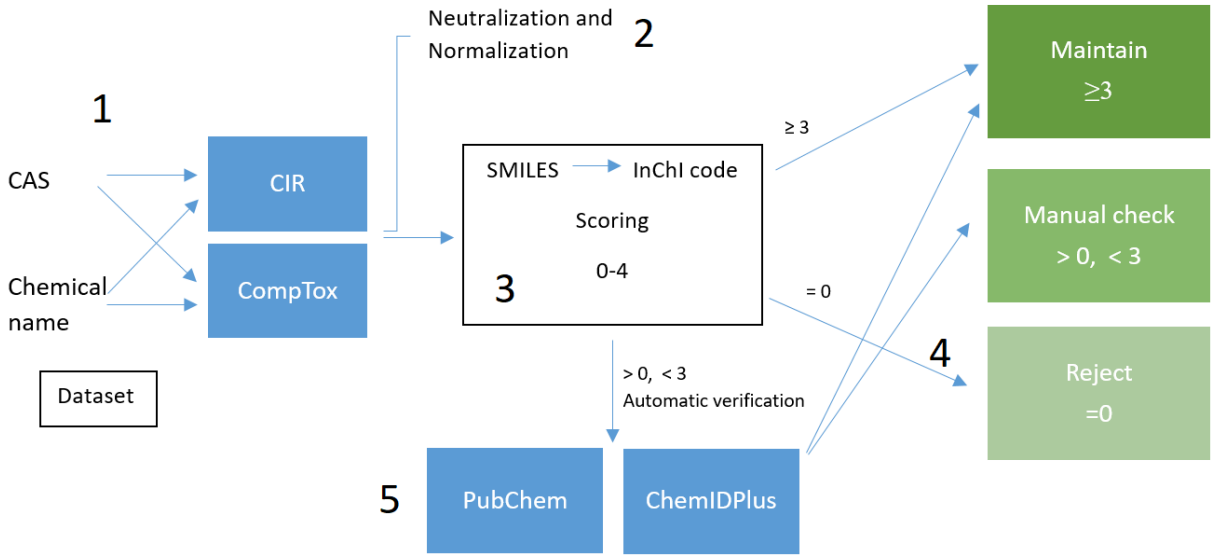


Figure 3. Details on the data curation step including the scoring system of the KNIME workflow.

1. Retrieval of SMILES

The workflow uses a HTML parser to find SMILES from two databases, the chemical identifier resolver (CIR)[32] and CompTox[33] (Figure 3, 1). The workflow is searching SMILES using both the chemical name and the CAS Number, with a maximum retrieval of four structures. If the data collected from the two databases are incongruent, an additional check is performed by the workflow to raise the level of confidence in structure retrieval using two additional databases; PubChem[34] and ChemIDPlus[35] (see more under point 5). This step is rather dependent on given CAS (if entered CAS yield two different structures, the entry is deleted), whereas misspelled names generally yield a larger set for manual inspection.

2. Neutralisation of salts and removal of stereochemistry

After the retrieval of SMILES, the data clean-up and identification of which substances are suitable for hazard screening starts (Figure 2, 2). The first step in data curation is neutralisation of salts and elimination of counter ions. When counter ions, both mono- and polyatomic, are removed the information is saved in the preliminary output file. The main molecule is identified as the one with the highest molecular weight, which is normalised to a neutral form. Since the tools for hazard screening typically apply two-dimensional molecular structure information, any information on stereoisomerism is removed. Such substances will be flagged in the final output file, to notify the user that this could be a compound with a specific molecular configuration. To assure that different tautomeric forms do not cause inconsistency in the output, the SMILES are converted to InChI form (Figure 3, 3). Two stereoisomeric forms of the same compound can be described with a single InChI and thus merged to one entry.

3. Scoring of results

The derived InChI codes are used for scoring the result from the SMILES retrieval (Figure 3, 3). In brief, a maximum score of 4 indicates that equal structures have been collected from name and CAS Number in both CIR and CompTox databases. If the score is ≥ 3 (the same InChI code have been derived 3 or 4 times), the compound will be directly maintained in the final output. A score < 3 activates the system to verify structure as explained below under 5. Errors within the databases are known to exist and this contributes to the importance of structures retrieved from several sources[36].

4. Removal of inorganics, unusual elements and polymers

The compounds are screened for the presence of unusual elements, as tools for hazard screening are not typically designed for these. The usual elements are H, C, N, O, F, Br, I, Cl, P, and S, and compounds containing atoms other than those will be rejected from the final output and flagged with a warning note stating “unusual element”. Small inorganic compounds are also removed in this step. The chemical names are screened for keywords indicating structures unsuitable for typical hazard screening tools, such as polymers, mixtures, metabolites, reaction masses/products, chemicals with variable composition etc. Further keywords can be, for example, “react”, “product”, “isomer”, “polymer”, “mix”. If the system finds such a keyword, it will flag the substance with a warning note in the final output. All of these substances will be given the score 0 and are thus rejected by the workflow.

5. Additional database check

A new retrieval of SMILES based on chemical name and CAS Registry Numbers is performed for substances with a score $0 > x < 3$ from the databases PubChem and ChemIDPlus (Figure 3, 5). If confirmation of structure can be made using these databases, the substance will be transferred to the maintain list. If conformation fails, the substance will go on to manual check.

3.3.3 Part 3. Output

A preliminary output file is derived from the workflow. It contains one part defined as “maintained”, which is ready for use. The file also contains one part that requires a manual check (Figure 3, 4) and one that is rejected. The compounds that are assorted for manual check are those with a scoring < 3. There are three different notes (i-iii) for these, depending on what kind of inconsistency the workflow identified. The procedure for handling these issues follow Gadaleta et al 2018[31] with minor modifications regarding e.g. applied databases.

(i) **Verify name** - There was no success in generating SMILES from the chemical name in any of the data bases. This error might be due to misspelling, or chemical name in another language than English. The name is manually inspected and verified. The output file presents a number of synonyms found based on the CAS Registry Number to simplify the verification in case the name is misspelled. The row in the Excel file is marked with "1" if the name is considered right.

(ii) **Search at least one confirmation** - To confirm the correct structure, the CAS Number of each chemical is searched in SciFinder[37]. If the software presents a structure of a molecule, the SMILES is run in OpenMolecule[38] and the two derived structures are compared. If the structures are identical, it is concluded a match. If SciFinder did not present any structure, the CAS Number is checked in ChemSpider[39]. If ChemSpider is able to present a structure, it is compared to the SMILES. The row in the Excel file will be marked with "1" if a match is found, and that compound will be maintained in the final output file.

(iii) **Search at least two confirmations** - The procedure is the same as in 'one confirmation' (ii); however, both SciFinder and ChemSpider need to present the same structure. The row is marked "2" for two verifications, and for maintaining in the final output. If only one verification is made, the row is marked with "1" and that compound will be rejected from the final output file.

In addition to the procedure described above, we developed following rules for the manual check;

- If no specific structure can be established, due to a substance containing a complex mixture of different compounds (such as petroleum products with several hydrocarbons), the substance will not be included in the final output file
- Substances that consist of one or more molecules combined with a polymer will be excluded from the final output file.
- Complex large molecules, such as proteins or clays, will not be suitable for hazard screening and are excluded.
- Salts and mixtures with counter ions will be included. The molecule with the highest molecular weight will be kept.
- Unspecific compounds that could have different isomers will be included. The structure that the SMILES presents will be considered a representative of that substance. One example is dichlorobenzenes, where one representative of the various congeners will be included.
- If the name and the CAS belongs to two different compounds, it will be excluded as we cannot establish which compound is the correct one.

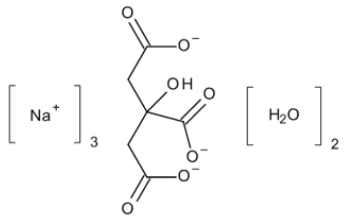
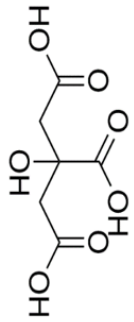
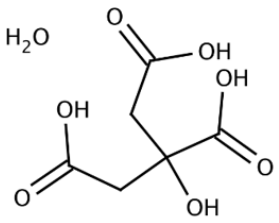
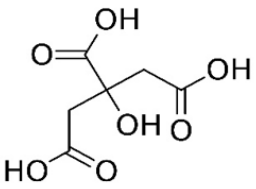
The SMILES of organic counter ions can be retrieved from the preliminary output file. If an organic counter ion is identified that substance will be marked with a warning "organic counter ion". The Excel file can be sorted after this warning and the substances can be checked manually to retrieve any additional SMILES of interest for hazard screening. To include these entries, they have to be added manually to the batch of SMILES after final output is retrieved.

After the manual check, the preliminary output file is saved, complemented with the number of web confirmations made, and reloaded into the workflow (see references in the Supplementary information for details and instructions on running the workflow). The final output file contains neutralised and standardised SMILES, ready for hazard screening.

3.3.4 A few examples on the effects of data curation

The curation process means that some structures are altered, e.g., when salts are curated, the counter ions are removed and the main molecule is kept neutral by addition of a proton. Data curation can be illustrated as in Table 2, where three different substances with different CAS Numbers ended up giving one SMILES; hence, one structure for hazard screening. These processes of data curation result in duplicates and triplicates in the final output file that has to be sorted out. Another illustration on potential effects of data curation is seen in Figure *4*, where the central atom is removed and the two fragments neutralised into a new molecule.

Table 2. *An example where three different entries gave one SMILES after data curation.*

CAS Number	Structure	SMILES after neutralisation	Structure for hazard screening
5949-29-1		OC(=O)CC(O)(CC(O)=O)C(O)=O	
6132-04-3			
77-92-9			

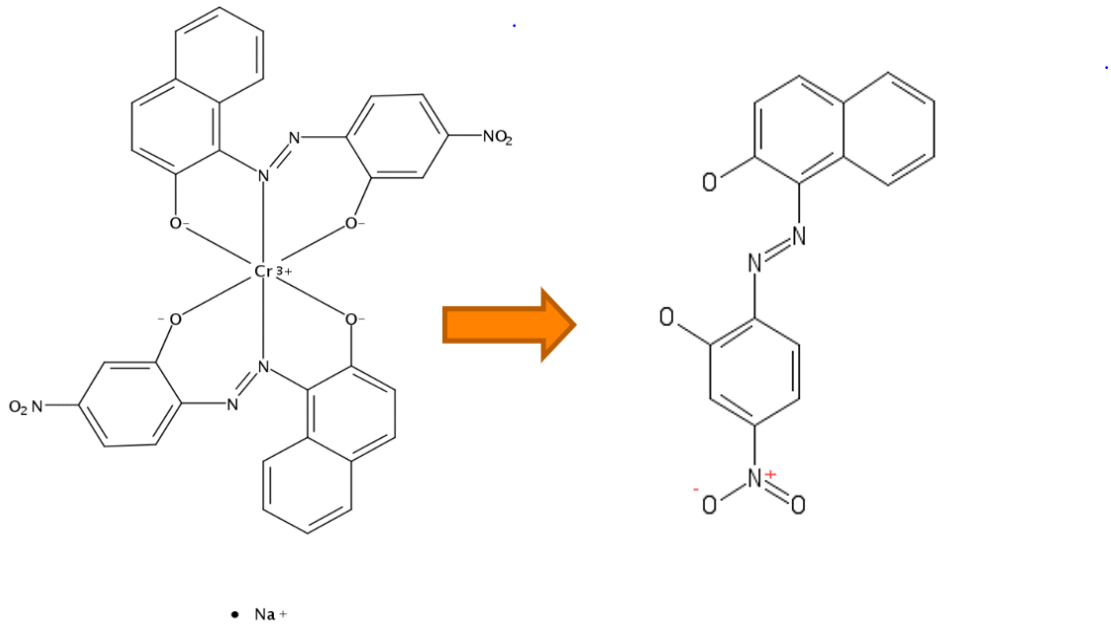


Figure 4. The effects of data curation of CAS 64611-73-0. The center atom is removed from the original molecule in the data curation step, yielding two fractions of the original molecule, which were merged into a new (neutral) molecule.

3.4 Hazard screening for potential NERCs

Quantitative structure–activity relationship models (QSARs) are developed to predict certain physicochemical, biological or environmental properties (endpoints) of a compound, based on chemical structure information. The accuracy of the different QSAR models is heavily dependent on the input data, and the initial step of data curation plays an important role in the final result of the predictions. Structural errors of the SMILES would give false predictions by the models[7]. EPI Suite[40] and VEGA[41] were chosen as QSAR platforms for hazard screening, and we focused mainly on persistence (P), bioaccumulation (B) and mobility (M). It was, however, also within the scope of this project to discuss modelled toxicity (T) predictions including carcinogenicity (C), mutagenicity (M) and reproductive toxicity (R). To assess the different hazard endpoints of the substances, several different hazard screening models were chosen and combined.

Persistence - For the persistence prediction, two different models were used: IRFMN (VEGA) and BIOWIN-3 (EPI Suite). The applied BIOWIN-3 (BIOWIN ultimate) model is a regression-based model with underlying expert-based judgements on complete biodegradation, whilst the IRFMN model classifies the compounds as; not persistent (nP), persistent (P) and very persistent (vP). Threshold values that were applied in scoring the persistence of chemicals were taken from REACH; a substance is persistent (P) if the half-time in seawater is longer than 60 days, or 40 days in fresh water (Table 3). The values for complete biodegradation are converted to half-lives in terms of days using EPI Suites conversion factors.

Bioaccumulation - The bioconcentration factor (BCF) describes the partitioning of the compound between biota and water. BCF was calculated in this study using three different models; Caesar, Arnot-Gobas (VEGA) and BCFBAF (EPI Suite). According to REACH, the criteria for bioaccumulation (B) is fulfilled at values > 2000 L/kg in aquatic species (Table 3). The evaluation of B can be expanded with the bioaccumulation factor (BAF), which takes

into account both exposure from the water phase and dietary exposure[42]. Such estimations have a particular impact on highly hydrophobic substances. However, BAF estimations are outside the scope of this study although models are available in EPI Suite. Note that models for BAF and BCF are known to be unsuitable for e.g. substances with a very high log K_{ow} (> 9) as well as poly- and perfluorinated substances, pigments and dyes[40].

Mobility - The mobility of a substance can be reflected using the organic carbon-water partition coefficient (K_{oc}). A high value of K_{oc} indicates that the substance is readily adsorbed to organic matter and has a low mobility[43,44]. For K_{oc} estimations, we used three different models. Two models from EPI Suite (one that uses molecular connectivity index (MCI) and one that uses log K_{ow}) and one model from Opera in VEGA. This information was combined with two models estimating water solubility (IRFMN from VEGA and Wskow in EPI Suite). As suggested by the German Environment Agency (UBA), mobile chemicals have high water solubility (≥ 0.15 mg/L) and low tendency to adsorb to solids[44].

Toxicity – Toxicity can be expressed in multiple ways using data from a large range of bioassays from different species and biological complexity. ECHA applies models for identification of PBT substances[45], and these were implemented for the positive controls. The toxicity criterion used in this study was NOEC values < 0.01 mg/L, and potential T by LC50 values < 0.1 mg/L. To analyse the positive controls, we used the following models included in the VEGA QSAR platform as a first attempt to screen for toxicity:

- LC50 estimations on crustacean and fish (Daphnia magna 48 h, Fathead minnow 96 h-EPA)
- NOEC fish (IRFMN)
- Developmental toxicity (Caesar)
- Carcinogenicity (Opera)
- Endocrine disruption (ER binding affinity model - IRFMN)
- Skin sensitisation (Caesar).

3.4.1 Scoring for hazard estimations

For hazard scoring, two different threshold limits were applied. In a first approach, threshold values established under REACH[46] were used ("Alarming" and "Probable" in Table 3), while a second approach applied more conservative limits, which were set below the established levels for each endpoint for identification of potential NERCs ("Possible" in Table 3). For P and B, two and three models were used, respectively, and an overall score was derived with combined results (Table S1 in the Supplementary information). Top scored chemicals for these measures indicate that several (or all) models predict P and B properties. The integration of several models to raise the quality of the predictions has been concluded successful by earlier studies[47,48]. For identification of NERCs to be analysed in the occurrence part of the study, a scoring system was developed using a conservative approach focused on only P and B. As an example, a score of 1 was given if the chemical qualifies for either P or B under the threshold for the hazard mark "Possible", and for being selected for occurrence study, a minimum score of 2 (minimum P+B = 1 + 1) was required.

Table 3. Threshold values for scoring chemicals, note that the hazard mark "Possible" was used as threshold for the conservative approach, and "Alarming" and "Probable" correspond to REACH limits.

Hazard mark	Persistence	Bioaccumulation	Mobility		Scoring
	Water (half-lives, days)	BCF (L/kg)	Log K_{oc}	S_w (mg/L)	
Alarming	160	> 5000	< 3	≥ 0.15	3
Probable	60	2000 - 5000	< 4	≥ 0.15	2
Possible	32.5	500 - 2000	> 4 - < 6	≥ 0.15	1
Unlikely	< 32.5	< 500	> 6	≥ 0.15	0
Top score	6	9	9		

3.5 Occurrence of identified potential NERCs

The occurrence of identified NERCs can be investigated in various environmental and human matrices, as well as in products and food items. The process for literature and data-base search used in this project is schematically described in Figure 5. Lists of potential NERCs created through the workflow described in Sections 3.3 and 3.4 were compared with compound databases summarizing earlier suspect and non-target screening efforts as well as with the open literature on target data.

For illustrative purposes on how such an occurrence study could be performed, NERCs derived from the datasets on paper and paperboard, as well as plastic additives were investigated in this report. The literature search focused on the compounds found to be both persistent and bioaccumulative, i.e., compounds that received a score ≥ 1 for both persistence and bioaccumulation during the hazard screening modelling (as listed in Table 3). A detailed summary of the created lists and their associated literature references is given in the Supplementary information.

3.5.1 Substances subject to SVHC assessment under ECHA

Prior to comparing the created lists of potential NERCs with available literature and data base records, compound lists were compared with ongoing substance-specific activities regarding REACH and the Classification, Labelling and Packaging (CLP) Regulation. The purpose of this step is to identify well-known chemical hazards in the created list of potential NERCs. It can be argued that already known hazardous chemicals do not match the definition of NERCs specifically and should therefore not be investigated further in the current report.

The following compound databases provided by the ECHA were taken into consideration:

- The public activities coordination tool (PACT)[15]
- Substances subject to POPs regulation[16]
- The candidate list of substances of very high concern (SVHC) for authorisation[17]

PACT represents and includes the other databases, and additionally contains information on dossier evaluation, substance evaluation, endocrine disruptor assessment, PBT assessment, harmonised classification and labelling, SVHC assessment, and current restrictions of the different compounds. Further comparisons of compound lists created in the current project and lists summarising ongoing substance-specific activities were therefore proceeded using PACT.

Created NERC lists were filtered to only contain compounds that are not subject to the SVHC assessment conducted by authorities under ECHA. In order to create lists of compounds that are not yet under regulatory consideration, compounds that are part of the SVHC assessment were not investigated regarding chemical occurrence during this specific study. Another possible approach could have been to investigate if the compounds are included in other regulatory assessments like the harmonised classification and labelling (CLH).

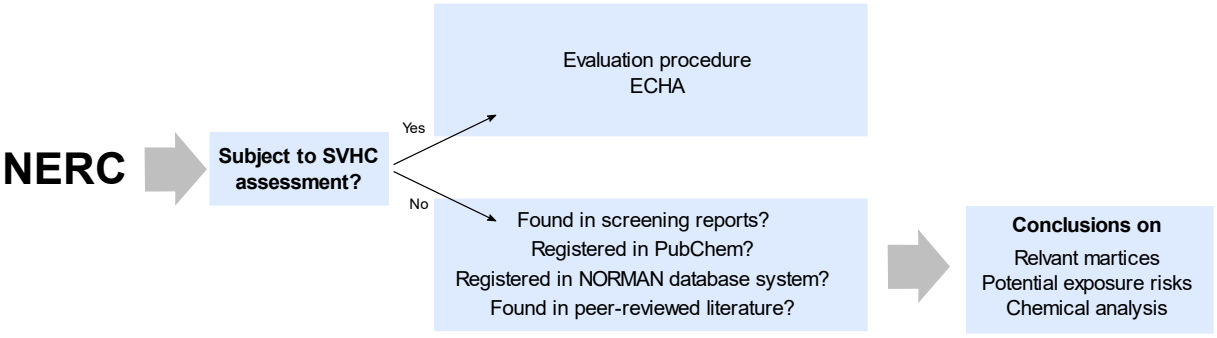


Figure 5. General workflow for retrieving data on the occurrence of compounds identified as potential NERCs in this study.

3.5.2 Reports and peer-reviewed literature

In the search of occurrence data for potential NERCs, the following literature sources were consulted:

- Datasets published in reports (number of available reports)
  - Nordic screening[48] (14)
  - Aarhus University[49], Denmark (54000)
  - Miljødirektoratet[50], Norway (7300)
  - Svenska Miljöinstitutet (IVL)[51], Sweden (2700)
  - Finnish Environment Institute (SYKE)[52], Finland (540)
- Access to peer-reviewed publications:
  - Google scholar[53]
  - Web of Science[54]

Relevant reports from report databases were downloaded for faster access. The databases were filtered according to the keywords depicted in Figure 6. A total of 208 reports were identified as relevant, and a detailed reference list can be found in the Supplementary information. Google Scholar and Web of Science were consulted for the compounds included in the investigation of chemical occurrence.

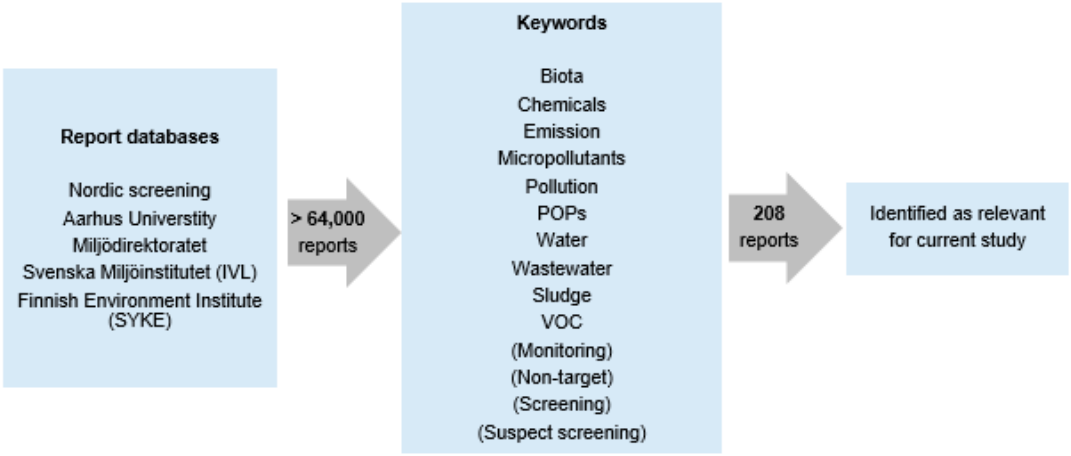


Figure 6. Description of the workflow for the identification of relevant screening reports from the selected report databases. Keywords in brackets could only be used to some extent due to extensive use of these keywords in research fields other than analytical chemistry.

3.5.3 Chemical databases

Chemical databases contain a large variety of information. Data on the use and manufacture of compounds, as well as literature references can help the investigation of chemical occurrence and analysis. Chemical databases chosen for retrieving relevant data on chemical occurrence and analysis in this project were PubChem[55] and the NORMAN database system[56]. PubChem is operated by the National Library of Medicine (National Center for Biotechnology Information) and contains information on > 1,000,000 chemical structures. Along with information on compound name, structure and identifiers, it summarises physicochemical properties, safety and hazards, toxicity, pharmacology and biochemistry, biological test results, spectral information, use and manufacturing, chemical vendors and existing literature. For the current project, information on existing literature and use/manufacturing patterns were found most useful, while other information collected in PubChem might be useful for future work on NERC candidates.

The NORMAN network[57] consists of researchers and stakeholders dealing with emerging substances. Numerous possibilities exist to share and access data on chemical compounds, including the extensive database system comprising information on chemical occurrence, bioassay testing, passive sampling, ecotoxicity, links to a suspect list exchange platform, MassBank Europe, substance factsheets and a digital sample freezing platform. For the current project, information on chemical occurrence was deemed most useful, while future investigations of identified NERCs might want to make use of other parts of the database system.

4 Results

4.1 Data curation

4.1.1 KNIME output from the different datasets

A summary of the results from the data curation process is given in Table 4. The table shows that the number of entries in the final output is substantially lower than the original number of entries (35-80%), illustrating the high number of substances not suitable for hazard screening within these datasets. The time spent on data curation for the four different datasets is related to the amount of input data, and whether verification in the two databases PubChem and ChemIDplus was necessary. The most time-consuming dataset was the PFAS list from OECD including 4730 entries. The list of everyday products consisted of ambiguous data with Danish spelling of the chemical name and erroneous CAS Numbers, causing a large number of substances to end up under manual check (26%). A detailed analysis on entries classified as maintain, reject, and manual check is given below (Section 4.1.3).

Table 4. Summary of the results from the data curation of the four selected datasets used within this study.

	Paper and paperboard	Plastic additives	PFAS	Everyday products
Input	1904	2663	4730	1773
Manual check	281	127	159	465
Reject	822	1644	1169	336
Duplicates	189	132	367	91
Maintain (percentage)	**893 (52%)**	**887 (35%)**	**3194 (73%)**	**1346 (80%)**
Time for workflow (h)	5.5	5	28	3.5
Time for manual check (h)	4	2	3	6

4.1.2 Evaluation of data curation

To evaluate the performance of the KNIME workflow used within this study, a number of studies were performed including comparisons of the analysis of the plastic additives dataset by Woodcock[18], the PFAS dataset by Chelsea et al[19], and the outcome of the curation of the positive controls. The study by Woodcock[18] was done without a systematic data curation. In that study, 1747 unique chemical structures were identified among the plastic additives. A manual rejection of 229 identified polymers and inorganics was performed, and EPI Suite was used indicating that 171 substances are both P and B. An assessment for T was carried out and in summary 24 substances were concluded PBT.

Among the 24 substances concluded as PBT in the study by Woodcock, seven were removed in this study in the data curation step (Table 5) as they lacked structures suitable for hazard screening. One of the substances (CAS 27253-29-8) ended up in the preliminary output file for manual inspection, but it is an unspecific mixture of carboxylic acids and incompletely defined, and was therefore excluded. Another example was CAS 147-14-8, which is a large pigment with a central copper atom and was rejected from the final output for several reasons. The overall result shows that six of the seven rejected compounds all have unspecified structures according to SciFinder[37] and have been correctly removed. The results illustrate the importance of data curation to obtain reliable SMILES suited for hazard screening as opposed to relying on datasets with no data curation, as was done in the study by Woodcock (2018)[18].

Table 5. Substances removed with the data curation methodology developed in this project that originally were included and defined as PBT compounds in the study by Woodcock[18].

CAS	Name/definition	Rejected as
64741-96-4	Distillates (petroleum), solvent-refined heavy naphthenic	Missing/ambiguous
64742-56-9	Distillates (petroleum), solvent-dewaxed light paraffinic	Missing/ambiguous
64742-65-0	Distillates (petroleum), solvent-dewaxed heavy paraffinic	Missing/ambiguous
72623-87-1	Lubricating oils (petroleum), C20-50, hydrotreated neutral oil-based	Missing/ambiguous
147-14-8	Copper, [29H,31H-phthalocyaninato(2-)-.kappa.N29,.kappa.N30,.kappa.N31,.kappa.N32]-, (SP-4-1)-	Organometallic, Unusual elements, Missing/ambiguous
27253-29-8	Neodecanoic acid, zinc salt (2:1)	Manual check
9016-87-9	Formaldehyde, oligomeric reaction products with aniline and phosgene	Missing/ambiguous

Data curation of the OECD list using the KNIME tool was recently published by Chelsea et al[19]. They extracted in total 3636 structures in comparison with the 3194 outputs that were extracted in present study. The former study only collected SMILES from one source and then estimated an error rate of 0.5% after a manual check of a random subset. One major difference was the number of excluded unusual elements; 132 (current study) versus 31 (Chelsea et al). It can be concluded that the generic model presented here is slightly more conservative and also less time consuming (Chelsea et al included manual checks of several databases). The data curation step was also examined using the ten positive controls, which showed that all chemicals passed the curation and were defined as 'maintain' with the exception of bisphenol A. BPA was rejected and classified as missing/ambiguous if entered as BPA.

4.1.3 Analysis of maintained, rejected and missing/ambiguous

Maintained. From the final output of the paper and paperboard chemicals and plastic additives, 100 randomly selected substances were manually controlled to identify any potential errors. The analysis revealed that 96 of the chemicals from the paper and paperboard list were correctly assigned, including seven mixtures, where the largest molecule in the respective mixture was kept (

Figure *7*). The wrongly maintained entries included two polymers, one substance with an erroneous CAS Registry Number and one organometallic compound. The random test of plastic additives also resulted in 96 correct structures, including one mixture. The wrongly maintained included one polymer and three undefined mixtures (fatty acids and benzenes). Overall, this analysis showed that the workflow has good capacity to identify correct structures.

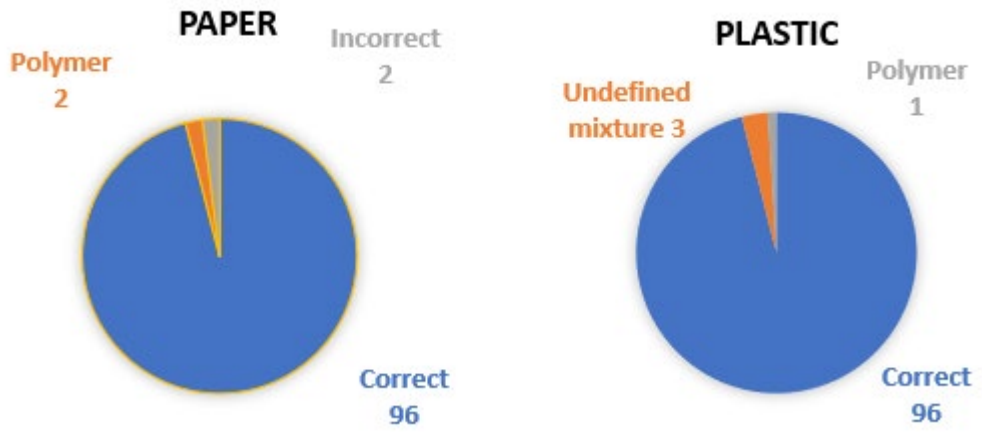


Figure 7. Distribution of a random subset of 100 substances from the 'paper and paperboard' (left) and 'plastic additives' (right) databased redefined as 'maintained' by the workflow.

Rejected. A large number of entries were rejected for various reasons by the applied curation methodology (Figure 8). It is important to note that the workflow may reject the same substance for several reasons. For example, a substance may be both a reaction product and have a missing/ambiguous SMILES. In the table below, the substances are classified only according to one rejection cause, which was the first one presented in the rejection file. Overall, the share of rejected chemicals varied widely from only 20% (everyday products) to 65% (plastic additives). A major reason for the low numbers of maintained entries from the plastic additives was the number of polymers (n=637) present in the dataset, which represents 24% of the total dataset. In each of the datasets, the number of SMILES that could not be found or that were ambiguous, i.e. substances without clear structure such as complex mixtures or substances where CAS Number and chemical name gave different structures, was rather large varying from 32 to 56% (see below for further analysis of these). The other indicators represent chemical characteristics that are unsuitable for hazard screening, and these have been correctly excluded from the final output (Figure 8). The category "manual reject" represents those substances assorted for manual check where no SMILES could be established, in accordance with the developed rules mentioned above. In this category several additional polymers were found. Of the 80 compounds under this post from paper and paperboard, 43 were identified as polymers by "poly" in the chemical name.

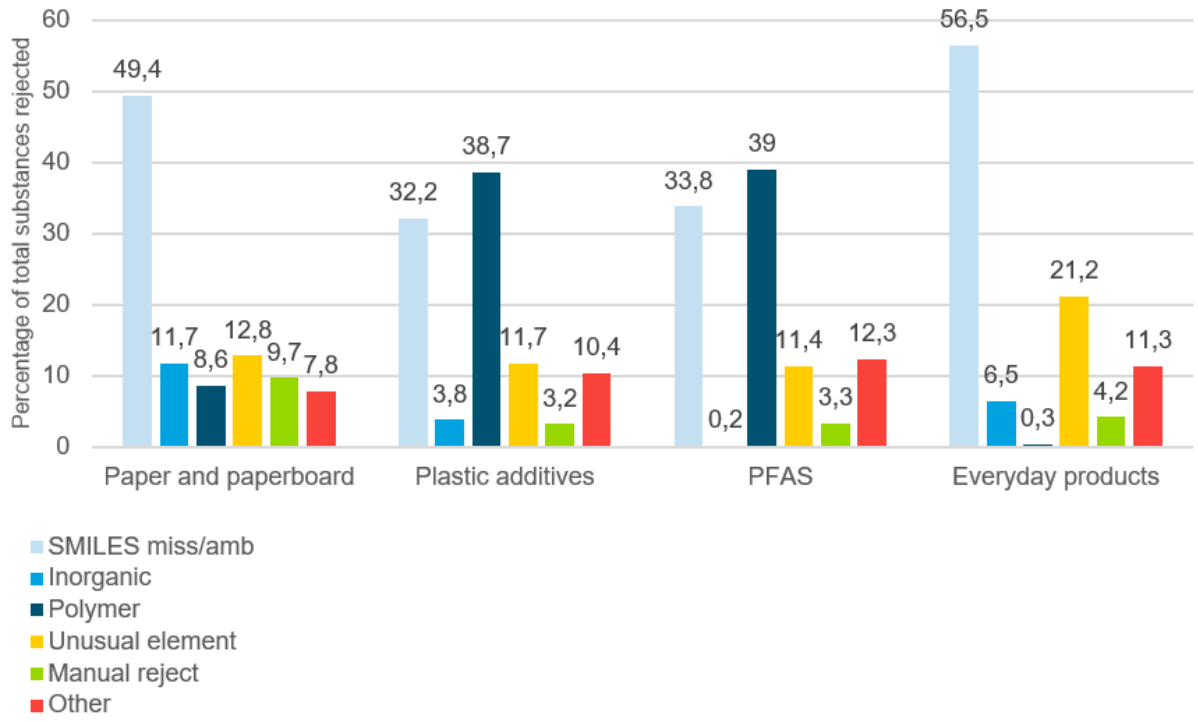


Figure 8. The distribution of rejection causes for each of the different data bases within this study. Organometallics, reaction products, mixtures and entries with unknown CAS Numbers are represented as "other".

Missing/ambiguous. To make an assessment on the false negative results, an analysis of substances removed as missing/ambiguous was carried out. A randomised test including 67 entries was performed on paper and paperboard chemicals classified as of high relevance. The analysis (using SciFinder) showed that the substances assorted as missing/ambiguous included complex combinations of hydrocarbons as well as polymers, mixtures, proteins and pigments. To get a general idea of what the mixtures consisted of, some of them were analysed in more detail and identified as asphalt, glass, waxes, oils and mixtures with cellulose and starch. Only one of the investigated substances, a pigment, was found to have a defined structure and SMILES. A similar analysis was performed on 75 substances from the final output from plastic additives, which resulted in seven compounds having a structure in SciFinder. The other substances were polymers, fatty acids, petroleum products, alcohols, waxes, derivatives, clays and substances where the CAS Number and chemical name gave two different structures and these were correctly assigned as missing/ambiguous. The conclusion of the random tests was that the program performed well in rejecting substances not suitable for hazard screening, with a false negative result of 1.5 and 10%.

4.2 Estimation of environmental and human health effects

Data curation was followed by hazard profiling using tools previously applied in Zheng et al[14]. We focused on persistence (P), bioaccumulation (B) and mobility (M) for the curated dataset of paper and paperboard and plastic additives, but also present data from estimating toxicity related properties (T) for the positive controls.

4.2.1 Positive controls

The analysis of the positive controls showed that, as expected, all chemicals reached in at least two categories the highest set level ("Alarming" or "Probable", Table 3 and Table S1). Most of the chemicals were predicted as toxic according to their predicted LC50 values and as expected, PFOS scored high for mobility and DDT was defined as a carcinogen and a vPvB (very Persistent, very Bioaccumulative) substance. Overall, the results indicate that the used models for hazard screening were able to identify chemicals known to be hazardous.

4.2.2 Listing of potential NERCs

The SMILES from the curated dataset of paper and paperboard and plastic additives were run in three models for bioconcentration, two models for persistence, three for the organic carbon-water partition coefficient (K_{oc}) and two models for water solubility (the last two for assessing mobility). Five positive controls (those listed as PB chemicals by ECHA) were used as a verification of the method, and they were all found within the top scored as PB substances (Table S1 in the Supplementary information). All substances predicted by at least one of the models to have both P and B properties were considered potential NERCs, yielding 138 substances from the plastic additives inventory and 134 from the paper and paperboard list (Figure 10). These compounds were further evaluated in step three of the project (the occurrence study). The number of substances identified with the applied conservative approach were compared to the number of substances that qualified as P, PB and PM according to REACH threshold values (Figure 9). Clearly, the number of hits decreases dramatically by using the threshold values set under REACH; from 138 to 47 for plastic additives and from 134 to 27 for paper and paperboard chemicals. By applying vPvM or vPvB criteria, the numbers decreased even more. The substances that were estimated as vPvB or vPvM by at least one of the models (two models for mobility) are listed in Table 6 and Table 7.

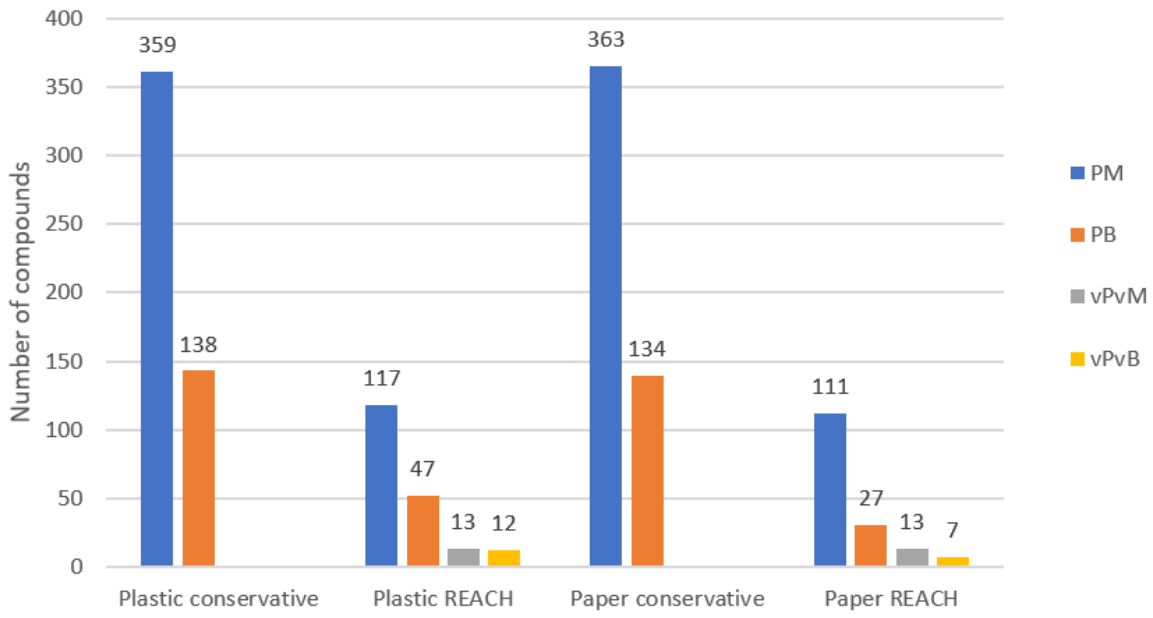


Figure 9. Distribution of chemicals defined as combinations of persistent (P), bioaccumulative (B), and mobile (M) according to the conservative settings and REACH, respectively.

Table 6. Substances ranked as vP by at least one of the models and vM by at least two of the models for each dataset.

CAS No	Name	Dataset
72496-88-9	Sodium bis[2-[[5-(aminosulphonyl)-2-hydroxyphenyl]azo]-3-oxo-N-phenylbutyramidato(2-)]cobaltate(1-)	Paper
26523-64-8	Trichlorotrifluoroethane	Paper
81-11-8	Diaminostilbenedisulphonicacid	Paper
86014-76-8	4-[2-(2,4-Dihydroxyphenyl)diazenyl]-5-hydroxy-2,7-naphthalenedisulfonic acid	Paper
108-78-1	Melamine	Paper
2706-28-7	Disodium 2-amino-5-[(4-sulphonatophenyl)azo]benzenesulphonate	Paper
375-73-5	Perfluorobutanesulfonic acid	Paper
15827-60-8	Diethylenetriaminepenta(methylenephosphonic acid)	Paper
52236-73-4	Lithium 4-[(5-amino-3-methyl-1-phenyl-1H-pyrazol-4-yl)azo]-2,5-dichlorobenzenesulphonate	Paper
6528-34-3	2-[(4-Methoxy-2-nitrophenyl)azo]-N-(2-methoxyphenyl)-3-oxobutyramide	Paper
6358-31-2	2-[(2-Methoxy-4-nitrophenyl)azo]-N-(2-methoxyphenyl)-3-oxobutyramide	Paper
6359-98-4	Disodium2,5-dichloro-4-(5-hydroxy-3-methyl-4-(sulphophenylazo)pyrazol-1-yl)benzenesulphonate	Paper
1052-38-6	4,4'-[1,3-Phenylenebis(azo)]bisbenzene-1,3-diamine	Paper
71832-85-4	Benzenesulfonic acid, 4-[2-[1-[[(2-chlorophenyl)amino]carbonyl]-2-oxopropyl]diazenyl]-3-nitro-, calcium salt (2:1)	Plastic
12286-66-7	Benzenesulfonic acid, 4-[2-[1-[[(2-methylphenyl)amino]carbonyl]-2-oxopropyl]diazenyl]-3-nitro-, calcium salt (2:1)	Plastic
80-51-3	Benzenesulfonic acid, 4,4'-oxybis-, 1,1'-dihydrazide	Plastic
108-78-1	1,3,5-Triazine-2,4,6-triamine	Plastic
115-27-5	4,7-Methanoisobenzofuran-1,3-dione, 4,5,6,7,8,8-hexachloro-3a,4,7,7a-tetrahydro-	Plastic
33329-35-0	1,3-Propanediamine, N1,N1-bis[3-(dimethylamino)propyl]-N3,N3-dimethyl-	Plastic
6358-31-2	Butanamide, 2-[2-(2-methoxy-4-nitrophenyl)diazenyl]-N-(2-methoxyphenyl)-3-oxo-	Plastic
15875-13-5	1,3,5-Triazine-1,3,5(2H,4H,6H)-tripropanamine, N1,N1,N3,N3,N5,N5-hexamethyl-	Plastic
74441-05-7	Benzamide, N-[4-(aminocarbonyl)phenyl]-4-[2-[1-[[(2,3-dihydro-2-oxo-1H-benzimidazol-5-yl)amino]carbonyl]-2-oxopropyl]diazenyl]-	Plastic
34454-97-2	1-Butanesulfonamide, 1,1,2,2,3,3,4,4,4-nonafluoro-N-(2-hydroxyethyl)-N-methyl-	Plastic
65212-77-3	Benzenesulfonic acid, 4,5-dichloro-2-[2-[4,5-dihydro-3-methyl-5-oxo-1-(3-sulfophenyl)-1H-pyrazol-4-yl]diazenyl]-, calcium salt (1:1)	Plastic
28768-32-3	2-Oxiranemethanamine, N,N'-(methylenedi-4,1-phenylene)bis[N-(2-oxiranylmethyl)-	Plastic
64265-57-2	1-Aziridinepropanoic acid, 2-methyl-, 1,1'-[2-ethyl-2-[[3-(2-methyl-1-aziridinyl)-1-oxopropoxy]methyl]-1,3-propanediyl] ester	Plastic

Table 7. Substances ranked as vPvB by at least one of the models for each dataset.

CAS No	Name	Dataset
596-84-9	Manoyloxide	Paper
5102-83-0	2,2'-[(3,3'-Dichloro[1,1'-biphenyl]-4,4'-diyl)bis(azo)]bis[N-(2,4-dimethylphenyl)-3-oxobutyramide]	Paper
375-95-1	Perfluorononanoic acid	Paper
1763-23-1	Perfluorooctane-sulphonic acid	Paper
335-67-1	Perfluorooctanoic acid	Paper
375-85-9	Perfluoroheptanoic acid	Paper
72479-28-8	Sodium 4-chloro-3-[4-[[5-chloro-2-(2-chlorophenoxy)phenyl]azo]-4,5-dihydro-3-methyl-5-oxo-1H-pyrazol-1-yl]benzenesulphonate	Paper
133-14-2	Peroxide, bis(2,4-dichlorobenzoyl)	Plastic
64338-16-5	7-Oxa-3,20-diazadispiro[5.1.11.2]heneicosan-21-one, 2,2,4,4-tetramethyl-	Plastic
2212-81-9	Peroxide, [1,3-phenylenebis(1-methylethylidene)]bis[(1,1-dimethylethyl)	Plastic
3864-99-1	Phenol, 2-(5-chloro-2H-benzotriazol-2-yl)-4,6-bis(1,1-dimethylethyl)-	Plastic
36437-37-3	Phenol, 2-(2H-benzotriazol-2-yl)-4-(1,1-dimethylethyl)-6-(1-methylpropyl)-	Plastic
78-63-7	Peroxide, 1,1'-(1,1,4,4-tetramethyl-1,4-butanediyl)bis[2-(1,1-dimethylethyl)	Plastic
6731-36-8	Peroxide, 1,1'-(3,3,5-trimethylcyclohexylidene)bis[2-(1,1-dimethylethyl)	Plastic
61260-55-7	1,6-Hexanediamine, N1,N6-bis(2,2,6,6-tetramethyl-4-piperidinyl)	Plastic
79-94-7	Phenol, 4,4'-(1-methylethylidene)bis[2,6-dibromo-	Plastic
509-34-2	Spiro[isobenzofuran-1(3h),9'-[9h]xanthen]-3-one,3',6'-bis(diethylamino)-	Plastic
4948-15-6	Anthra[2,1,9-def:6,5,10-d'e'f']diisoquinoline-1,3,8,10(2H,9H)-tetrone, 2,9-bis(3,5-dimethylphenyl)-	Plastic
83524-75-8	Anthra[2,1,9-def:6,5,10-d'e'f']diisoquinoline-1,3,8,10(2H,9H)-tetrone, 2,9-bis[(4-methoxyphenyl)methyl]-	Plastic

4.3 Occurrence of identified NERC in various matrices

The occurrence of NERCs derived through the hazard screening from the curated datasets on paper and paperboard, as well as plastic additives was investigated. The literature search focused on the compounds found to be both persistent and bioaccumulative, i.e., compounds that received a score ≥ 1 during the hazard screening modelling for both P and B as listed in Table 3, composing 139 (paper and paperboard) and 143 (plastic additives) compounds. The derived lists were further filtered to only include compounds not included in a SVHC assessment by ECHA15 (Figure 10). A summary including substance-specific details on where the individual compounds have been identified (matrix), instrumental analysis as well as a complete list of the respective literature references (screening reports and peer-reviewed literature) is given in the Supplementary information.

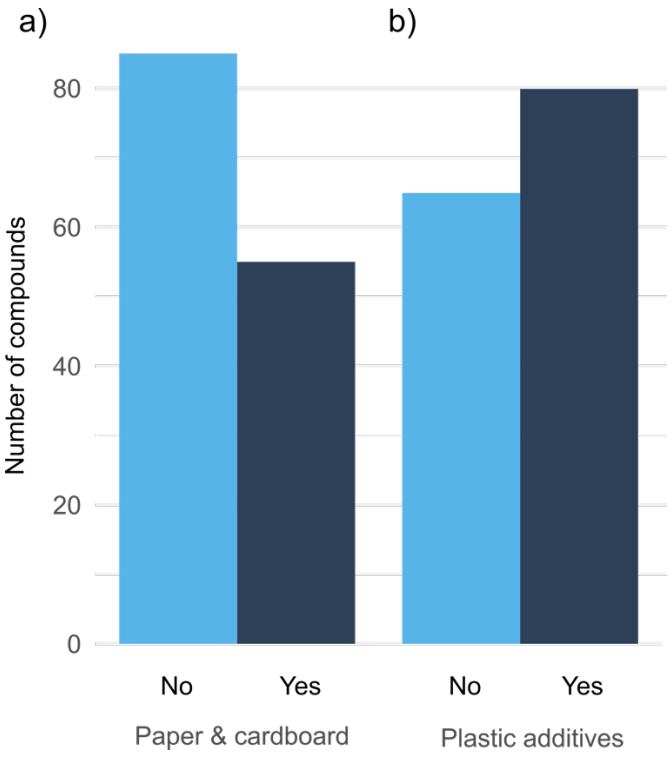


Figure 10. Number of compounds included ("Yes", dark blue) and not included ("No", lighter blue) in SVHC assessment by ECHA among those identified as NERCs (persistent and bioacculumative (PB)) in the (a) paper and paperboard dataset and (b) plastic additives dataset.

4.3.1 Paper and paperboard

For paper and paperboard, 55 of the 139 prioritised compounds were covered in a SVHC assessment by ECHA leaving 84 compounds for further investigation on occurrence ("No" in Figure 10a). For the majority of compounds included in the occurrence analysis, relevant literature on chemical occurrence was found. All 84 compounds were listed in PubChem, and 16 compounds were found in at least one of the selected screening reports. A total of 64 compounds could be found in either reports and/or peer-reviewed literature, whereas 29 compounds could not be found in any literature investigating chemical occurrence.

Conclusions on the analysability of the compounds can be gained by investigating the applied analytical methods reported in the literature. As for instrumental analysis, it was found that 43 of the 84 compounds were reported as analysed with gas chromatography coupled to mass spectrometry (GC-MS). 7 could be reported as analysed via liquid chromatography coupled to mass spectrometry (LC-MS) and 5 compounds were analysed by other methods (e.g. spectrophotometric) leaving 30 compounds with no data on instrumental analysis.

For the identified NERCs from the paper and paperboard dataset, it was found that most compounds were detected in food (12 compounds) and oils and fats (12), followed by biota (11), surface water (11) and sediment (10) (Figure 12). Many compounds were also found in one-time-use food packaging in the form of paper wrappings (8) as well as in polymeric consumer products (5), including plastic containers for food. Note, that reusable plastic containers for food were counted as consumer products in polymeric materials, while the category "food packaging" summarises single-use food packaging such as disposable paper wrappings. Many compounds were reported in several matrices, which is why the total number of compounds in Figure 11 and Figure 12 is larger than the number of potential NERCs investigated. For more information on the individual compounds' occurrence, sample preparation methods, instrumental analysis, and literature references, see Supplementary information.

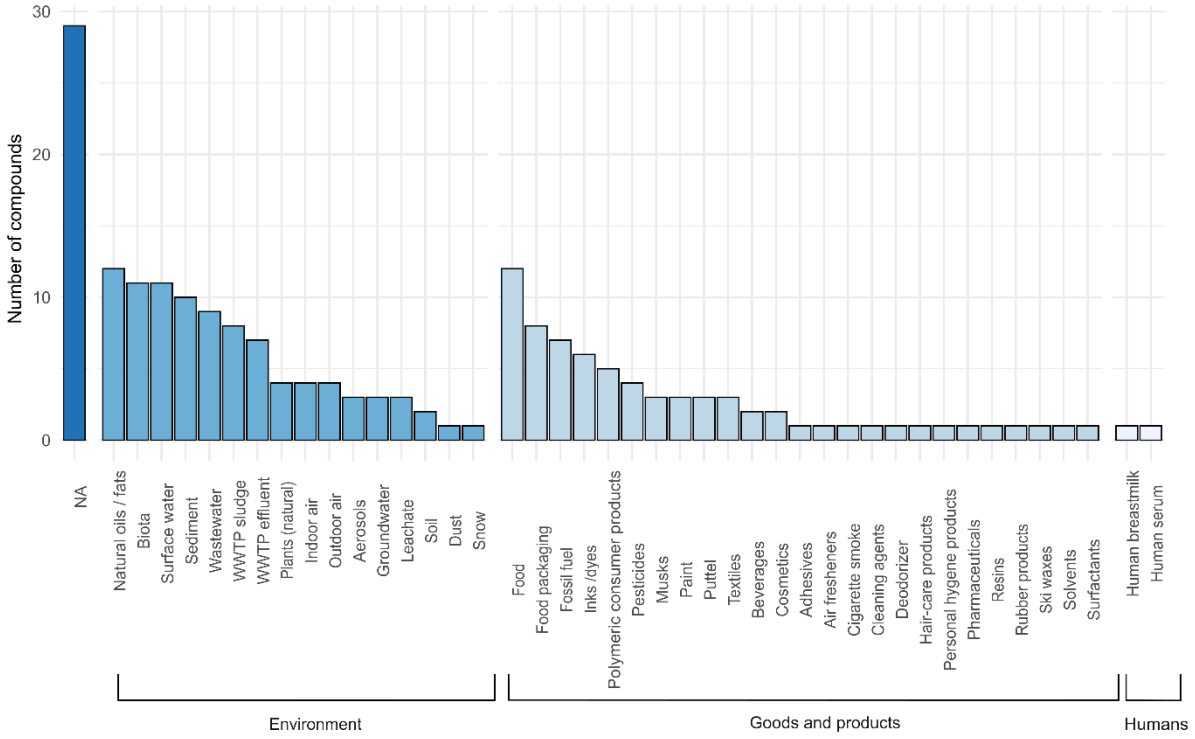


Figure 11. Occurrence of identified NERCs from the paper and paperboard dataset not included in a SVHC assessment by ECHA in various matrices. Note that compounds often were reported in several matrices and the total "number of compounds" is larger than the number of compounds investigated (84).

4.3.2 Plastic additives

For the plastic additives, 80 of the prioritised compounds were covered in a SVHC assessment by ECHA leaving 63 compounds for further investigation on occurrence (Figure 10b). Among these, 61 were listed in PubChem, 2 compounds were found in at least one of the selected screening reports, and 1 compound was analysed but not found in one report. In total, 45 compounds were found in either the selected screening reports and/or peer-reviewed literature leaving 12 compounds that could not be found in any literature investigating chemical occurrence. 8 of the compounds were reported as analysed with GC-MS, 5 were reported as analysed via LC-MS methods. 5 compounds were analysed by other methods (e.g. spectrophotometric), and for 45 of the 63 investigated compounds no data on analysis could be found.

A compilation of the number of compounds found in various matrices can be seen in Figure 12. The identified NERCs from the plastic additives dataset were mainly found in consumer products in polymeric materials, such as reusable food containers, toys or housings of electronic devices (26 compounds) followed by inks and dyes (17), textiles (9), and food (6). The large total number of compounds found in textiles, inks and dyes is related to the use of these compounds as colouring agents, which to a certain extent also goes for the occurrence of the same compounds in consumer products in polymeric materials. For more information on occurrence, sample preparation methods, instrumental analysis, and respective literature reference, see Supporting Information.

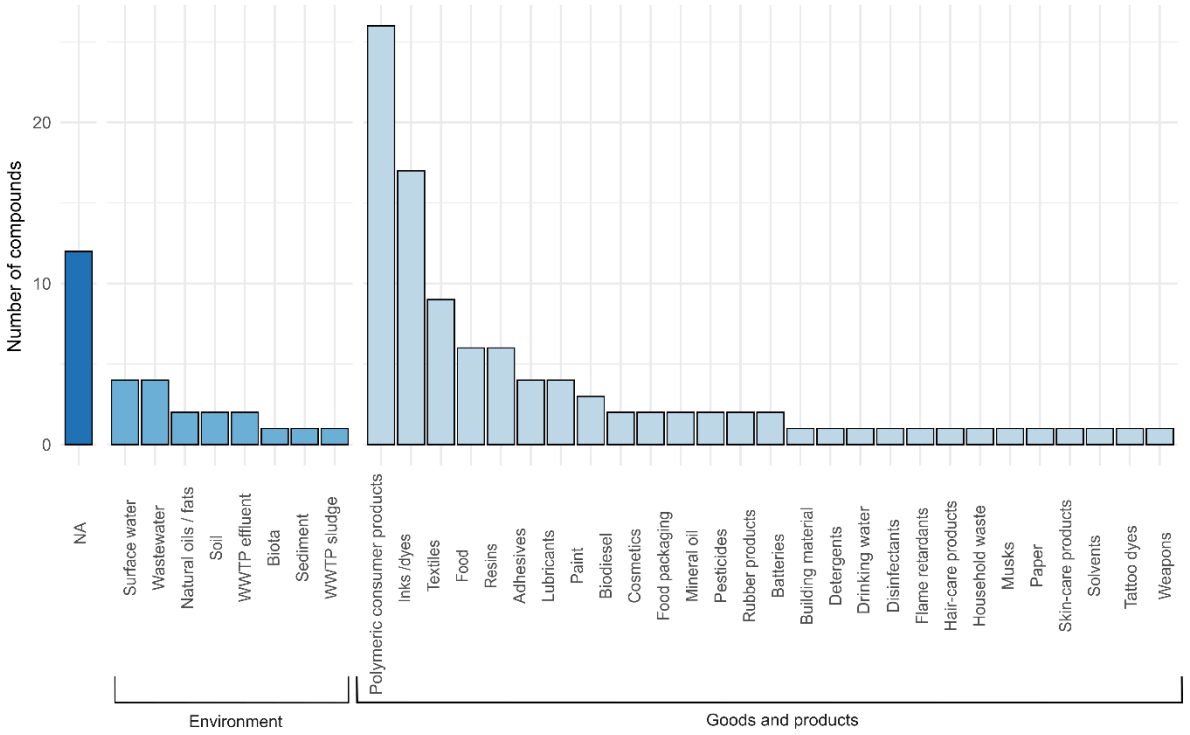


Figure 12. Occurrence of identified NERCs from the plastic additives dataset not included in a SVHC assessment by ECHA in various matrices. Note that compounds often were reported in several matrices and the total "number of compounds" is larger than the number of compounds investigated (63).

4.3.3 Overlap of different lists

One way to prioritise compounds for future screening studies is to compare the lists of potential NERCs derived through curation and hazard screening from various datasets and to prioritise compounds present in several of those lists. In the examples used in this report (PB compounds in paper and cardboard products and plastic additives), this overlap included 24 compounds (Table 8). Among these, 5 are not currently considered in a SVHC assessment under ECHA.

Table 8. Compounds present in both investigated lists of potential NERCs, i.e., compounds estimated to be persistent and bioaccumulative from the paper and cardboard and the plastic additives datasets. The column "SVHC" indicate if the compound currently is part of a SVHC assessment under ECHA.

CAS No	Names	SVHC
13475-82-6	2,2,4,6,6-Pentamethylheptane	Yes
597-82-0	Triphenylphosphorothioate,o,o,o-	Yes
111-65-9	Octane	Yes
142-82-5	Heptane	Yes
101-02-0	Phosphorous acid, triphenyl ester	Yes
84-61-7	Phthalic acid, dicyclohexyl ester	Yes
9007-13-0	Resin acids and rosin acids, calcium salts	Yes
84852-15-3	Phenol, 4-nonyl-, branched	Yes
128-37-0	2,6-Di-tert-butyl-p-cresol	Yes

75980-60-8	Diphenyl-(2,4,6-trimethylbenzoyl)phosphinoxide	Yes
6846-50-0	2,2,4-Trimethyl-1,3-pentanedioldiisobutyrate	Yes
116-37-0	Bisphenolabis(2-hydroxypropyl)ether	Yes
4098-71-9	Isophorondiisocyanat	Yes
162881-26-7	Phenyl bis(2,4,6-trimethylbenzoyl)-phosphine oxide	Yes
5873-54-1	2,4'-Diisocyanatodiphenylmethan	Yes
101-68-8	4,4-Methylenediphenyl diisocyanate	Yes
96-76-4	2,4-Bis(1,1-dimethylethyl)phenol	Yes
5567-15-7	2,2'-[(3,3'-Dichloro[1,1'-biphenyl]-4,4'-diyl)bis(azo)]bis[N-(4-chloro-2,5-dimethoxyphenyl)-3-oxobutyramide]	Yes
112-84-5	(Z)-Docos-13-enamide	Yes
6358-30-1	8,18-Dichloro-5,15-diethyl-5,15-dihydrodiindolo[3,2-b:3,2-m]triphenodioxazine	No
4531-49-1	2,2'-[(3,3'-Dichloro[1,1'-biphenyl]-4,4'-diyl)bis(azo)]bis[N-(2-methoxyphenyl)-3-oxobutyramide]	No
112-55-0	Dodecylmercaptan	No
26544-23-0	Phosphorous acid, isodecyl diphenyl ester	No
7128-64-5	2,5-Thiophenediylbis(5-tert-butyl-1,3-benzoxazole)	No

5 Discussion and Conclusions

A generic method for identification of potential NERCs was established and applied on a variety of different datasets, using an initial data curation workflow followed by hazard screening, where a combination of several computational tools and models were used. The generated lists of potential NERCs were then screened for occurrence in different matrices. Data curation was found to be a crucial step in the workflow to generate input data suitable for computational analysis, like hazard screening using QSAR models. This results in removing a vast part of the input data that today's tools are not compatible with, including polymers and inorganics. Substances that are removed could still be of potential hazard, and methodologies to analyse these are needed. Polymers will over time degrade to smaller units including oligomers and monomers that could pose a larger hazard. In addition, polymers typically include a minor share of monomers as impurities. Today's QSAR models used in the hazard screening are designed to merely make predictions on small neutral organic molecules. Further development is required to assess risks of inorganics, ionisable chemicals and larger molecules including polymers.

The developed and applied workflow is constructed for data curation based on data mining from several sources to raise the level of confidence in yielding correct structures, which is critical prior to hazard screening[6,7]. The workflow showed to yield fast curation of data and most parts are automatic. The conservative approach of searching several databases and comparing results is both the strength and weakness of the method. Structures are valid with a higher confidence, but might also be excluded incorrectly if the databases contains variation in structures for the same CAS[36]. Furthermore, it has been illustrated that there is a need for data curation to remove substances including undefined complex chemical mixtures that are incompatible with the hazard screening models18. Model developments are ongoing including the application of various machine learning methods that may increase the applicability domain of future hazard screening tools[59].

Several aspects of the workflow could be developed including for example the treatment of polymers as some were not excluded properly. Data could be searched automatically for identification of polymers and other unwanted structures by including keywords such as "poly", "acids", "petroleum", "deriv" and "alcohols". This would decrease the number of substances that are found under missing/ambiguous and could make that subset of data easier for manual control. Another area of improvement is the procedure to handle organic counter ions where both parts could be of interest for a hazard assessment where today only the largest part was kept. The workflow is heavily dependent on CAS Numbers, and chemical names are not enough for a successful data curation. An improvement would be to include an option to retrieve CAS from existing SMILES or SMILES generated from names. It would also be good to flag compounds presented with abbreviations such as BPA for bisphenol-A. The workflow failed to identify BPA if presented with the abbreviation and the compound was deleted (as studied databases returned different structures for the abbreviation).

An investigation of derived potential NERCs was completed using scientific literature, reports and regulatory databases to gain more information on individual compounds regarding chemical occurrence in various matrices and methodologies for chemical analysis. The PACT list provided by ECHA was consulted in order to determine which compounds were subject to a SVHC assessment. It was argued that substances investigated in a SVHC assessment could not be defined as NERCs, as these are already identified as potentially hazardous chemicals and undergoing detailed investigations. In addition to cross-comparisons with the PACT lists, this stage could also include comparisons with other databases, like the harmonised

classification system or similar that could be accomplished automatically if databases can be downloaded.

In contrast to most of the other steps in the described workflow, the investigation of current literature and screening databases as well as the search in non-peer-reviewed reports is a time-consuming and purely manual step. The download of relevant reports published on previous screening studies is a time-consuming first step and the completeness is highly dependent on the (online) availability of reports, the investigator's skill to detect relevance from reports titles or abstracts only, and the number and nature of databases consulted. In the current study, reports were drawn from report databases throughout the Scandinavian countries, leaving many large European and international databases untouched. Next to the factor of time, language is a limiting factor, as many reports are published in the language of the responsible authorities. The factor of language further hampers the search for compounds within the reports themselves and is therefore highly dependent on included CAS Numbers (as applied in this study) or other internationally comprehensible and unique compound identifiers.

The search for compounds in peer-reviewed literature is also challenging and subject to biased results. It was found that CAS Numbers are not commonly printed in scientific articles, even if studies focused on chemical screening. Even though most peer-reviewed literature is published in English, a compound's name can be expressed in various ways and the investigator has to know which of the compound's chemical names or trade names are most relevant to search for in the literature. Similar to the search for reports in report database, the search for literature in peer-reviewed article databases is dependent on the investigator's skill to detect relevant articles from the title or the abstract. Many articles purely concern synthesis and manufacturing of the compounds, and adding certain search words additionally to the compounds' names is necessary to limit the number of articles in the search output. Adding search words, like "environment", "water", "atmosphere" or "screening", rarely led to a relevant limitation of the output, as articles on synthesis of compounds make use of these words.

The occurrence study focused on derived lists of NERCs estimated to have persistent and bioaccumulative (PB) properties compiled from the datasets focusing on paper/cardboard materials and plastic additives. A summary of matrices in which NERCs previously have been found (as done in Figure 11 and Figure 12 in this report) can help decision makers to identify relevant matrices for screening studies. However, it should be noted that this kind of summary can give a skewed picture on the true occurrence of the NERCs. Many matrices such as surface water or wastewater are sampled and investigated in many studies, while other matrices, such as human urine or personal care products are covered in fewer studies. It is also possible that the most relevant matrices are not investigated yet for a certain set of compounds. Care should therefore be taken when drawing conclusions from the summary of matrices in which compounds previously have been identified. Expert judgement and strategies to gain information on migration patterns and analysability of the compounds (sample preparation, general instrumental analysis, limits of detection, etc.) should be applied in future studies in order to be able to give better recommendations for analysis of potential NERCs in different matrices. Machine learning techniques for automatic data retrieval from abstracts and full-texts should be investigated for widening the literature search in a systematic manner.

The challenge of inhomogeneous datasets will most likely persist in the future and data curation methods like the one presented are thus necessary. Data curation is crucial for obtaining lists of compounds suitable for the computational approaches applied for hazard screening. While the method presented in this report can be regarded as a fast and robust

procedure, further development is required for handling complex chemical mixtures, organic counterions and degradation products of studied chemicals. Additionally, computational models applied for hazard screening is currently not applicable for a large range of the entries in chemical inventories including polymers, ionisable compounds and inorganics. For the time-consuming stage of searching for occurrence data, more advanced data mining techniques could potentially help speeding this step up. However, potential lists of NERCs were derived applying developed workflow from various datasets. These lists of chemicals from the paper and plastics industry warrants further investigations and could provide insights in designing future screening and monitoring campaigns. Albeit discussed challenges, the developed workflow can be recommended as a prototype for an early warning system to identify potential NERCs.

6 References

1. Study for the strategy for a non-toxic environment of the 7th Environment Action Programme - Publications Office of the EU. Accessed December 1, 2020. https://op.europa.eu/sv/publication-detail/-/publication/89fbbb74-969c-11e7-b92d-01aa75ed71a1/language-en
2. Palmen N. New and Emerging Risks of Chemicals (NERCs) and the future of work. Accessed November 27, 2020. http://www.ilo.org/global/topics/safety-and-health-at-work/events-training/events-meetings/world-day-for-safety/33thinkpieces/WCMS_681596/lang--en/index.htm
3. Regulation (EC) No 1907/2006 - Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) - Arbetsmiljö – EU-Osha. Accessed November 27, 2020. https://osha.europa.eu/sv/legislation/directives/regulation-ec-no-1907-2006-of-the-european-parliament-and-of-the-council
4. Hogendoorn EA. Progress report on New or Emerging Risks of Chemicals (NERCs).
5. Wang Z, Walker GW, Muir DCG, 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. doi:10.1021/acs.est.9b06379
6. Fourches D, Muratov E, Tropsha A. Trust, but Verify II: A Practical Guide to Chemogenomics Data Curation. *J Chem Inf Model*. 2016;56(7):1243-1252. doi:10.1021/acs.jcim.6b00129
7. Fourches D, Muratov E, Tropsha A. Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J Chem Inf Model*. 2010;50(7):1189-1204. doi:10.1021/ci100176x
8. Muratov EN, Varlamova E V., Artemenko AG, Polishchuk PG, Kuz'Min VE. Existing and developing approaches for QSAR analysis of mixtures. *Mol Inform*. 2012;31(3-4):202-221. doi:10.1002/minf.201100129
9. QSAR-modeller - ECHA. Accessed November 27, 2020. https://echa.europa.eu/sv/support/registration/how-to-avoid-unnecessary-testing-on-animals/qsar-models
10. PM 4/19: Chemical Substances in Paper and Paperboard - Kemikalieinspektionen. Accessed December 3, 2020. https://www.kemi.se/publikationer/pm/2019/pm-4-19-chemical-substances-in-paper-and-paperboard
11. PM2/19: Survey on substances in plastics in the Swedish product register - Kemikalieinspektionen. Accessed December 3, 2020. https://www.kemi.se/publikationer/pm/2019/pm2-19-survey-on-substances-in-plastics-in-the-swedish-product-register
12. OECD Portal on Per and Poly Fluorinated Chemicals - OECD Portal on Per and Poly Fluorinated Chemicals. Accessed December 2, 2020. http://www.oecd.org/chemicalsafety/portal-perfluorinated-chemicals/
13. Kortlægning af forbrugerprodukter. Accessed December 3, 2020. https://mst.dk/kemi/kemikalier/forskning-og-kortlaegning/kortlaegning-af-forbrugerprodukter/

14. Zheng Z, Peters GM, Arp HPH, Andersson PL. Combining in Silico Tools with Multicriteria Analysis for Alternatives Assessment of Hazardous Chemicals: A Case Study of Decabromodiphenyl Ether Alternatives. *Environ Sci Technol*. 2019;53(11):6341-6351. doi:10.1021/acs.est.8b07163

15. Agency EC. The public activities coordination tool (PACT). *https://echa.europa.eu/de/pact*. Published online 2020.

16. Agency EC. List of substances subject to POPs Regulation. *https://echa.europa.eu/list-of-substances-subject-to-pops-regulation*. Published online 2020.

17. Agency EC. Candidate List of substances of very high concern for Authorisation. *https://echa.europa.eu/candidate-list-table*. Published online 2020.

18. Louise Woodcock. *Screening Av Ämnen Med Potentiella PBT/ VPvB- Egenskaper*.; 2018.

19. Chelcea IC, Ahrens L, Örn S, Mucs D, Andersson PL. Investigating the OECD database of per- and polyfluoroalkyl substances – chemical variation and applicability of current fate models. *Environ Chem*. 2020;17(7):498. doi:10.1071/EN19296

20. Triclosan - Substance Information - ECHA. Accessed December 7, 2020. https://echa-term.echa.europa.eu/et/web/guest/substance-information/-/substanceinfo/100.020.167

21. Heptadecafluorooctane-1-sulphonic acid - Substance Information - ECHA. Accessed December 7, 2020. https://echa.europa.eu/sv/substance-information/-/substanceinfo/100.015.618

22. Bis(2-ethylhexyl) phthalate - Substance Information - ECHA. Accessed December 7, 2020. https://echa.europa.eu/sv/substance-information/-/substanceinfo/100.003.829

23. Tris(2-chloroethyl) phosphate - Substance Information - ECHA. Accessed December 7, 2020. https://echa.europa.eu/sv/substance-information/-/substanceinfo/100.003.744

24. Clofenotane - Substance Information - ECHA. Accessed December 7, 2020. https://echa.europa.eu/sv/substance-information/-/substanceinfo/100.000.023

25. 1,2,5,6,9,10-hexabromocyclododecane - Substance Information - ECHA. Accessed December 7, 2020. https://echa.europa.eu/sv/substance-information/-/substanceinfo/100.019.724

26. Benzo[def]chrysene - Substance Information - ECHA. Accessed December 7, 2020. https://echa.europa.eu/sv/substance-information/-/substanceinfo/100.000.026

27. 2,2’,6,6’-tetrabromo-4,4’-isopropylidenediphenol - Substance Information - ECHA. Accessed December 7, 2020. https://echa.europa.eu/sv/substance-information/-/substanceinfo/100.001.125

28. Triphenyl phosphate - Substance Information - ECHA. Accessed December 7, 2020. https://echa.europa.eu/sv/substance-information/-/substanceinfo/100.003.739

29. 4,4’-isopropylidenediphenol - Substance Information - ECHA. Accessed December 7, 2020. https://echa.europa.eu/sv/substance-information/-/substanceinfo/100.001.133

30. KNIME | Open for Innovation. Accessed December 2, 2020. https://www.knime.com/

31. Gadaleta D, Lombardo A, Toma C, Benfenati E. A new semi-automated workflow for chemical data retrieval and quality checking for modeling applications. *J Cheminform*. 2018;10(1):60. doi:10.1186/s13321-018-0315-6

32. NCI/CADD Chemical Identifier Resolver. Accessed March 9, 2020. https://cactus.nci.nih.gov/chemical/structure

33. Williams AJ, Grulke CM, Edwards J, et al. The CompTox Chemistry Dashboard: A community data resource for environmental chemistry. *J Cheminform*. 2017;9(1):61. doi:10.1186/s13321-017-0247-6

34. PubChem. Accessed December 2, 2020. https://pubchem.ncbi.nlm.nih.gov/

35. ChemIDplus Advanced - Chemical information with searchable synonyms, structures, and formulas. Accessed December 2, 2020. https://chem.nlm.nih.gov/chemidplus/

36. Young D, Martin T, Venkatapathy R, Harten P. Are the chemical structures in your QSAR correct? *QSAR Comb Sci*. 2008;27(11-12):1337-1345. doi:10.1002/qsar.200810084

37. SciFinder® Login. Accessed December 9, 2020. https://sso.cas.org/as/authorization.oauth2?response_type=code&client_id=SciFinderWeb&redirect_uri=https%3A%2F%2Fscifinder.cas.org%3A443%2Fpa%2Foidc%2Fcb&state=eyJ6aXAiOiJERUYiLCJhbGciOiJkaXIiLCJlbmMiOiJBMTI4Q0JDLUhTMjU2Iiwia2lkIjoiOXkiLCJzdWZmaXgiOiJUZFpLZGkuMTYwNzc3MTA5MCJ9..yHnd0gcMugqO8_wAnQ5A9g.exBMd1MUhTNRDmo-Fll-q8m3sD6CGVZB8-n9an1tAraSKprQNkqDOGE513TfQLH2gT2EMlPB7Co7RCCfaOqI2A._R3qPApn_xqT22Jh_KODQA&nonce=OSI_awwUtEvKKjYgK5mRLje_gwDPmD61jxsa0iTj91M&scope=openid address email phone profile&vnd_pi_requested_resource=https%3A%2F%2Fscifinder.cas.org%3A443%2Fscifinder&vnd_pi_application_name=SciFinderWebIDF

38. www.openmolecules.org. Accessed December 3, 2020. http://www.openmolecules.org/

39. ChemSpider | Search and share chemistry. Accessed December 3, 2020. http://www.chemspider.com/

40. US EPA O. EPI Suite™-Estimation Program Interface. Accessed November 27, 2020. https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface

41. Benfenati E, Manganaro A, Gini G. VEGA-QSAR: AI inside a platform for predictive toxicology. Published online 2013:21-28. Accessed November 27, 2020. https://moh-it.pure.elsevier.com/en/publications/vega-qsar-ai-inside-a-platform-for-predictive-toxicology

42. Differences between BCF, BAF and BMF. Accessed November 27, 2020. https://www.chemsafetypro.com/Topics/CRA/definition_differences_BCF_BAF_and_BMF.html

43. Umweltbundesamt. *Protecting the Sources of Our Drinking Water*.; 2017. Accessed November 27, 2020. www.umweltbundesamt.de

44. Crookes MJ, Fisk P. *Evaluation of Using Mobility of Chemicals in the Environment to Fulfil Bioaccumulation Criteria of the Stockholm Convention FINAL REPORT*. Vol 5758319.; 2018.

45. Chapter R.11: PBT/vPvB assessment Guidance on Information Requirements and Chemical Safety Assessment Chapter R.11: PBT/vPvB assessment. Published online 2017. doi:10.2823/128621

46. EU Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 199 1-849. No Title.

47. Grisoni F, Consonni V, Villa S, Vighi M, Todeschini R. QSAR models for bioconcentration: Is the increase in the complexity justified by more accurate predictions? *Chemosphere*. 2015;127:171-179. doi:10.1016/j.chemosphere.2015.01.047

48. Gissi A, Nicolotti O, Carotti A, Gadaleta D, Lombardo A, Benfenati E. Integration of QSAR models for bioconcentration suitable for REACH. *Sci Total Environ*. 2013;456-457:325-332. doi:10.1016/j.scitotenv.2013.03.104

49. Screening N. Nordic screening report database. *https://nordicscreening.org/?page_id=18*. Published online 2020.

50. University A. Aarhus University report database. *https://dce.au.dk/udgivelser/*. Published online 2020.

51. Miljødirektoratet. Miljødirektoratet report database. *https://www.miljodirektoratet.no/publikasjoner/?Type=12*. Published online 2020.

52. (IVL) SM. Svenska Miljöinstitutet report database. *https://www.ivl.se/publikationer.html*. Published online 2020.

53. (SYKE) FEI. Finnish Environment Institute report database. *https://www.syke.fi/en-US/Research__Development/Consumption_and_production_and_sustainable_use_of_natural_resources/Publications*. Published online 2020.

54. Scholar G. Google Scholar. *https://scholar.google.com/*. Published online 2020.

55. Clarivate. Web of Science. *https://clarivate.com/webofsciencegroup/solutions/web-of-science/*. Published online 2020.

56. Medicine NL of. The PubChem database. *https://pubchem.ncbi.nlm.nih.gov/*. Published online 2020.

57. network N. NORMAN Database System. *https://www.norman-network.com/nds/*. Published online 2020.

58. NORMAN. The NORMAN network. *https://www.norman-network.com/*. Published online 2020.

59. Sun X, Zhang X, Muir DCG, Zeng EY. Identification of Potential PBT/POP-Like Chemicals by a Deep Learning Approach Based on 2D Structural Features. *Environ Sci Technol*. 2020;54(13):8221-8231. doi:10.1021/acs.est.0c01437

60. Castro M, Breitholtz M, Yuan B, Athanassiadis I, Asplund L, Sobek A. Partitioning of Chlorinated Paraffins (CPs) to Daphnia magna Overlaps between Restricted and in-Use Categories. *Environ Sci Technol*. 2018;52(17):9713-9721. doi:10.1021/acs.est.8b00865

Supplementary Information

Information and data on the literature survey, predicted environmental fate and human health endpoints and scoring of hits can be found in the Excel file entitled "Supplementary information on new or emerging risk chemicals"

KNIME workflow

To run the workflow, the KNIME software work tool must be downloaded. In this study version 4.2.1 was used. All available extensions were installed. The node HTML-parser was not available for automated update but collected from https://nodepit.com/node/ws.palladian.nodes.retrieval.parser.HtmlParserNodeFactory.

The node is needed to collect the SMILES and names from the external sources (CIR, CompTox, PubChemID and ChemID). The workflow and instructions on how to run the program is found online at https://github.com/DGadaleta88/data_curation_workflow

Figures and Tables

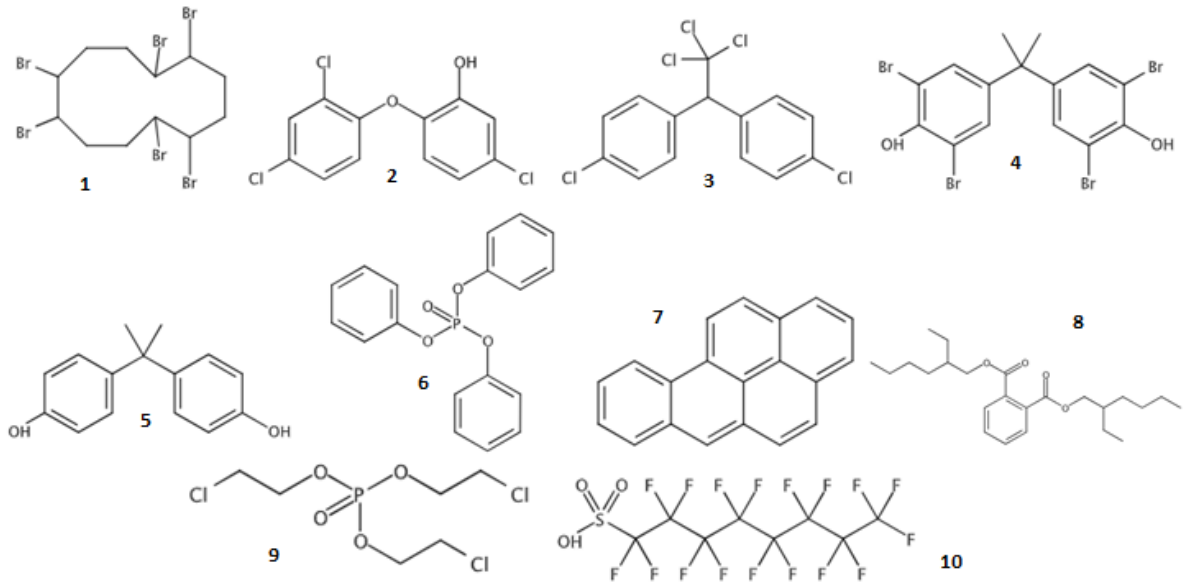


Figure S1. Structures of the positive controls used in this study.

Table S1. Summary of positive controls including the canonical SMILES used for hazard screening and hazard properties according to REACH.

Common name	Product group	CAS	Canonical SMILES	Properties[1]
HBCDD Hexabrom-cyclododecane	Brominated flame retardant	3194-55-6	BrC1CCC(Br)C(Br)CCC(C(CCC1Br)Br)Br	PBT POP R Ss
Triclosan	Biocide	3380-34-5	Clc1ccc(c(c1)O)Oc1ccc(cc1Cl)Cl	PBT ED
DDT Clofenotane	Pesticide	50-29-3	ClC(C(c1ccc(cc1)Cl)c1ccc(cc1)Cl)(Cl)Cl	POP C
TBBPA Tetrabromo-bisphenol A	Brominated flame retardant	79-94-7	CC(c1cc(Br)c(c(c1)Br)O)(c1cc(Br)c(c(c1)Br)O)C	PBT ED
BPA Bisphenol A	Monomer	80-05-7	CC(c1ccc(cc1)O)(c1ccc(cc1)O)C	ED Ss R
TPhP Triphenyl-phosphate	Organophosphate Plasticizer and flame retardant	115-86-6	O=P(Oc1ccccc1)(Oc1ccccc1)Oc1ccccc1	ED
Benzo[a]pyrene	Polycyclic aromatic hydrocarbon	50-32-8	c1ccc2c(c1)c1ccc3c4c1c(c2)ccc4ccc3	PBT POP R M C Ss
DEHP	Phthalate Plasticizer	117-81-7	CCCCC(COC(=O)c1ccccc1C(=O)OCC(CCCC)CC)CC	ED R
TCEP	Organophosphate	115-96-8	ClCCOP(=O)(OCCCl)OCCCl	R C
PFOS perfluorooctane sulfonate	PFAS	1763-23-1	FC(C(C(C(C(S(=O)(=O)O)(F)F)(F)F)(F)F)(F)F)(C(C(C(F)(F)F)(F)F)(F)F)F	R C

[1]PBT: Persistent, Bioaccumulative and Toxic; POP: Persistent organic pollutant; ED: Endocrine disrupting; CMR: Carcinogenic, Mutagenic and Reprotoxic; Ss: Skin sensitizing.

Toxikologiska rådet

– expertorgan för rådgivning och samråd i toxikologiska frågor

The Toxicological Council

– body of experts for advice and consultation on toxicological issues