Environment International 152 (2021) 106511

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

An annotation database for chemicals of emerging concern in exposome research

Jeroen Meijer [a,b], Marja Lamoree [b], Timo Hamers [b], Jean-Philippe Antignac [c], Sébastien Hutinet [c], Laurent Debrauwer [d,e], Adrian Covaci [f], Carolin Huber [g], Martin Krauss [g], Douglas I. Walker [h], Emma L. Schymanski [i], Roel Vermeulen [a], Jelle Vlaanderen [a,*]

[a] Institute for Risk Assessment Sciences (IRAS), Utrecht University, Utrecht, the Netherlands
[b] Department Environment & Health, Vrije Universiteit, Amsterdam, the Netherlands
[c] Oniris, INRAE, LABERCA, Nantes, France
[d] Toxalim (Research Centre in Food Toxicology), Toulouse University, INRAE, ENVT, INP-Purpan, Toulouse, France
[e] Metatoul-AXIOM Platform, National Infrastructure for Metabolomics and Fluxomics: MetaboHUB, Toxalim, INRAE, Toulouse, France
[f] Toxicological Center, University of Antwerp, Belgium
[g] Department Effect-Directed Analysis, Helmholtz Centre for Environmental Research – UFZ, Leipzig, Germany
[h] Department of Environmental Medicine and Public Health, Icahn School of Medicine, Mount Sinai, New York, NY, USA
[i] Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Belvaux, Luxembourg

ARTICLE INFO

Handling Editor: Lesa Aylward

Keywords:
Annotation database
High-resolution mass spectrometry
Chemicals of emerging concern
Exposome research

ABSTRACT

Background: Chemicals of Emerging Concern (CECs) include a very wide group of chemicals that are suspected to be responsible for adverse effects on health, but for which very limited information is available. Chromatographic techniques coupled with high-resolution mass spectrometry (HRMS) can be used for non-targeted screening and detection of CECs, by using comprehensive annotation databases. Establishing a database focused on the annotation of CECs in human samples will provide new insight into the distribution and extent of exposures to a wide range of CECs in humans.

Objectives: This study describes an approach for the aggregation and curation of an annotation database (CECscreen) for the identification of CECs in human biological samples.

Methods: The approach consists of three main parts. First, CECs compound lists from various sources were aggregated and duplications and inorganic compounds were removed. Subsequently, the list was curated by standardization of structures to create "MS-ready" and "QSAR-ready" SMILES, as well as calculation of exact masses (monoisotopic and adducts) and molecular formulas. The second step included the simulation of Phase I metabolites. The third and final step included the calculation of QSAR predictions related to physicochemical properties, environmental fate, toxicity and Absorption, Distribution, Metabolism, Excretion (ADME) processes and the retrieval of information from the US EPA CompTox Chemicals Dashboard.

Results: All CECscreen database and property files are publicly available (DOI: https://doi.org/10.5281/zenodo.3956586). In total, 145,284 entries were aggregated from various CECs data sources. After elimination of duplicates and curation, the pipeline produced 70,397 unique "MS-ready" structures and 66,071 unique QSAR-ready structures, corresponding with 69,526 CAS numbers. Simulation of Phase I metabolites resulted in 306,279 unique metabolites. QSAR predictions could be performed for 64,684 of the QSAR-ready structures, whereas information was retrieved from the CompTox Chemicals Dashboard for 59,739 CAS numbers out of 69,526 inquiries. CECscreen is incorporated in the *in silico* fragmentation approach MetFrag.

Discussion: The CECscreen database can be used to prioritize annotation of CECs measured in non-targeted HRMS, facilitating the large-scale detection of CECs in human samples for exposome research. Large-scale detection of CECs can be further improved by integrating the present database with resources that contain CECs (metabolites) and meta-data measurements, further expansion towards *in silico* and experimental (e.g., MassBank) generation of MS/MS spectra, and development of bioinformatics approaches capable of using correlation patterns in the measured chemical features.

* Corresponding author at: Institute for Risk Assessment Sciences (IRAS), Utrecht University, Yalelaan 2, 3584 CM Utrecht, the Netherlands.
E-mail address: j.j.vlaanderen@uu.nl (J. Vlaanderen).

https://doi.org/10.1016/j.envint.2021.106511
Received 1 September 2020; Received in revised form 3 February 2021; Accepted 6 March 2021
Available online 24 March 2021

1. Introduction

Environmental exposures are a key contributor to disease and premature death (Landrigan et al., 2018). Yet, only a fraction of chemicals produced in significant amounts with potential to enter the environment have sufficient information to characterize human exposure levels and toxicity. Furthermore, information on potential by-products as well as biotic and abiotic transformation products is almost unknown (Landrigan et al., 2018; Vermeulen et al., 2020). Chemicals for which limited information is available in terms of exposure and toxicity and chemicals for which we do not know the existence off are commonly referred to as "known unknowns" and "unknown unknowns" (Little et al., 2012), respectively. Characterizing these exposures are critical to identify the role of the chemical exposome in human health and to advance non-targeted analytical platforms for universal chemical surveillance screening (Vermeulen et al., 2020). Of particular interest are Chemicals of Emerging Concern (CECs) (Sauvé and Desrosiers, 2014). This label is used by various agencies around the world to classify chemicals of potential concern to human or environmental health due to expected toxicity or exposure levels, but for which very limited information related to either quantity is available. Due to this broad classification, the number of CECs is extremely high and ever changing and includes compounds with a wide variety of chemical properties and uses (*e.g.* pesticides, flame retardants, surfactants, pharmaceuticals, etc). Approaches such as suspect screening and effect-directed analysis (EDA) try to shed light on the presence of CECs in human biological matrices and other complex samples.

The goal of this work is to develop a universal screening database of CECs for exposome research (Pourchet et al., 2020; Vermeulen et al., 2020), with emphasis on detection of CECs in biological samples collected from human populations. An example of the possible application of this database in exposome research is included in Pourchet et al. (2021). Due to the large number of CECs, targeted assessment of exposure levels in human samples is too costly, time consuming and often requires large sample volumes (Wishart et al., 2018). Chromatography techniques, including liquid chromatography (LC) or gas chromatography (GC), combined with high-resolution mass spectrometry (HRMS) provides an opportunity to screen for the presence of a wide range of CECs in human samples.

A major challenge in the use of HRMS to detect CECs in human samples is the assignment of chemical identities to the several thousands of mass spectral signals ("chemical features") that are detected in a typical non-targeted HRMS analysis. While identification of these chemical features requires validation with authentic standards (i.e., comparison of retention times and MS/MS spectra), etc. (Schymanski et al., 2014), an initial annotation that provides potential chemical identities is possible by comparing the accurate mass from the detected chemical features to a database that contains exact masses of known compounds. Although this technique requires a priori information on compounds that might be present, this is a useful approach within exposome research to 'zoom in' on a specific compound class/group of compounds of interest and may facilitate the selection of relevant features to focus on for identification. Further insight into the biological impact of the annotated compounds can be acquired by combining this approach with other tools such as the fully untargeted screening of the metabolome. However, many chemical features will remain unannotated and subsequent steps are needed to identify "unknown unknowns" specifically.

Although many databases exist, few are suitable for the annotation of CECs in biological samples. Metabolomics databases, such as HMDB, (Wishart et al., 2018; Wishart et al., 2007) the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000) and METLIN (Guijas et al., 2018), primarily incorporate compounds that have been reported in literature. Both HMDB and KEGG were primarily designed as a resource for compounds present in the human endogenous metabolome, while METLIN includes both endogenous metabolites and environmental chemicals, including >700 000 chemicals from the US EPA "Distributed Structure-Searchable Toxicity (DSSTox)" database. (Grulke et al., 2019) The U.S. Environmental Protection Agency has created the CompTox Chemicals Dashboard (Williams et al., 2017), which combines physicochemical properties, reference counts and a wide-range of additional meta-data for over 882,000 chemicals, providing one of most comprehensive chemical databases currently available for exposome research. The suspect lists created under the NORMAN network (https://www.norman-network.com/nds/SLE/) also provides a large source of potential CECs, however these lists are mostly focused on environmental monitoring. Additional databases include PubChem (Kim et al., 2019) and ChemSpider (Pence and Williams, 2010), both of which contain information on millions of chemical structures; however, both PubChem and ChemSpider suffer from the potential for false-positive matches, due to chemical entries that have an extremely low probability of representing true environmental exposures. While all of these databases provide critical resources for annotation of non-targeted HRMS features, the large number of chemicals, focus on endogenous metabolites or chemicals found in environmental samples, and limited ability to define compounds as CECs limits application for annotation of CECs in biological samples.

Two other databases, the Blood Exposome Database (Barupal and Fiehn, 2019) and the Exposome Explorer (Neveu et al., 2017), focus specifically on compounds present in the exposome. The Exposome Explorer (now included in HMDB v4.0) contains a systematically collected and manually curated collection of 908 compounds that are reported in the literature as measured in human samples. The high-quality standards of the Exposome Explorer, due to manual curation, have resulted in the small scale. The Blood Exposome Database is larger (n = 65,957) and incorporates a text mining approach to include potentially relevant information from the literature. However, this database is tailor made to blood only, not other matrices such as urine and includes many endogenous compounds.

Although a large number of data sources dealing with CECs exist we believe there is still need for a comprehensive annotation database for CECs focused on screening in human matrices specifically. In this work, performed in the frame of the Human Biomonitoring for Europe (HBM4EU) initiative (https://www.hbm4eu.eu/), CECscreen, a database of CECs including their predicted Phase I metabolites that can be used as an annotation source in exposome research is introduced. This database is a compilation of existing data sources dealing with CECs considered to be important for screening in human samples and includes annotation databases from the environmental field and regulatory agencies among others. Furthermore, to ensure applicability for annotation in HRMS data, curation and standardization of the retrieved CECs are performed in the form of "MS-ready" structures (McEachran et al., 2018). The use of "MS-ready" structures help link the common substance form to the form detected by HRMS in annotation databases. As most CECs will undergo metabolism in the human body and are therefore likely more prominently present as metabolite(s) than as a parent compound, the inclusion of Phase I metabolites would also improve the applicability of this database as an annotation source in human samples. The database was populated by collating a large number of lists of CECs produced by (regulatory) agencies and academic initiatives around the world. We describe here the applied qualitative consolidation and harmonization approach, steps that were taken to make the database "MS-ready", identification of relevant metadata, and generation of predicted metabolites. Further, it describes how the CECscreen compares with existing annotation resources, such as Exposome Explorer and the Blood Exposome Database and provides perspective on future developments that will further improve the identification of CECs in human biological samples.

2. Methods

2.1. Identification and compilation of CECs data sources

A general overview of the steps included in the curation work of this database is provided in Fig. 1. As an initial screening, data sources containing CECs were identified in a questionnaire in 2017 by members of the HBM4EU initiative. Compound name, Chemicals Abstract Service (CAS) number, and Simplified Molecular-Input Line Entry Specification (SMILES, when available) were extracted from the individual databases. Databases were merged based on CAS number. Using the CAS number as input, the "webchem" package (Szöcs et al., 2020) in R was used to retrieve missing SMILES. Entries for which no SMILES could be retrieved were excluded from the list.

2.2. Data curation and preparation for CECs screening

Correctness of the SMILES was evaluated by checking whether the valency matches the structures. Subsequently, adjusted KNIME workflows of Mansouri et al. (2016) and McEachran et al. (2018) were used to generate "MS- (Mansouri et al., 2016) and QSAR-ready" structures (McEachran et al., 2018) (*Figure SM-1*). The first step of the workflow was the removal of inorganic compounds, which are defined here as all substances without the element carbon. Subsequently, organic mixtures were split and each mixture constituent was entered into CECscreen as a separated entry with a link to the other constituents. In addition, any resulting inorganic components related to salts, plus water were removed. The final step to create "MS-ready" structures was chemical transformation and neutralization of the compounds to the forms that could be found in HRMS. To create "QSAR-ready" structures, stereochemistry information was removed from the "MS-ready" structures. A more detailed description of the standardization and transformation steps in the workflows was published by McEachran et al. (2018). The original workflows were adjusted by removing the organometallic filter and retaining the stereochemistry information for the "MS-ready" structures. Furthermore, log P values were predicted based on the "MS-ready" structures using the XLogP (Wang et al., 1997) node from the KNIME CDK plug-in (Beisken et al., 2013).

OpenBabel (O'Boyle et al., 2011) was used to convert the "MS- and QSAR-ready" SMILES into canonical SMILES, InChIs and InChIKeys. The "doBy" package (Højsgaard and Halekoh, 2006) in R was used to summarize the CECs list based on unique "MS-ready" InChIKeys and duplicates were removed. The first CAS number and compound name that was reported for the structures were put in a separate column to be used as main identifier. The other variants were reported in another column to avoid losing this information. Monoisotopic mass, masses of commonly occurring adducts, molecular formula and remaining charge were calculated using the RChemMass package (Schymanski, 2019) in R. The adducts included were $[M+H]^+$, $[M+Na]^+$, $[M+NH_4]^+$, $[M-H]^-$ and $[M+CH_3COO]^-$.

2.3. Simulation of metabolites

The comprehensive and open source software tool BioTransformer was used for the prediction of metabolites (Djoumbou-Feunang et al., 2019). CDK objects representing structures were generated using the "QSAR-ready" canonical SMILES and saved in an SDF file that was imported in the BioTranformer software (Djoumbou-Feunang et al., 2019). Metabolites were predicted using the Phase I (CYP450) transformation option and the program was run in Windows via the command line. Data retrieved included metabolite ID, predicted Phase I metabolite InChI and InChIKey, biotransformation reaction, enzymes involved and precursor "QSAR-ready" InChIKeys. Further data curation was performed similar to the curation work of the aggregated CECs list. In addition, compound name and CAS numbers were retrieved based on the InChIKeys using the "webchem" package (Szöcs et al., 2020) in R.

2.4. Generation of physicochemical properties, environmental fate, and toxicity endpoints

2.4.1. CompTox Chemicals Dashboard

The CompTox Chemicals Dashboard (Williams et al., 2017) was used to retrieve existing and predicted data on toxicity, exposure and prevalence in the literature of the entries in the aggregated list. Batch searches based on CAS numbers were performed containing 5000 compounds per batch. Information extracted from the CompTox Chemicals Dashboard included predicted median exposure in mg/kg bodyweight /day, availability of Toxicity Value data, ToxCast data (number of assays and percentage active), number of PubMed articles, chemicals and products database count and TEST toxicity predictions including 48 h Daphnia magna exposure LC50, developmental toxicity, Ames mutagenicity, oral rat LD50, Tetrahymena pyriformis IGC50 and 96 h fathead minnow LC50.

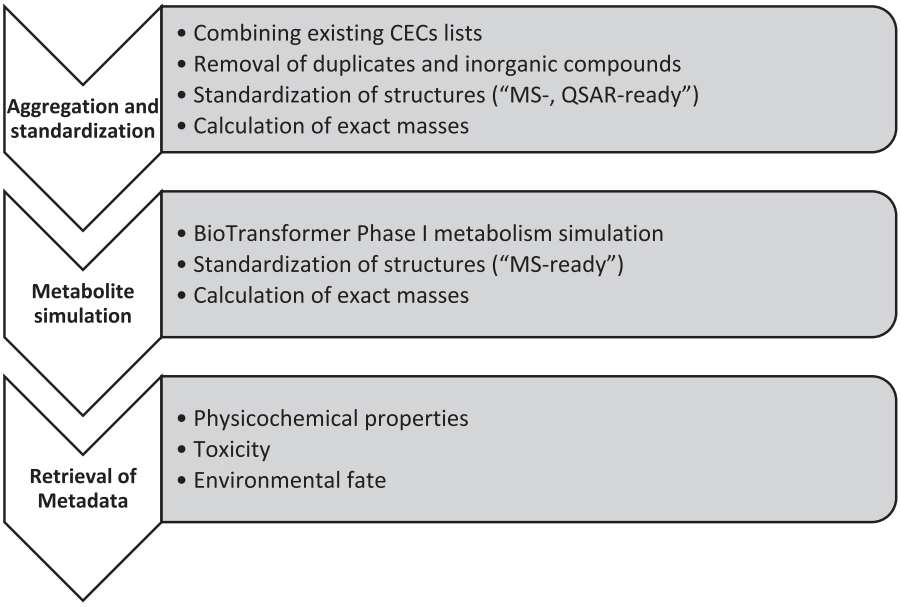

Fig. 1. The three major steps and sub steps in the pipeline to create a "MS-ready" database of CECs for screening in human samples. The first step is the aggregation of existing CECs lists and standardization to produce structure forms that can be measured with MS or are compliant with QSAR models. In the second step, metabolites are simulated and their structures are standardized. In the final step, metadata is collected and predicted for physicochemical properties, environmental fate, and toxicity.

2.4.2. OPERA

The OPEn (q)saR App (OPERA) (Mansouri et al., 2018) was used for the prediction of physicochemical properties, environmental fate, toxicity endpoints and ADME properties. MDL molfiles (v2000) were created from the "QSAR-ready" smiles and saved in an SDF file and imported into the OPERA User Interface (UI). Physicochemical properties that were predicted included the octanol–water partitioning coefficient (log P), melting point, boiling point at 760 mm Hg, vapor pressure, water solubility at 25 °C, Henry's Law constant, octanol–air partitioning coefficient, HPLC retention time, logarithmic acid dissociation constant (pK_a) and octanol–water distribution coefficient (log D). The environmental fate predictions included the fish bioconcentration factor (log BCF), OH rate constant for atmospheric gas-phase reactions, biodegradation half-life, ready biodegradability of organic chemicals, whole body primary biotransformation rate constant and soil adsorption coefficient of organic compounds. Toxicity endpoints predicted included acute toxicity, agonistic and antagonistic estrogen, as well as androgen receptor activity. Finally, the ADME endpoints that were predicted included the human plasma fraction unbound and the human hepatic intrinsic clearance.

2.5. Software and hardware

All data mining, curation and modelling was performed on a HP desktop computer with an Intel® Core™ i7-4790 CPU 3.60 GHz processor and 16.0 GB RAM with a 64-bit Windows 10 operating system. Data handling and curation was performed using version 3.5.1 of the R Project for Statistical Computing software (Team R, 2013) with the "webchem" (Szöcs et al., 2020), "doBy" (Højsgaard and Halekoh, 2006), "RChemMass" (Schymanski, 2019), enviPat (Loos et al., 2015) and rcdk (Guha, 2007) packages installed. Rcdk was installed to access the functionalities of the JAVA Chemistry Development Kit (Steinbeck et al., 2003). Furthermore, the KNIME Analytics Platform version 4.0.0 (Berthold et al., 2009) was used with the Marvin Chemistry (ChemAxon. Marvin., 2014) and CDK extensions (Beisken et al., 2013) to run the prebuild "MS-ready" (Mansouri, 2017) and "QSAR-ready" (Mansouri, 2017) workflows as well as OpenBabel version 2.4.1 (O'Boyle et al., 2011). BioTransformer version 1.1.0 (Djoumbou-Feunang et al., 2019) was used to predict metabolites and the Open (q)saR App (OPERA) (Mansouri et al., 2018) was used for the prediction of several physicochemical properties, environmental fate, toxicity endpoints and Absorption, Distribution, Metabolism, Excretion (ADME) properties.

3. Results

3.1. Identification, aggregation and curation of CECscreen

The questionnaire provided to HBM4EU members resulted in the identification of 11 databases of CECs (Table 1), as well as the CompTox Chemicals Dashboard, which included at that stage 40 CECs lists (*Table SM-1*), several of which were integrated via the NORMAN Suspect List Exchange (https://www.norman-network.com/nds/SLE/). In total 145,284 entries were included in the databases, which contained 107,655 unique CAS numbers and 109,237 unique SMILES.

For 30,740 entries with missing structural information in the databases, no SMILES structure could be retrieved with the "webchem" package (Szöcs et al., 2020) in R and these entries were therefore excluded (as described in the methods). Furthermore, another 3227 inorganics were excluded. Multi-component entries were separated and the organic mixture constituents were reinserted, resulting in a total number of 116,758 entries in the compiled list. Removal of duplicates and standardization of structures resulted in a total of 70,397 unique "MS-ready" structures and 66,071 unique QSAR-ready structures, which corresponds to 29,504 unique chemical formulas (DOI: https://doi.org/10.5281/zenodo.3956586). The monoisotopic masses range based on the "MS-ready" structures was from 16 to 6981 Da (Fig. 2).

Table 1
List names, compound count and list description of the CECs data sources proposed by members of the HBM4EU project.

List name	Compound count	List description
NORMAN merged suspects list dated 24/05/2017	14,633	Collection of NORMAN suspect lists for environmental monitoring (NORMAN, 2015; Network et al., 2020)
EFSA FoodToxDB	5812	Compilation of chemical and toxicological information on compounds assessed by EFSA (EFSA, 2002)
ECHA Candidate list	203	list of emerging chemicals of high concern for authorization (ECHA, 2017)
COSING (cosmetic ingredient database)	25,267	European Commission list with information on cosmetic chemicals and ingredients (European Commission, 2017)
REACH ANNEX III	64,900	List of substances likely to meet the criteria of Annex III of the REACH regulation (ECHA, 2017)
List of PBT/vPvBsubstances (67/548/EECd and 793/93/EECe)	128	List of suspected PBT and vPvB chemicals under the previous EU chemicals legislation (ECHA, 2007)
T3DB	3674	Database containing target information about toxins (TMIC, 2014; Wishart et al., 2015)
US EPA CPCAT database (includes SPIN 2000)	43,600	Database containing chemical information and usage in consumer products (Epa, 2017)
OECD HPV 2007	4645	OECD list of high production volume chemicals (IOMC, 2009)
EC EDS list	433	European Commission List of possible endocrine disrupting chemicals (European Commission, 2017)
TEDX list	1409	List of potential endocrine disrupting compounds (TEDX, 2017)
US EPA Chemicals Dashboard lists	74,705	Collection of CECs lists retrieved from the US EPA Chemistry Dashboard (Epa, 2017)

CECscreen and related metabolites are also directly incorporated into MetFrag (Ruttkies et al., 2016) (https://msbi.ipb-halle.de/MetFrag/).

3.2. Metabolite predictions: CECscreen_Metabolite_DB

BioTransformer was able to predict Phase I metabolites for 51,094 out of the 66,071 (77.3%) QSAR-ready structures, resulting in 434,953 transformations. After duplicate removal and standardization of structures 306,279 unique metabolites remained. 205,765 products were formed by hydroxylation, 43,429 by epoxidation, 22,121 by dealkylation, 13,029 by oxidation, 12,039 by desaturation, 6878 by reduction. The remaining 3018 products were formed by other types of reaction like ring opening, deisopropylation, dearylation, desulfurization, dehydrogenation, NS-cleavage, dechloroethylation, deboronation, dealkoxylation, depropargylation and formation of iminium or pyridinium (Fig. 3). As many of these predicted compounds may or may not already be known in databases, the R package "webchem" (Szöcs et al., 2020) was used to retrieve compound name and CAS numbers based on the InChIKeys, which resulted in the retrieval of 24,872 (8.1%) names and 8257 (2.7%) CAS numbers. For 8233 (2.7%) metabolites, both the name and CAS number were retrieved.

3.3. CompTox Chemicals Dashboard results: CECscreen_Comptox_DB

CompTox Chemicals Dashboard information was retrieved for 56,445 out of the 69,526 CAS numbers reported in CECscreen (Table 2). 5.895 (8.5%) of the matches included estimations of exposure and

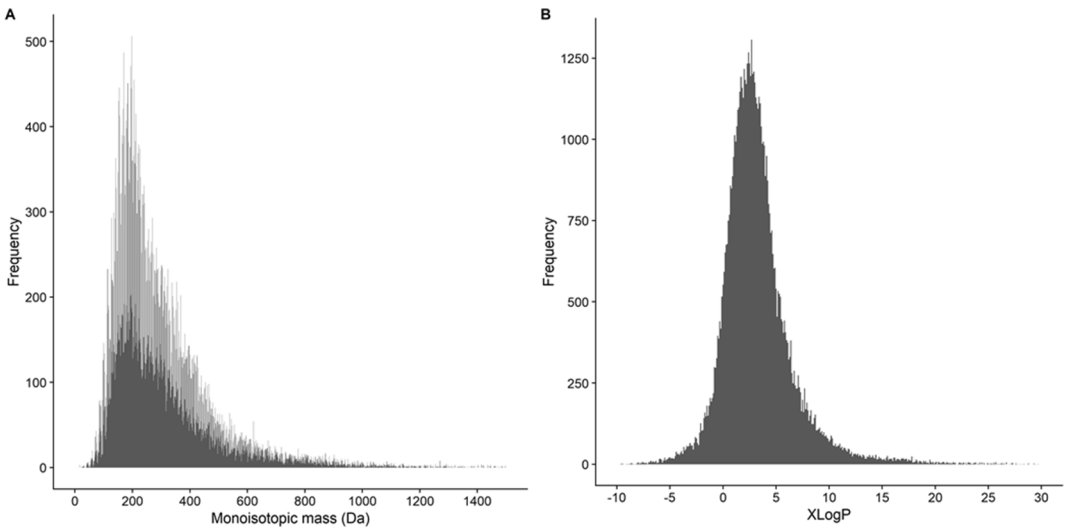

Fig. 2. Histogram of the monoisotopic masses (A) and the predicted XLogP values (B) of the entries in the combined CECs list. Bin width is 1 Da and 0.1 for A and B, respectively.

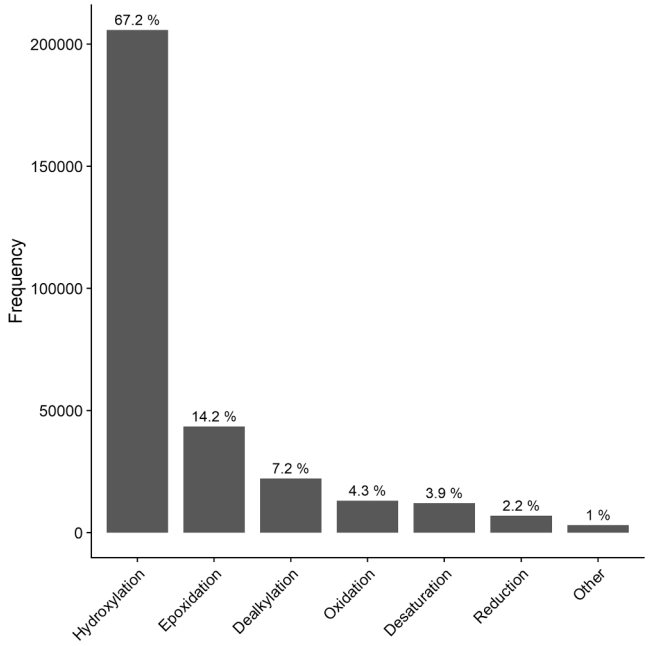

Fig. 3. Bar graph of different transformation reaction frequencies resulting from the BioTransformer predictions. Numbers above the bars represent the total percentage of metabolites predicted through that specific transformation reaction. "Other" includes ring opening, deisopropylation, dearylation, desulfurization, dehydrogenation, NS-cleavage, dechloroethylation, deboronation, dealkoxylation, depropargylation and formation of iminium or pyridinium.

18,553 (26.7%) also contained information from the chemicals and products database. For 11,266 (16.2%) of the matches ToxVal data was available, while 6544 (9.4%) of the matches also included ToxCast data. Source links to the ToxVal data is also included in the curated database. TEST toxicity prediction could be retrieved for between 24,639 (35.4%) and 37,541 (54.0%) of the entries depending on the endpoint. Furthermore, for 15,190 (21.8%) to 39,812 (57.3%) entries predictions related to several chemical properties could be retrieved. Finally, 11,939 (17.2%) of the matches included information on the number of times it is referenced in articles found on PubMed.

Table 2
Summary of the data retrieved from the CompTox Chemicals Dashboard batch search of 69,526 CAS numbers. Detailed information with regards to the extracted information from the CompTox Chemicals Dashboard can be retrieved from Williams et al. (United States Environmental Protection Agency (US EPA), 2020).

	Count	% of inquiry	Min	Max	Median
Matches	56,445	81.2	–	–	–
ExpoCast median exposure (mg/kg bodyweight/day)	5895	8.48	2.96e−09	2.81e−4	1.67e−07
ToxVal availability	11,266	16.2	–	–	–
ToxCast (% active)	6544	9.4	0	73.84	2.98
PubMed articles (count)	11,939	17.2	1	280,169	22
CPDat (count)	18,553	26.7	1	522,354	4
TEST predictions (Tox)					
BCF	32,431	46.6	1.92e−05	529,633	14.4
48hr Daphnia LC50 (M)	33,607	48.3	1.75e−13	0.19	3.17e−05
96hr Fathead minnow LC50 (M)	32,562	46.8	5.12e−12	0.89	2.97e−05
40hr T. pyriformis IGC50 (M)	24,639	35.4	5.61e−11	1.63	1.33e−04
Oral rat LD50 (mol/kg)	31,084	44.7	2.50e−07	0.25	6.67e−03
DevTox	37,541	54.0	−1.63	1.55	0.65
AMES mutagenicity	37,208	53.5	−0.59	1.48	0.26
TEST predictions (physical properties)					
Boiling point (°C)	34,836	50.1	−70.02	1018.6	308.49
Density (g/cm^3)	39,356	56.6	0.57	4.78	1.21
Flash point (°C)	38,021	54.7	−47.26	1431.4	157.25
Melting point (°C)	39,812	57.3	−140.83	904.32	89.28
Surface tension (dyn/cm)	15,190	21.8	6.66	64.29	33.45
Thermal conductivity (mW/mK)	18,571	26.7	24.15	275.91	141.85
Viscosity (viscosity cP)	19,890	28.6	0.13	986.28	6.40
Vapor pressure (mmHg)	35,041	50.4	6.15e−27	27,605.8	3.09e−05
Water solubility (M)	38,018	54.7	3.16e−14	19.68	9.84e−04

3.4. OPERA results: CECscreen_OPERA_Predictions

OPERA predictions were executed on 64,684 QSAR-ready structures, excluding organometallic compounds and those with a mass >1000 Da (Table 3). Data retrieved included the predictions but also information on whether a compound is within the applicability domain (AD) of the QSAR model as well as confidence indices that represents the reliability of the prediction based on the the accuracy of prediction of the five most similar compounds (Mansouri et al., 2018). Predictions were performed for 96.3% to 100% of the structures depending on the endpoint with exception of pK_a predictions, which were achieved for 42% for acidic pK_a and 33.8% for basic pK_a. The number of compounds within the AD varied notably between endpoints. For the physical chemical properties, the number of compounds within the AD ranged from 53.7% to 88.8%. Similar figures were found for predicted environmental fate endpoints with 48.5% to 85.5% of the compounds falling within the AD. A larger proportion of the compounds were included in the AD for the toxicity and ADME endpoints with numbers ranging from 97.5% to 100% and 86.1% to 86.9%, respectively. The same pattern was also observed for the confidence indices, as the average confidence index of the toxicity endpoints is abundantly higher than for the other endpoints and with lower standard deviation.

Table 3
Results from the running of the OPERA QSAR models on the QSAR-ready smiles. Data shown include the number of predictions made and percentage of the complete list. Furthermore, data are shown about the number of compounds within the applicability domain (AD) of the QSAR model and the average confidence index regarding the reliability of the prediction based on the accuracy of prediction of the five most similar compounds. Mansouri et al. provides a detailed description of all included prediction endpoints (Mansouri et al., 2018).

	Predictions		Within AD		Confidence Index	
	Count	%	Count	%	Average	SD
Physicochemical properties						
Boiling point	64,641	99.9	34,708	53.7	0.53	0.19
Henry's Law constant	64,606	99.9	35,333	54.6	0.39	0.17
K_{OA}	64,681	100.0	35,974	55.6	0.62	0.22
log P	64,593	99.9	56,614	87.5	0.58	0.13
Melting point	64,641	99.9	57,527	88.9	0.56	0.15
Vapor pressure	64,672	100	48,957	75.7	0.46	0.17
Water solubility	64,578	99.8	56,401	87.2	0.53	0.14
HPLC RT	64,565	99.8	53,813	83.2	0.52	0.07
pK_a a	27,151	42	52,537	81.2	0.57	0.15
pK_a b	21,883	33.8	52,537	81.2	0.57	0.15
log D	64,493	99.7	50,117	77.5	0.57	0.11
Environmental fate						
log BCF	64,571	99.8	51,196	79.2	0.54	0.10
AOH	64,601	99.9	31,393	48.5	0.54	0.10
BioDeg	62,320	96.4	10,433	16.1	0.55	0.13
RBioDeg	62,304	96.3	55,498	85.8	0.60	0.16
KM	64,583	99.8	46,691	72.2	0.51	0.12
KOC	62,299	96.3	48,879	75.6	0.52	0.16
Toxicological endpoints						
CERAPP-Binding	64,680	100	64,518	99.7	0.92	0.07
CERAPP-Agonist	64,671	100	64,370	99.5	0.94	0.07
CERAPP-Antagonist	64,680	100	64,334	99.5	0.93	0.07
CoMPARA-Binding	64,594	99.9	64,593	99.9	0.93	0.07
CoMPARA-Agonist	64,592	99.9	64,580	99.8	0.97	0.05
CoMPARA-Antagonist	64,594	99.9	64,517	99.7	0.93	0.07
CATMoS-VT	64,680	100	64,657	100	0.96	0.04
CATMoS-NT	64,671	100	64,669	100	0.91	0.04
CATMoS-EPA	64,676	100	64,649	100	0.84	0.06
CATMoS-GHS	64,607	99.9	64,605	99.9	0.85	0.08
CATMoS-LD50	64,679	100	63,083	97.5	0.84	0.10
ADME endpoints						
FuB	64,679	100	55,721	86.1	0.63	0.07
Clint	64,678	100	56,233	86.9	0.44	0.15

4. Discussion

The aim of the present study was to establish a database of CECs (CECscreen) for screening of these compounds in human biological samples, using non-targeted HRMS and thus to provide a resource for chemical annotation of the exposome. The database includes not only the CECs themselves, but also their corresponding Phase I metabolites that are likely to be present in human samples. The database was established by aggregating and curating existing chemicals databases containing CECs selected via a survey and by taking necessary steps to make the database "MS-ready". Furthermore Phase I metabolites for all parent compounds included in the database were simulated and information related to physicochemical properties, environmental fate, and toxicity were retrieved and predicted.

Combining and aggregating the CECs lists resulted in 70,397 unique "MS-ready" (i.e., desalted structures, with mixtures separated into components) and 66,071 unique "QSAR-ready" structures in CECscreen. Furthermore, primary metabolite simulation resulted in an additional 306,279 compounds in CECscreen_Metabolite_DB, i.e. 376,676 compounds all together. However, the chemical space that is classified as CECs changes over time due to our increasing understanding of CECs and their health effects. Therefore, this list should be periodically updated to incorporate new CECs. Many compounds in CECscreen were also included in the CompTox Chemicals Dashboard (CECscreen_CompTox_DB), but the amount of data available for the different endpoints varied widely. The diversity in data availability highlights the need for agnostic and systematic non-targeted approaches to gain more insight into CECs (Vermeulen et al., 2020). CECscreen also includes predictions of physicochemical properties, environmental fate and toxicity (CECscreen_OPERA_Predictions). In addition to the predicted properties, data was included whether the chemical falls within the AD of that particular QSAR model Predictions were performed for a major part of the list, although the portion of compounds falling within theAD of the QSARs varied between endpoints.

There is a degree of uncertainty in all predictions and this information is important for evaluation and to facilitate prioritization. Therefore, information with regards to confidence of prediction is included in the metadata files of CECscreen where available. Measures of confidence of predicted values in the CompTox Chemicals Dashboard were not available. However, for EXPOCAST predictions, a hyperlink is provided in CECscreen that links to the CompTox Chemicals Dashboard page which includes 95% confidence intervals of the predicted exposure values. For OPERA predictions a confidence index is provided which is a measure of the reliability of prediction when the query chemical falls within the AD and is calculated based on the prediction accuracy of the five most similar compounds of the query chemical. A detailed description of the implementation of such an approach is reported in Sahigara et al. (2012).

The current state of CECscreen and associated files function as a first step to facilitate screening for CECs in human populations. The overlap based on unique InChIKey between CECscreen and other exposome focused databases like the Human Blood Exposome Database, Exposome Explorer and DSSTox is illustrated in Fig. 4. The largest overlap of CECscreen is with the DSSTox database, which is also incorporated into METLIN (72.4%). However, this is still only a minor fraction of the total DSSTox database (6.4%) which also incorporates a large number of endogenous compounds and predicted metabolites. Furthermore, over half of the Exposome Explorer database is also included into CECscreen. The overlap of CECscreen with the Blood Exposome Database is relatively small. A total of 13,175 compounds overlap between the databases which accounts for 18.7% and 20.2% of CECscreen and the Blood Exposome Database, respectively. The difference between the databases is most likely the result of the large number of blood related endogenous markers that are included in the Blood Exposome Database, which are outside the scope of CECscreen and the different aggregation methods used. The limited overlap between the Blood Exposome Database and

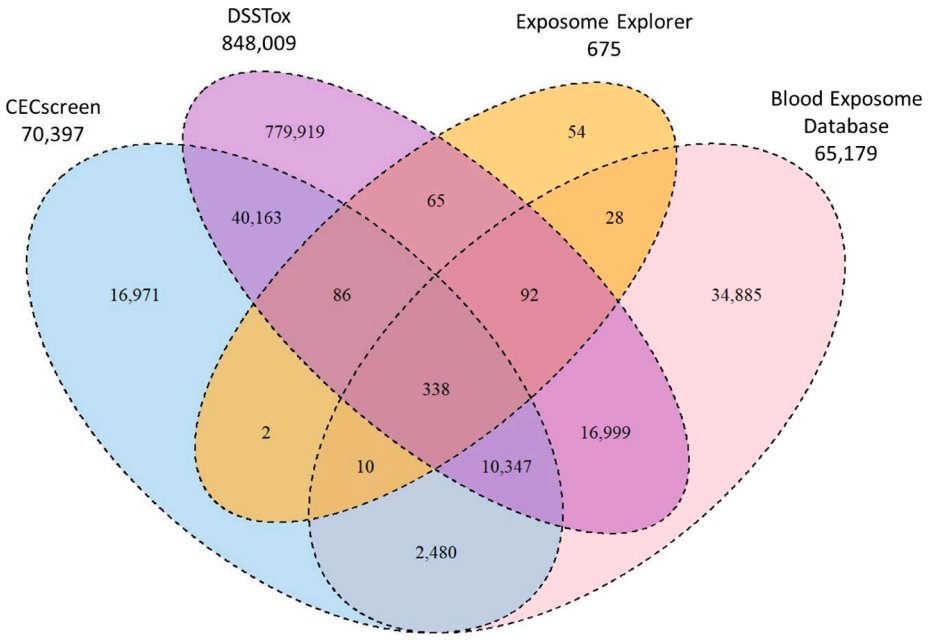

Fig. 4. Venn diagram of the overlapping unique InChIKeys between CECscreen, DSSTox, Exposome Explorer and Blood Exposome Database. These databases contain both CECs and endogenous metabolites.

CECscreen, which are both of approximately equal size, illustrates the added value of these aggregation approaches.

A strength of CECscreen is the inclusion of "MS- and QSAR-ready" structures as well as the addition of simulated Phase I metabolites to ensure applicability for screening in human biological samples using non-targeted HRMS. Furthermore, monoisotopic mass and the exact masses of several common adduct species such as $[M+H]^+$, $[M+Na]^+$, $[M+NH_4]^+$, $[M-H]^-$ and $[M+CH_3COO]^-$ are included for ease of use in several workflows. CECscreen also provides structural information as SMILES, which can be used to generate additional adduct species if required (Schymanski, 2019). Finally, the inclusion of metadata regarding physicochemical properties, environmental fate, and toxicity allows for prioritization of CECs of interests most appropriate for the research question, and/or to enhance the annotation confidence. For instance, selection could be based on the chemical properties that would make it active in humans according to Lipinski's rule of five (Hann and Kesër, 2012) or the probability of toxicity. Furthermore, information such as log P or log D can be used in retention time prediction under certain conditions and can potentially increase the annotation confidence (Aalizadeh et al., 2019). Metadata terms should be used with caution during prioritization, however, as the absence of toxicity information does not necessarily imply a lack of toxicity. Furthermore, due to the limitation of the models, there is no distinction between enantiomers which might behave differently.

Additional approaches that reduce data dimensionality using endpoint information can also be used for chemical prioritization, and to limit the number of features requiring annotation in a study. For instance, an EDA approach (Jonker et al., 2019; Zwart et al., 2020) can be used to extract only those features that are related to a measured biological effect of interest or by applying variable selection approaches to select features associated with a toxicological mechanism (Warth et al., 2017; Shen et al., 2014). Both of these techniques greatly reduce the total amount of features to be considered and therefore facilitate compound annotation. For an illustration of such an application of CECscreen in an annotation approach, see *Figure SM-2*. There are many other prioritization approaches available, e.g. signal intensity, characteristic isotope pattern or selection based on functional groups (Hollender et al., 2017). However, careful consideration should be given to the type of approach to be used to reach optimal efficiency. For instance, prioritization based on the highest peak intensities would not be optimal to screen for lower intensity CECs in a matrix with a high intensity of endogenous markers (Andra et al., 2017).

CECscreen and CECscreen_Metabolite_DB are large: by aggregating all available information in various existing databases, the number of CECs is high and this number is further expanded due to the inclusion of simulated Phase I metabolites. As a result, the screening process will produce a large number of annotations including a high fraction of false positives. Compound identification requires comparison of multiple orthogonal properties, including retention time, accurate mass and MS/MS spectra, to authentic standards analysed under the same analytical conditions (Schymanski et al., 2014). This process is costly and cumbersome and identification of all annotated compounds is not feasible. However, several approaches are available to improve the likelihood of correctly annotating features, which reduces the number of false positives and can aid in the selection of which annotated compounds should be identified. The incorporation into MetFrag (see below) is one way of facilitating these steps.

The use of tandem mass spectral information especially appears necessary to increase annotation confidence. Several tandem mass spectral libraries are available online, such as MassBank (Horai et al., 2010) and METLIN (Guijas et al., 2018), which already contain MS/MS fragmentation information including different fragmentation conditions and collision modes for a large number of chemicals. In addition, several resources and initiatives have also started to aggregate MS/MS libraries for different systems and under different acquisition conditions, including the HBM4EU initiative, complementarity to the present work (Horai et al., 2010; Oberacher et al., 2020). Despite these repositories and efforts, collection of MS/MS spectra for chemicals for which no standards are available, for instance, CECs metabolites produced by the human body, remains difficult. In order to get more insight in these types of compounds, *in silico* generation of MS/MS spectra approaches have been suggested (Chao et al., 2020; Dührkop et al., 2020; Dührkop et al., 2019). The *in silico* fragmenter MetFrag can combine database searches and MS/MS fragmentation prediction together and in this regard increase annotation confidence (Ruttkies et al., 2016). To facilitate this process, CECscreen is directly incorporated into MetFrag under the names of "HBM4EU_CECscreen_MF_Jul2020" and "HBM4EU_CECscreen_MF_Jul2020plusTPs" without and with Phase I metabolites,

respectively.

Several further steps can be undertaken in an attempt to reduce the percentage of false positive annotations. Based on the features generated by LC-HRMS, these include the use of information on retention time characteristics, mass defect, and isotope/adduct patterns (Scheubert et al., 2013) to characterize modules of features with correlated intensity levels found by hierarchical clustering (Uppal et al., 2017). If sufficient pathway level information is available, such clusters can be assessed for pathway level correlations to increase the annotations confidence for correlated features present in the same cluster. The approach has been described successfully with databases that contain pathways of endogenous metabolites such as KEGG (Uppal et al., 2017). Similar databases containing information on CECs metabolic pathways would need to be developed to apply this approach on a large scale.

The effective application of this database in screening human biological samples for CECs is dependent on the compartment that is measured. Phase I metabolites might be detected more frequently in blood samples whereas the metabolites detected in urine are often the result of Phase II metabolism. The BioTransformer software used to simulate the Phase I metabolites also includes a module to create Phase II metabolites, which can be easily applied to the "QSAR-ready" structures included in the CECscreen (Djoumbou-Feunang et al., 2019), however, considering the possibility that multiple Phase II metabolites are simulated from the Phase I metabolites, this will enlarge the volume of the database dramatically and could slow down the annotation procedure significantly. Furthermore, there are uncertainties in the validity of the metabolites as these were the product of simulation and not measurement. For example, multiple metabolites were simulated for the same compound but no information was provided on the likeliness of occurrence. As a result, some predicted metabolites may be detected more frequently, whilst others might not occur at all. Therefore, when available, CAS numbers and compound names were retrieved for the simulated metabolites in order to provide users a means to perform literature searches to determine if an annotated match is feasible. However, subsequent steps are still necessary to confirm the annotated metabolite (e.g. *in vitro* metabolism studies (Jia and Liu, 2007), MS/MS mass shifts (Boix et al., 2016)).

BioTransformer covers a wide range of chemical substrates including xenobiotics and different metabolic biotransformation reactions. Therefore, the simulated Phase I metabolites is considered to be an appropriate first step to facilitate CECs metabolite identification in exposome research (Djoumbou-Feunang et al., 2019). Initiatives are in progress that measure and document existing metabolites of a wide range of CECs such as the activities within the US EPA, NORMAN-SLE and PubChem (US EPA, 2004; LCSB-ECI, 2020). These measured metabolites should in time also be included in CECscreen. The presence of predicted metabolites in annotation databases such as CECscreen will help finding new metabolites in samples, which in turn will add to the knowledge on existing metabolites.

The approach described here resulted in the first version of the CECscreen database tailored for annotation of CECs in human samples and offers users metadata to prioritize compounds based on their own research question. Most of the generated metadata was the result of prediction models and simulations. As the understanding of CECs increases over time, actual measured values should be added to the databases. The addition of information on compound class, usage and production volumes would also be a valuable contribution to the database for prioritization and application (Fischer, 2017). Furthermore, including adduct species information that are specific for certain LC-HRMS methods is valuable as well, although can cause file size inflation. All information required to produce these species are included in the database. The current list is mainly focused on LC-HRMS types of data acquisitions, but can be extended for GC-HRMS. The chemical space investigated with GC–MS can also be important in the scope of studying the exposome, especially to investigate the link between volatile parent compounds and their metabolites with health effects.

4.1. Hosting

CECscreen, CECscreen_Metabolite_DB, CECscreen_CompTox_DB and CECscreen_OPERA_Predictions are hosted by the NORMAN Suspect List Exchange (NORMAN, 2015) and also available on Zenodo (Meijer et al., 2020) via DOI: https://doi.org/10.5281/zenodo.3956586. The database is also included in the CompTox Chemicals Dashboard (Williams et al., 2017)(https://comptox.epa.gov/dashboard/chemical_lists/CECSCREEN), PubChem via the NORMAN-SLE Classification Browser (https://pubchem.ncbi.nlm.nih.gov/classification/#hid=101) and is incorporated in MetFrag (https://msbi.ipb-halle.de/MetFrag/) (Ruttkies et al., 2016).

CRediT authorship contribution statement

Jeroen Meijer: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing - original draft, Visualization, Writing - review & editing. **Marja Lamoree:** Conceptualization, Writing - review & editing. **Timo Hamers:** Conceptualization, Writing - review & editing. **Jean-Philippe Antignac:** Conceptualization, Writing - review & editing, Supervision, Project administration. **Sébastien Hutinet:** Data curation, Validation. **Laurent Debrauwer:** Writing - review & editing. **Adrian Covaci:** Writing - review & editing. **Carolin Huber:** Writing - review & editing. **Martin Krauss:** Writing - review & editing, Data curation. **Douglas I. Walker:** Writing - review & editing. **Emma L. Schymanski:** Writing - review & editing, Validation, Resources, Data curation. **Roel Vermeulen:** Conceptualization, Writing - review & editing, Supervision. **Jelle Vlaanderen:** Conceptualization, Methodology, Data curation, Investigation, Supervision, Project administration, Writing - original draft, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

JM, ML, TH, JA, SH, LD, AC, CH, MK, RV and JV acknowledge financial support by the HBM4EU project funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 733032. In addition, JV and RV acknowledge funding by EXPANSE (EC H2020, grant agreement No 874627) and EXPOSOME-NL. EXPOSOME-NL is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.004.017). ELS acknowledges funding support from the Luxembourg National Research Fund (FNR) for project A18/BM/12341006. Support was also provided to DIW by the US National Institute of Health, award number U2C ES030859. ELS thanks Evan Bolton, Jeff Zhang, Paul Thiessen (PubChem), Antony Williams (US EPA), Steffen Neumann (IPB Halle) and Natalia Glowacka for their assistance in the PubChem, CompTox, MetFrag and NORMAN-SLE website integration of CECscreen, respectively.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2021.106511.

References

Aalizadeh, R., Nika, M.C., Thomaidis, N.S., 2019. Development and application of retention time prediction models in the suspect and non-target screening of emerging contaminants. J. Hazard. Mater. https://doi.org/10.1016/j.jhazmat.2018.09.047.

Alygizakis, N., 2016. S6|ITNANTIBIOTIC|Antibiotic List. ITN MSCA ANSWER. https://doi.org/10.5281/zenodo.2621957.

Alygizakis, N., Samanipour, S., Thomas, K., 2017. S12|NORMANEWS|NormaNEWS for Retrospective Screening of New Emerging Contaminants. https://doi.org/10.5281/zenodo.2623816.

Alygizakis, N.A., Samanipour, S., Hollender, J., et al., 2018. Exploring the potential of a global emerging contaminant early warning network through the use of retrospective suspect screening with high-resolution mass spectrometry. Environ. Sci. Technol. 52 (9), 5135–5144. https://doi.org/10.1021/acs.est.8b00365.

Andra, S.S., Austin, C., Patel, D., Dolios, G., Awawda, M., Arora, M., 2017. Trends in the application of high-resolution mass spectrometry for human biomonitoring: an analytical primer to studying the environmental chemical space of the human exposome. Environ. Int. 100, 32–61. https://doi.org/10.1016/j.envint.2016.11.026.

Bade, R., Schymanski, E., 2015. S4|UJIBADE|University of Jaume I Bade et al List. November 2015. https://doi.org/10.5281/zenodo.2621917.

Bade, R., Bijlsma, L., Miller, T.H., Barron, L.P., Sancho, J.V., Hernández, F., 2015. Suspect screening of large numbers of emerging contaminants in environmental waters using artificial neural networks for chromatographic retention time prediction and high resolution mass spectrometry data analysis. Sci. Total Environ. 538, 934–941. https://doi.org/10.1016/j.scitotenv.2015.08.078.

Barupal, D.K., Fiehn, O., 2019. Generating the blood exposome database using a comprehensive text mining and database fusion approach. Environ. Health Perspect. https://doi.org/10.1289/EHP4713.

Beisken, S., Meinl, T., Wiswedel, B., de Figueiredo, L.F., Berthold, M., Steinbeck, C., 2013. KNIME-CDK: workflow-driven cheminformatics. BMC Bioinf. https://doi.org/10.1186/1471-2105-14-257.

Berthold, M., Cebron, N., Dill, F., et al., 2009. KNIME – the Konstanz information miner: Version 2.0 and Beyond. SIGKDD Explor. 11, 26–31.

Boix, C., Ibáñez, M., Bagnati, R., et al., 2016. High resolution mass spectrometry to investigate omeprazole and venlafaxine metabolites in wastewater. J. Hazard. Mater. https://doi.org/10.1016/j.jhazmat.2015.09.059.

Chao, A., Al-Ghoul, H., McEachran, A.D., et al., 2020. In silico MS/MS spectra for identifying unknowns: a critical examination using CFM-ID algorithms and ENTACT mixture samples. Anal. Bioanal. Chem. 412 (6), 1303–1315. https://doi.org/10.1007/s00216-019-02351-7.

ChemAxon. Marvin. 2014. http://www.chemaxon.com/products/marvin/marvinsketch/%0A%0A.

Djoumbou-Feunang, Y., Fiamoncini, J., Gil-de-la-Fuente, A., Greiner, R., Manach, C., Wishart, D.S., 2019. BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. J. Cheminform. 11 (1), 1–25. https://doi.org/10.1186/s13321-018-0324-5.

Dührkop, K., Nothias, L.-F., Fleischauer, M., et al., 2020. Classes for the masses: systematic classification of unknowns using fragmentation spectra. bioRxiv. 2020.04.17.046672. https://doi.org/10.1101/2020.04.17.046672.

Dührkop, K., Fleischauer, M., Ludwig, M., et al., 2019. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. Nat. Methods. https://doi.org/10.1038/s41592-019-0344-8.

Dulio, V., 2017. S15|NORMANPRI|NORMAN. Priority List. https://doi.org/10.5281/zenodo.2624273.

Dulio, V., Aalizadeh, R., 2017. S16|FRENCHLIST|French. Monitor. List. https://doi.org/10.5281/zenodo.2624325.

ECHA, 2007. PBT/vPvB assessments under the previous EU chemicals legislation. https://echa.europa.eu/information-on-chemicals/pbt-vpvb-assessments-under-the-previous-eu-chemicals-legislation. Published 2007 (accessed October 23, 2017).

ECHA, 2017. Candidate List of substances of very high concern for Authorisation. https://echa.europa.eu/candidate-list-table (accessed October 23, 2017).

ECHA, 2017. Annex III inventory. https://echa.europa.eu/information-on-chemicals/annex-iii-inventory (accessed October 23, 2017).

EFSA, 2017. Chemical hazards data – OpenFoodTox. https://www.efsa.europa.eu/en/data/chemical-hazards-data. Published 2002 (accessed October 23, 2017).

US EPA. Select List. https://comptox.epa.gov/dashboard/chemical_lists (accessed November 3, 2017).

US EPA, 2017. Chemical and Products Database (CPDat). https://www.epa.gov/chemical-research/chemical-and-products-database-cpdat (accessed October 23, 2017).

European Commission, 2017. Cosmetic ingredient database. https://ec.europa.eu/growth/sectors/cosmetics/cosing_en (accessed October 23, 2017).

European Commission, 2017. Which substances are of concern? https://ec.europa.eu/environment/chemicals/endocrine/strategy/substances_en.htm#priority_list (accessed October 23, 2017).

Fischer, S., 2017. S17|KEMIMARKET|KEMI. Market List. https://doi.org/10.5281/zenodo.3959394.

Fischer, S., 2016. KEMI Market List: Organic Chemicals Potentially Identified on the EU Market.

Fischer, S., 2017. S14|KEMIPFAS|PFAS Highly Fluorinated Substances List: KEMI. https://doi.org/10.5281/zenodo.3544805.

Gago Ferrero, P., 2016. S8|ATHENSSUS|University of Athens Surfactants and Suspects List. https://doi.org/10.5281/zenodo.2621980.

Gago-Ferrero, P., Schymanski, E.L., Bletsou, A.A., Aalizadeh, R., Hollender, J., Thomaidis, N.S., 2015. Extended suspect and non-target strategies to characterize emerging polar organic contaminants in raw wastewater with LC-HRMS/MS. Environ. Sci. Technol. 49 (20), 12333–12341. https://doi.org/10.1021/acs.est.5b03454.

Grulke, C.M., Williams, A.J., Thillanadarajah, I., Richard, A.M., 2019. EPA's DSSTox database: history of development of a curated chemistry resource supporting computational toxicology research. Comput. Toxicol. https://doi.org/10.1016/j.comtox.2019.100096.

Guha, R., 2007. Chemical informatics functionality in R. J. Stat. Softw. https://doi.org/10.18637/jss.v018.i05.

Guijas, C., Montenegro-Burke, J.R., Domingo-Almenara, X., et al., 2018. METLIN: a technology platform for identifying knowns and unknowns. Anal. Chem. https://doi.org/10.1021/acs.analchem.7b04424.

Hann, M.M., Kesër, G.M., 2012. Finding the sweet spot: the role of nature and nurture in medicinal chemistry. Nat. Rev. Drug. Discov. https://doi.org/10.1038/nrd3701.

Højsgaard, S., Halekoh, U., 2006. Groupwise Statistics. LSmeans, Linear Contrasts, Utilities.

Hollender, J., Schymanski, E.L., Singer, H.P., Ferguson, P.L., 2017. Nontarget screening with high resolution mass spectrometry in the environment: ready to go? Environ. Sci. Technol. https://doi.org/10.1021/acs.est.7b02184.

Horai, H., Arita, M., Kanaya, S., et al., 2010. MassBank: a public repository for sharing mass spectral data for life sciences. J. Mass Spectrom. 45 (7), 703–714. https://doi.org/10.1002/jms.1777.

IOMC, 2009. The 2007 Oecd list of high production volume chemicals. Ser Test Assess. 112 (112), 1–104. ENV/JM/MONO(2007)10.

Jia, L., Liu, X., 2007. The conduct of drug metabolism studies considered good practice (II) vitro experiments. Curr. Drug Metab. https://doi.org/10.2174/138920007782798207.

Jonker, W., de Vries, K., Althuisius, N., et al., 2019. Compound identification using liquid chromatography and high-resolution noncontact fraction collection with a solenoid valve. SLAS Technol. 24 (6), 543–555. https://doi.org/10.1177/2472630319848768.

Kanehisa, M., Goto, S., 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. https://doi.org/10.1093/nar/28.1.27.

Kim, S., Chen, J., Cheng, T., et al., 2019. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res. https://doi.org/10.1093/nar/gky1033.

Landrigan, P.J., Fuller, R., Acosta, N.J.R., et al., 2018. The Lancet Commission on pollution and health. Lancet 391 (10119), 462–512. https://doi.org/10.1016/S0140-6736(17)32345-0.

LCSB-ECI, Krier, J., Schymanski, E., et al., 2020. S68|HSDBTPS|Transformation Products Extracted from HSDB Content in PubChem. https://doi.org/10.5281/zenodo.3890392.

Letzel, T., Grosse, S., Sengel, M., 2017. S2|STOFFIDENT|HSWT/LfU STOFF-IDENT Database of Water-Relevant Substances. September 2017. https://doi.org/10.5281/zenodo.3900133.

Little, J.L., Williams, A.J., Pshenichnov, A., Tkachenko, V., 2012. Identification of "known unknowns" utilizing accurate mass data and chemspider. J. Am. Soc. Mass Spectrom. https://doi.org/10.1007/s13361-011-0265-y.

Loos, M., Gerber, C., Corona, F., Hollender, J., Singer, H., 2015. Accelerated isotope fine structure calculation using pruned transition trees. Anal. Chem. https://doi.org/10.1021/acs.analchem.5b00941.

Mansouri, K., Grulke, C.M., Richard, A.M., Judson, R.S., Williams, A.J., 2016. An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modelling. SAR QSAR Environ. Res. 27 (11), 911–937. https://doi.org/10.1080/1062936X.2016.1253611.

Mansouri, K., Grulke, C.M., Judson, R.S., Williams, A.J., 2018. OPERA models for predicting physicochemical properties and environmental fate endpoints. J. Cheminform. 10 (1), 10. https://doi.org/10.1186/s13321-018-0263-1.

Mansouri, K., 2019. kmansouri/MS-ready. https://github.com/kmansouri/MS-ready. Published 2017 (accessed August 19, 2019).

Mansouri, K., 2017. kmansouri/QSAR-ready. https://github.com/kmansouri/QSAR-ready. Published 2017 (accessed August 19, 2019).

McEachran, A.D., Mansouri, K., Grulke, C., Schymanski, E.L., Ruttkies, C., Williams, A.J., 2018. "MS-Ready" structures for non-targeted high-resolution mass spectrometry screening studies. J. Cheminform. 10 (1), 45. https://doi.org/10.1186/s13321-018-0299-2.

Meijer, J., Lamoree, M., Hamers, T., et al., 2020. S71|CECSCREEN|HBM4EU CECscreen: Screening List for Chemicals of Emerging Concern Plus Metadata and Predicted Phase 1 Metabolites. https://doi.org/10.5281/zenodo.3956587.

Mistrik, R., Aligizakis, N., 2019. S19|MZCLOUD|mzCloud Compounds. April 2019. https://doi.org/10.5281/zenodo.3542104.

Moschet, C., 2017. S11|SWISSPEST|Swiss Insecticides. Fungicides and TPs. https://doi.org/10.5281/zenodo.2623741.

Moschet, C., Piazzoli, A., Singer, H., Hollender, J., 2013. Alleviating the reference standard dilemma using a systematic exact mass suspect screening approach with liquid chromatography-high resolution mass spectrometry. Anal. Chem. 85 (21), 10312–10320. https://doi.org/10.1021/ac4021598.

Network, N., Aalizadeh, R., Alygizakis, N., et al., 2020. S0|SUSDAT|Merged NORMAN Suspect List: SusDat. https://doi.org/10.5281/zenodo.3900203.

Neveu, V., Moussy, A., Rouaix, H., et al., 2017. Exposome-Explorer: a manually-curated database on biomarkers of exposure to dietary and environmental factors. Nucleic Acids Res. https://doi.org/10.1093/nar/gkw980.

NORMAN, 2017. NORMAN Suspect List Exchange. https://www.norman-network.com/nds/SLE/. Published 2015 (accessed May 24, 2017).

O'Boyle, N.M., Banck, M., James, C.A., Morley, C., Vandermeersch, T., Hutchison, G.R., 2011. Open babel: an open chemical toolbox. J Cheminform. 3 (1), 33. https://doi.org/10.1186/1758-2946-3-33.

Oberacher, H., Sasse, M., Antignac, J.P., et al., 2020. A European proposal for quality control and quality assurance of tandem mass spectral libraries. Environ. Sci. Eur. 32 (1) https://doi.org/10.1186/s12302-020-00314-9.

Paulus, G.K., Hornstra, L.M., Alygizakis, N., Slobodnik, J., Thomaidis, N., Medema, G., 2019. The impact of on-site hospital wastewater treatment on the downstream communal wastewater system in terms of antibiotics and antibiotic resistance genes. Int. J. Hyg. Environ. Health 222 (4), 635–644. https://doi.org/10.1016/j.ijheh.2019.01.004.

Pence, H.E., Williams, A., 2010. Chemspider: an online chemical information resource. J. Chem. Educ. https://doi.org/10.1021/ed100697w.
Pourchet, M., Debrauwer, L., Klanova, J., et al., 2020. Suspect and non-targeted screening of chemicals of emerging concern for human biomonitoring, environmental health studies and support to risk assessment: from promises to challenges and harmonisation issues. Environ. Int. https://doi.org/10.1016/j.envint.2020.105545.
Pourchet, M., Narduzzi, L., Jean, A., et al., 2021. Non-targeted screening methodology to characterise human internal chemical exposure: application to halogenated compounds in human milk. Talanta 225, 121979. https://doi.org/10.1016/j.talanta.2020.121979.
Rostkowski, P., Fischer, S., 2017. S20|BISPHENOLS|Bisphenols. September 2017. https://doi.org/10.5281/zenodo.3779854.
Ruttkies, C., Schymanski, E.L., Wolf, S., Hollender, J., Neumann, S., 2016. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. J Cheminform. https://doi.org/10.1186/s13321-016-0115-9.
Sahigara, F., Mansouri, K., Ballabio, D., Mauri, A., Consonni, V., Todeschini, R., 2012. Comparison of different approaches to define the applicability domain of QSAR models. Molecules. https://doi.org/10.3390/molecules17054791.
Sauvé, S., Desrosiers, M., 2014. A review of what is an emerging contaminant. Chem. Cent. J. 8 (1), 15. https://doi.org/10.1186/1752-153X-8-15.
Scheubert, K., Hufsky, F., Ocker, S.B., 2013. REVIEW Open Access Computational Mass Spectrometry for Small Molecules.
Schymanski, E., 2014. S7|EAWAGSURF|eawag surfactants. Suspect List. https://doi.org/10.5281/zenodo.3549934.
Schymanski, E.L., Jeon, J., Gulde, R., et al., 2014. Identifying small molecules via high resolution mass spectrometry: communicating confidence. Environ. Sci. Technol. https://doi.org/10.1021/es5002105.
Schymanski, E., Schulze, T., Alygizakis, N., Meier, R., 2019. S1|MASSBANK|NORMAN. Compd. MassBank. https://doi.org/10.5281/zenodo.3247682.
Schymanski, E.L., Singer, H.P., Longrée, P., et al., 2014. Strategies to characterize polar organic contamination in wastewater: exploring the capability of high resolution mass spectrometry. Environ. Sci. Technol. 48 (3), 1811–1818. https://doi.org/10.1021/es4044374.
Schymanski, E.L., Singer, H.P., Slobodnik, J., et al., 2015. Non-target screening with high-resolution mass spectrometry: critical review using a collaborative trial on water analysis. Anal. Bioanal. Chem. 407 (21), 6237–6255. https://doi.org/10.1007/s00216-015-8681-7.
Schymanski, E., 2016. S3|NORMANCT15|NORMAN Collaborative Trial Targets and Suspects. https://doi.org/10.5281/zenodo.3723469.
Schymanski, E.L., 2019. RChemMass: Various Cheminformatic, Curation and Mass Spectrometry Functions. https://github.com/schymane/RChemMass.
Shen, H., Xu, W., Peng, S., et al., 2014. Pooling samples for "top-down" molecular exposomics research: the methodology. Environ. Heal A Glob. Access Sci. Source 13 (1). https://doi.org/10.1186/1476-069X-13-8.
Singer, H.P., Wössner, A.E., McArdell, C.S., Fenner, K., 2016. Rapid screening for exposure to "Non-Target" pharmaceuticals from wastewater effluents by combining HRMS-based suspect screening and exposure modeling. Environ. Sci. Technol. 50 (13), 6698–6707. https://doi.org/10.1021/acs.est.5b03332.
Sjerps, R., 2016. S5|KWRSJERPS|KWR drinking water. Suspect List. https://doi.org/10.5281/zenodo.2621942.
Sjerps, R.M.A., Vughs, D., van Leerdam, J.A., ter Laak, T.L., van Wezel, A.P., 2016. Data-driven prioritization of chemicals for various water types using suspect screening LC-HRMS. Water Res. 93, 254–264. https://doi.org/10.1016/j.watres.2016.02.034.
Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., Willighagen, E., 2003. The Chemistry Development Kit (CDK): an open-source Java library for chemo- and bioinformatics. J. Chem. Inf. Comput. Sci. https://doi.org/10.1021/ci025584y.
Stravs, M.A., Schymanski, E.L., Singer, H.P., Hollender, J., 2013. Automatic recalibration and processing of tandem mass spectra using formula annotation. J. Mass Spectrom. 48 (1), 89–99. https://doi.org/10.1002/jms.3131.
Szöcs, E., Stirling, T., Scott, E.R., Scharmüller, A., Schäfer, R.B., 2020. Webchem: an R package to retrieve chemical information from the web. J. Stat. Softw. https://doi.org/10.18637/jss.v093.i13.
Team R., 2013. The R Project for Statistical Computing. https://doi.org/10.1159/000323281.
TEDX. Search the TEDX List. https://endocrinedisruption.org/interactive-tools/tedx-list-of-potential-endocrine-disruptors/about-the-tedx-list (accessed October 23, 2017).
TMIC, 2017. The Toxin and Toxin Target Database (T3DB). http://www.t3db.ca/. Published 2014 (accessed October 23, 2017).
Trier, X., Lunderberg, D., 2015. S9|PFASTRIER|PFAS Suspect List: fluorinated substances. November 2015. https://doi.org/10.5281/zenodo.3542121.
United States Environmental Protection Agency (US EPA). US EPA CompTox Chemistry Dashboard. https://comptox.epa.gov/dashboard (accessed May 5, 2020).
Uppal, K., Walker, D.I., Jones, D.P., 2017. xMSannotator: an R package for network-based annotation of high-resolution metabolomics data. Anal. Chem. 89 (2), 1063–1067. https://doi.org/10.1021/acs.analchem.6b01214.
US EPA. Biotransformation and ToxCast. https://cfpub.epa.gov/si/si_public_record_report.cfm?Lab=NCCT&TIMSType=&count=10000&dirEntryId=211445&searchAll=&showCriteria=2&simpleSearch=0&startIndex=20001. Published 2004. (accessed May 1, 2020).
Vermeulen, R., Schymanski, E.L., Barabási, A.L., Miller, G.W., 2020. The exposome and health: Where chemistry meets biology. Science (80-) 367 (6476), 392–396. https://doi.org/10.1126/science.aay3164.
von der Ohe, P., Aalizadeh, R., 2020. S13|EUCOSMETICS|Combined Inventory of Ingredients Employed in Cosmetic Products (2000) and Revised Inventory (2006). https://doi.org/10.5281/zenodo.3959386.
Wang, R., Fu, Y., Lai, L., 1997. A new atom-additive method for calculating partition coefficients. J. Chem. Inf. Comput. Sci. 37 (3), 615–621. https://doi.org/10.1021/ci960169p.
Warth, B., Spangler, S., Fang, M., et al., 2017. Exposome-scale investigations guided by global metabolomics, pathway analysis, and cognitive computing. Anal. Chem. https://doi.org/10.1021/acs.analchem.7b02759.
Williams, A.J., Grulke, C.M., Edwards, J., et al., 2017. The CompTox chemistry dashboard: a community data resource for environmental chemistry. J Cheminform. https://doi.org/10.1186/s13321-017-0247-6.
Wishart, D., Arndt, D., Pon, A., et al., 2015. T3DB: the toxic exposome database. Nucleic Acids Res. 43 (D1), D928–D934. https://doi.org/10.1093/nar/gku1004.
Wishart, D.S., Tzur, D., Knox, C., et al., 2007. HMDB: The human metabolome database. Nucleic Acids Res. 35 (SUPPL. 1), 521–526. https://doi.org/10.1093/nar/gkl923.
Wishart, D.S., Feunang, Y.D., Marcu, A., et al., 2018. HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Res. https://doi.org/10.1093/nar/gkx1089.
Wössner, A., Singer, H., 2017. S10|SWISSPHARMA|pharmaceutical list with. Consump. Data. https://doi.org/10.5281/zenodo.2623485.
Zwart, N., Jonker, W., ten Broek, R., et al., 2020. Identification of mutagenic and endocrine disrupting compounds in surface water and wastewater treatment plant effluents using high-resolution effect-directed analysis. Water Res. 168, 115204. https://doi.org/10.1016/j.watres.2019.115204.