Letter to the Editor: FAIR-ifying the Exposome Journal: Templates for Chemical Structures and Transformations

Emma L. Schymanski[1]* and Evan E. Bolton[2]*

[1]Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 6 avenue du Swing, 4367, Belvaux, Luxembourg. ORCID 0000-0001-6868-8145

[2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, 20894, USA. ORCID 0000-0002-5959-6190

* Correspondence: ELS: emma.schymanski@uni.lu and EEB: bolton@ncbi.nlm.nih.gov

Running title: FAIRifying Chemical Structures and Transformations

Abstract

The exposome, the totality of lifetime exposures, is a new and highly complex paradigm for health and disease. Tackling this challenge requires an effort well beyond single individuals or laboratories, where every piece of the puzzle will be vital. The launch of this new Exposome journal coincides with the evolution of the exposome through its teenage years and into a growing maturity in an increasingly open and FAIR (***findable, accessible, interoperable, reusable***) world. This letter discusses how both authors and the Exposome journal alike can help increase the ***FAIR***ness of the chemical structural information and the associated metadata in the journal, aiming to capture more details about the chemistry of exposomics. The proposed chemical structure template can serve as an ***interoperable*** supplementary format that is made ***accessible*** through the website and more ***findable*** by linking the DOI of this data file to the article DOI metadata, supporting further ***reuse***. An additional Transformations template provides authors with a means to connect predecessor (parent, substrate) molecules to successor (transformation product, metabolite) molecules and thus provide ***FAIR*** connections between observed (*i.e.,* experimental) chemical exposures and biological responses, to help improve the public knowledgebase on exposome-related transformations. These connections are vital to extend current biochemical knowledge and to fulfil the current Exposome definition of “the cumulative measure of environmental influences and associated biological responses throughout the lifespan including exposures from the environment, diet, behaviour, and endogenous processes”.

Keywords

Open science, chemical information, FAIR, transformation products, data workflows, data sharing

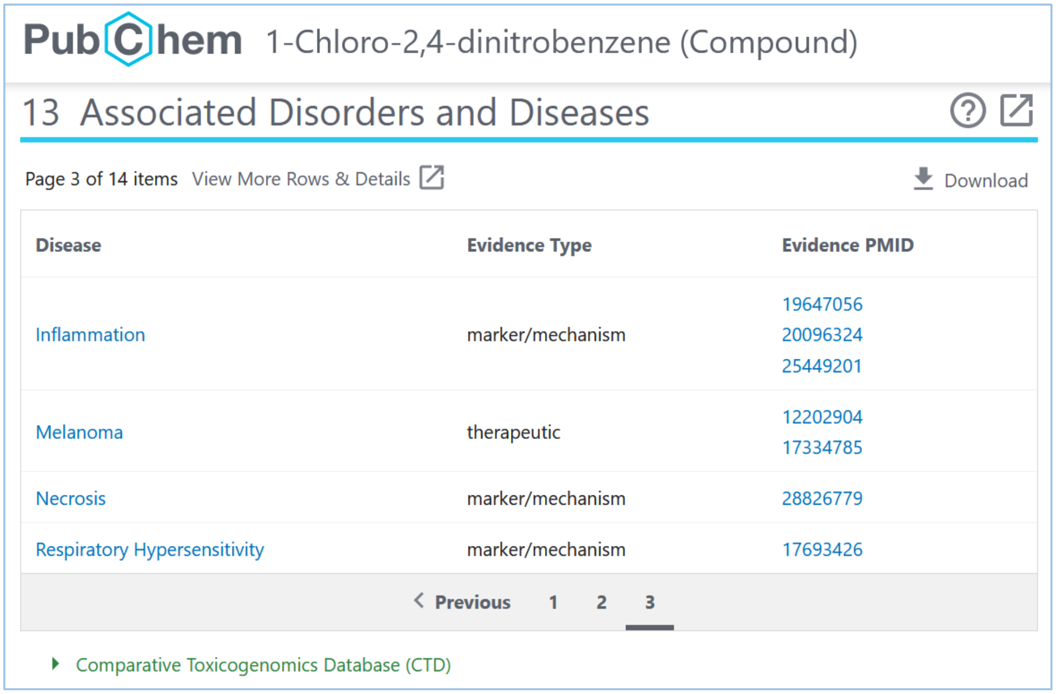
Main Text

Motivation

The “exposome” is a concept first mentioned in 2005 by Wild[1] to offer an environmental complement to the genome[2] in considering health and disease. Now that the exposome is in its adolescence and “emerging from the primordial swamp” sufficiently to warrant its own journal[2], it is a good time to reflect on what steps are required to enable exposomics to mirror the achievements of genomics. A quick search reveals, for instance, that global investment in genomics is projected into the tens of billions in the coming years[3,4], while the global investment in the exposome or exposomics is rather of the order of tens of millions. Yet, exposomics is an extraordinarily complex paradigm that will certainly require concerted global effort comparable to that of the human genome[5]. Although capturing “the cumulative measure of environmental influences and associated biological responses throughout the lifespan including exposures from the environment, diet, behaviour, and endogenous processes”[6] may seem unachievable for some, sequencing the human genome was also considered an almost impossible task only a few decades ago. While the success of genomics is arguably due to many factors (including extensive investment), one very significant factor in its success is the open exchange of genomics data and the ecosystem of open resources that has been built around genomics, enabling scientists around the world to achieve extraordinary progress in a relatively short time. Can exposomics achieve the same?

With this letter, we provide some perspectives and guidance on how both authors of articles in Exposome and the Exposome journal itself can contribute to the cumulative efforts needed to tackle the exposomics challenge from a chemical information and chemical informatics standpoint. Exposomics is inherently a data-driven discipline. The interlinking of chemical, disease and reference information is already providing support to exposomics efforts, as shown in Figure 1 using an examples from PubChem[7] and the Comparative Toxicogenomics Database (CTD)[8], as well as from the CompTox Chemicals Dashboard[9,10]. Such information gathering and cross-resource integration efforts are much easier if data is both open and ***FAIR*** (***findable, accessible, interoperable, reusable***). Providing guidance and coordinating at a journal level is one way to enable such information gathering; genomics data deposition is mandated in most major journals and this has been key to building the open genomics data resources that are so critical for food-based pathogen surveillance, COVID-19 disease variant tracking, and so much more. If sufficient information for exposomics was available, what can we as a community achieve?

Authors need guidance to properly and uniformly capture and report chemical structure information and transformations, *i.e.,* connecting either endogenous or exogenous chemicals with their metabolites – thus helping capture the associated biological responses. The flexible templates provided here (see sections “Chemical Structure Data” and “Transformations Data”) show how authors can consistently submit this information to the Exposome journal as supplementary materials with their articles. These templates are designed such that authors can include as much or as little information as is available, yet still contribute their knowledge and outcomes to the exposomics “pool” (and beyond) in an open and FAIR manner. The “Chemical Structure Data” template is identical to the template introduced recently in the Journal of Cheminformatics[11].



PubChem 1-Chloro-2,4-dinitrobenzene (Compound)

13 Associated Disorders and Diseases

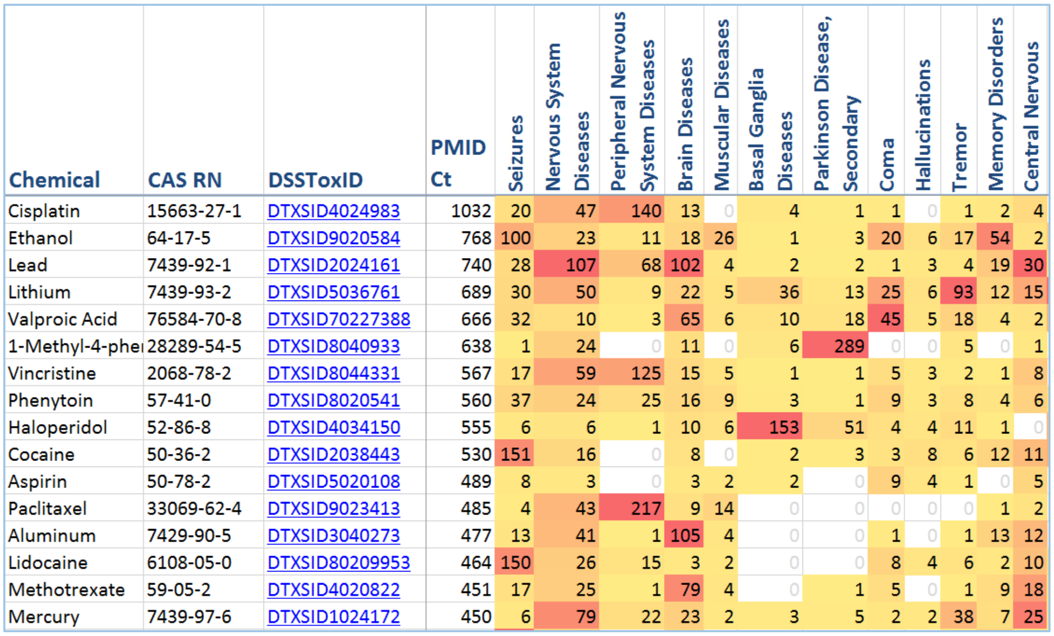
Page 3 of 14 items View More Rows & Details

Download

Disease	Evidence Type	Evidence PMID
Inflammation	marker/mechanism	19647056 20096324 25449201
Melanoma	therapeutic	12202904 17334785
Necrosis	marker/mechanism	28826779
Respiratory Hypersensitivity	marker/mechanism	17693426

< Previous 1 2 3

▸ Comparative Toxicogenomics Database (CTD)



Chemical	CAS RN	DSSToxID	PMID Ct	Seizures	Nervous System Diseases	Peripheral Nervous System Diseases	Brain Diseases	Muscular Diseases	Basal Ganglia Diseases	Parkinson Disease, Secondary	Coma	Hallucinations	Tremor	Memory Disorders	Central Nervous
Cisplatin	15663-27-1	DTXSID4024983	1032	20	47	140	13	0	4	1	1	0	1	2	4
Ethanol	64-17-5	DTXSID9020584	768	100	23	11	18	26	1	3	20	6	17	54	2
Lead	7439-92-1	DTXSID2024161	740	28	107	68	102	4	2	2	1	3	4	19	30
Lithium	7439-93-2	DTXSID5036761	689	30	50	9	22	5	36	13	25	6	93	12	15
Valproic Acid	76584-70-8	DTXSID70227388	666	32	10	3	65	6	10	18	45	5	18	4	2
1-Methyl-4-phe	28289-54-5	DTXSID8040933	638	1	24	0	11	0	6	289	0	0	5	0	1
Vincristine	2068-78-2	DTXSID8044331	567	17	59	125	15	5	1	1	5	3	2	1	8
Phenytoin	57-41-0	DTXSID8020541	560	37	24	25	16	9	3	1	9	3	8	4	6
Haloperidol	52-86-8	DTXSID4034150	555	6	6	1	10	6	153	51	4	4	11	1	0
Cocaine	50-36-2	DTXSID2038443	530	151	16	0	8	0	2	3	3	8	6	12	11
Aspirin	50-78-2	DTXSID5020108	489	8	3	0	3	2	2	0	9	4	1	0	5
Paclitaxel	33069-62-4	DTXSID9023413	485	4	43	217	9	14	0	0	0	0	0	1	2
Aluminum	7429-90-5	DTXSID3040273	477	13	41	1	105	4	0	0	1	0	1	13	12
Lidocaine	6108-05-0	DTXSID80209953	464	150	26	15	3	2	0	0	8	4	6	2	10
Methotrexate	59-05-2	DTXSID4020822	451	17	25	1	79	4	0	1	5	0	1	9	18
Mercury	7439-97-6	DTXSID1024172	450	6	79	22	23	2	3	5	2	2	38	7	25

Figure 1: FAIRifying and opening up exposomics information is critical to "big data" exposomics, empowering information discovery and cross-resource integration. Top (A): associated disorders and diseases (and references) for a single chemical, 1-chloro-2,4-dinitrobenzene in PubChem[7], with information sourced from the Comparative Toxicogenomics Database (CTD)[8]. Source: https://pubchem.ncbi.nlm.nih.gov/compound/6#section=Associated-Disorders-and-Diseases. Bottom (B): Individual chemical – disease endpoint mappings via Name, Chemical Abstract Services Registry Numbers (CAS RN), CompTox Chemicals Dashboard identifiers (DSSToxID or DTXSIDs), plus total and endpoint-specific reference counts in the context of neurotoxicity, embedded in an excel macro[10,12].

An incredible amount of knowledge relevant for exposomics has already been gathered, yet current studies are based primarily on using public resources to find existing information. To extend exposomics into the future, we need to enable the discovery and reporting of new findings via rapid integration into public resources. Thus, author contributions, no matter how small, will gradually help build the bigger picture needed to unravel and comprehend the exposome. Before we launch into the template descriptions, a few definitions are covered in the next section.

Definitions

While "FAIR" and "Open" are used somewhat interchangeably in this article as we strongly believe that chemical data should be both where possible, there is a distinction that is particularly relevant for exposomics, as sensitive human data cannot necessarily be made open. Data can be "open" but not "FAIR", and vice versa. Open science has many facets; of most relevance to this article is open access. Open access (OA) is a set of principles and a range of practices through which research outputs are distributed online, free of cost or other access barriers[13]. The FAIR principles for digital assets, on the other hand, include guidance on how to make data more Findable, Accessible, Interoperable and Reusable[14,15]. For example, if you have open data that is not findable, no one can use it; whereas if you have "FAIR" data that is not "open", it is not available for integration into open community resources. Thus, the most powerful data is both open and FAIR

In Table 1, we provide some definitions of chemical and transformation terms used later in this article.

Table 1: Definition of chemical and transformation terms used in this article and/or templates.

Concept	***Definition***
Biosystem	The medium in which the predecessor is transformed into the successor (e.g., environment, human liver, etc.)
Identifier	An identifier or name that you (the author) have for a chemical structure
InChI	IUPAC International Chemical Identifier is a descriptor of a chemical structure[16]
InChIKey	A 27-character long, layered "hash" of an InChI[16]
PubChem CID	PubChem Compound Identifier
Predecessor	Substrate/parent that is transformed (somehow) into a successor product
SMILES	Chemical structure notation expressed as a string
Successor	Transformation product/metabolite resulting from transformation (somehow) of a substrate/parent

Templates for FAIR Exposomics Chemical Data

Chemical Structure Data

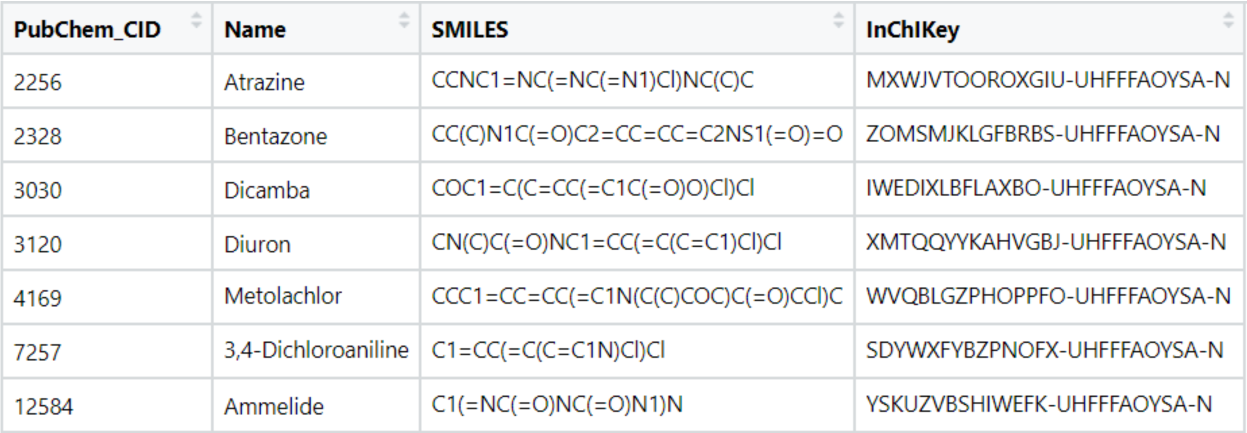
Better consideration of chemical factors in the exposome requires high-quality chemical information in research articles. Many exposomics resources are based (mostly) on literature mining using name and synonym matching, which can be notoriously prone to errors. In this section, we provide some guidance on what information authors should consider providing, as well as the pros and cons of various choices. Since this Chemical Structure Data template was presented recently to the Journal of Cheminformatics[11], some of the material in this section overlaps with the previous article.

Authors should consider submitting their chemical structure information with their manuscript as Supplementary Material using the suggested template as comma separated value (CSV; *.csv); or, alternatively, as tab-separated value (TSV; *.tsv) or structure data file (SDF; *.sdf) formats. These formats ensure maximum interoperability between resources and operating systems. The popular XLS(X) format is not truly interoperable (options to save as CSV or TSV are offered), while the extraction of information from PDF format is difficult without introducing errors. The content below describes the CSV/TSV formats, SDF instructions are available elsewhere[17] (however, the SD fields should match the CSV/TSV headers). In our experience, so far CSV often proves most interoperable for the widest audience, although the other formats also have certain advantages.

For CSV/TSV files, the header (first row) indicates the data content of each column; each subsequent row corresponds to a complete chemical record description: chemical structure, chemical names, identifiers, comments, and any other data the authors wish to provide (as additional columns). The interoperable case-insensitive template CSV/TSV column headers (or SDF SD fields) are: ***SMILES, InChI***, and ***InChIKey*** for chemical structure; ***Name*** and ***Synonym*** for chemical names; and ***Comment*** for textual comments. Any additional columns headers (*e.g.*, for data, additional identifiers, or desired metadata) are up to the author (*e.g.,* the ***PubChem_CID*** identifier header in Figure 2). Note that there may be many ***Synonym*** and ***Comment*** columns in the file to provide space for more chemical names and metadata, respectively.

The author-submitted template file[18] should contain **at least one** of the following columns: ***SMILES, InChI, Name*** or ***InChIKey***. The ***Name*** column corresponds to a single primary name for the chemical structure. Each ***Synonym*** column corresponds to an additional chemical name (one name entry per column). Each ***Comment*** column can be added to provide additional text that may be important to the downstream user. Authors can also provide additional CSV/TSV columns (or SDF SD fields) containing information about their chemical substances (with unique, descriptive headers) for additional context. Chemical database identifiers or registry numbers could be included in this manner (as additional columns or fields), or as a ***Synonym***. Note that chemical records indicating chemical structure with only ***InChIKey*** or ***Name*** will not contain sufficient information to describe a chemical structure; and *can only be mapped to existing entries* in destination resources. Batch services are available (*e.g.* from PubChem[7,19] or CompTox[9,20]) for authors to add, *e.g.,* ***SMILES*** and/or ***InChI*** to their records, based upon the ***Name*** or other identifiers.

Figure 1 in Schymanski & Bolton 2021[11] shows the template file, which is available for download[18] and as Supporting Information with this article. Figure 2 below shows an example submission according to the proposed template, created by sub-setting the “HSDBTPS” dataset of literature-mined and curated transformation products from the Hazardous Substance Data Bank (HSDB) in PubChem[21,22]. This example provides the ***Name***, ***SMILES*** and ***InChIKey*** fields as suggested, and an identifier (the PubChem Compound Identifier, CID) as an additional (optional) column (***PubChem_CID***) with a unique and easily recognizable header that can be processed by other resources as they choose, helping with interoperability.



PubChem_CID	Name	SMILES	InChIKey
2256	Atrazine	CCNC1=NC(=NC(=N1)Cl)NC(C)C	MXWJVTOOROXGIU-UHFFFAOYSA-N
2328	Bentazone	CC(C)N1C(=O)C2=CC=CC=C2NS1(=O)=O	ZOMSMJKLGFBRBS-UHFFFAOYSA-N
3030	Dicamba	COC1=C(C=CC(=C1C(=O)O)Cl)Cl	IWEDIXLBFLAXBO-UHFFFAOYSA-N
3120	Diuron	CN(C)C(=O)NC1=CC(=C(C=C1)Cl)Cl	XMTQQYYKAHVGBJ-UHFFFAOYSA-N
4169	Metolachlor	CCC1=CC=CC(=C1N(C(C)COC)C(=O)CCl)C	WVQBLGZPHOPPFO-UHFFFAOYSA-N
7257	3,4-Dichloroaniline	C1=CC(=C(C=C1N)Cl)Cl	SDYWXFYBZPNOFX-UHFFFAOYSA-N
12584	Ammelide	C1(=NC(=O)NC(=O)N1)N	YSKUZVBSHIWEFK-UHFFFAOYSA-N

Figure 2: An example chemical structure data file constructed according to the proposed template[18] by taking a subset of the HSDBTPS structure data[21]. Image created in RStudio (Version 1.2.5042). The HSDBTPS efforts resulted in the deposition of 5 new structures to PubChem all documented in HSDB text snippets, CIDs 146035700, 146035701, 146035702, 146035703 and 146037633.

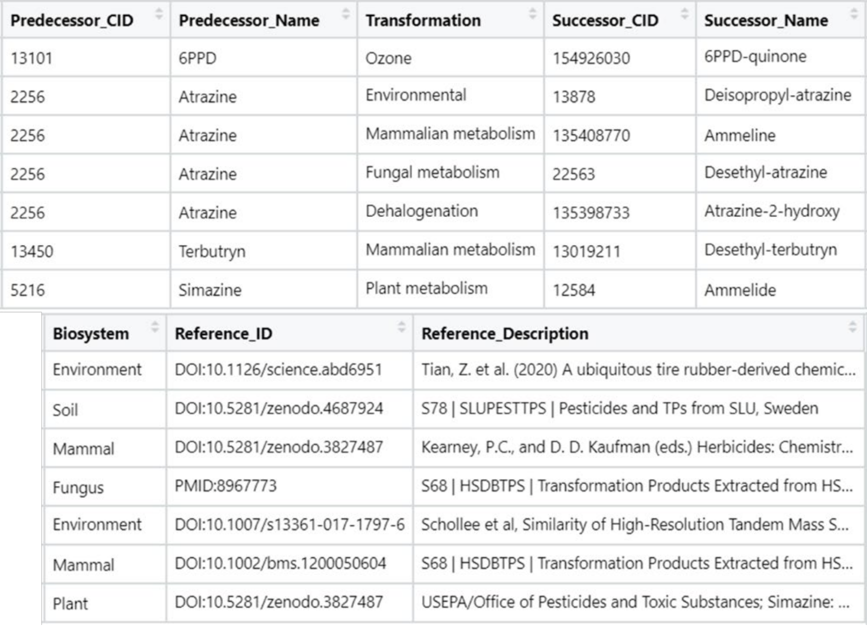
Transformations Data

The advancement of modern science is data driven[23,24]. Providing key data in a ready to use format helps to assist in its reuse in research articles, regulatory reports, or machine learning data models. Exposomics especially needs access to ready-to-use, high-quality chemical information from individual research articles (*e.g.,* such as the connection of detected chemicals with the disease endpoint investigated or the aggregation of known metabolites of thousands of common chemicals). For instance, HSDB contains metabolites and metabolism information for 3220 chemicals gathered over 40 years, but these are only available as text snippets that need to be matched to chemical structures by synonyms followed by manual curation (initial efforts have covered only 1/100th of this dataset[22]). However, as mentioned above, a key challenge in exposomics is to connect chemicals (*e.g.,* of anthropogenic origin, but also endogenous or exogenous chemicals) that are associated with exposures with their biological response. Since metabolism is the most dynamic of the biological responses, and metabolites per definition fall into the same molecular mass category as many anthropogenic chemicals of concern, a key gap in exposomics knowledge is the connection between chemicals and their metabolites. The efforts of many will be needed to help fill this knowledge gap, and the timing could not be better for exposomics with several recent studies emerging using *in vitro* enzymes to investigate parent-metabolite relationships of drugs and other relevant chemicals[25,26].

The Transformations template provided here has been designed on the basis of recent efforts to fill the gaps of transformation products in PubChem using literature data[27], in collaboration with the NORMAN Suspect List Exchange (NORMAN-SLE)[28–30]. Several datasets from a variety of sources have now been processed. Transformations from the NORMAN-SLE, where S## refers to the list number, followed by the list code, include: S60 SWISSPEST19[31,32], S66 EAWAGTPS[33,34], S68 HSDBTPS[21,22], S73 METXBIODB[35,36], S74 REFTPS[37], S78 SLUPESTTPS[38,39], S79 UACCSCEC[40,41] and S81 THSTPS[42] (list available from https://git-r3lab.uni.lu/eci/pubchem/-/raw/master/annotations/tps/Transformation_Datasets.txt). Of these, MetXBioDB also contains enzyme information, while the rest are primarily environmental data. Figure 3 shows an example "environmental" dataset compiled from several of these lists, using the proposed template. In addition to the NORMAN-SLE datasets, a dataset of more than 1200 transformations from

ChEMBL[43] has also been added, including enzyme, gene and protein information (where available). An example of Transformations with more biological information available is given in Figure 4.

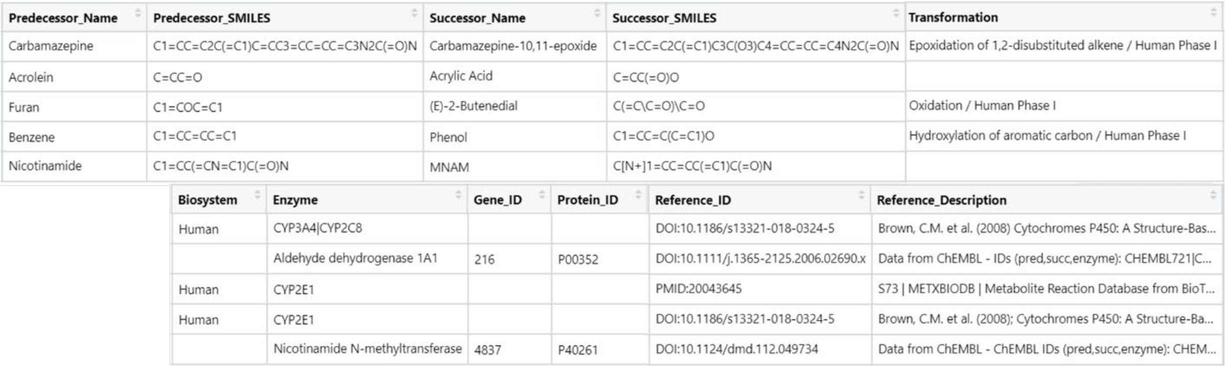
Information about both the predecessor (parent/precursor) and successor (transformation product/metabolite) must be given for a valid transformation. The template can accept *at least one* of ***Name***, ***SMILES*** or ***PubChem CID*** for each, where ***SMILES*** or ***CID*** is preferred, and ***SMILES*** will be the most interoperable. Note that these need not be consistent – for instance, it is possible to provide ***SMILES*** of the successor and a ***CID*** of the predecessor if a ***Name*** or ***CID*** is not available for the successor. It is preferable to give two fields, Figure 3 shows the example of ***Name*** and ***CID***, while Figure 4 an example of ***SMILES*** and ***Name*** (top panel on each figure).



Predecessor_CID	Predecessor_Name	Transformation	Successor_CID	Successor_Name
13101	6PPD	Ozone	154926030	6PPD-quinone
2256	Atrazine	Environmental	13878	Deisopropyl-atrazine
2256	Atrazine	Mammalian metabolism	135408770	Ammeline
2256	Atrazine	Fungal metabolism	22563	Desethyl-atrazine
2256	Atrazine	Dehalogenation	135398733	Atrazine-2-hydroxy
13450	Terbutryn	Mammalian metabolism	13019211	Desethyl-terbutryn
5216	Simazine	Plant metabolism	12584	Ammelide

Biosystem	Reference_ID	Reference_Description
Environment	DOI:10.1126/science.abd6951	Tian, Z. et al. (2020) A ubiquitous tire rubber-derived chemic...
Soil	DOI:10.5281/zenodo.4687924	S78 | SLUPESTTPS | Pesticides and TPs from SLU, Sweden
Mammal	DOI:10.5281/zenodo.3827487	Kearney, P.C., and D. D. Kaufman (eds.) Herbicides: Chemistr...
Fungus	PMID:8967773	S68 | HSDBTPS | Transformation Products Extracted from HS...
Environment	DOI:10.1007/s13361-017-1797-6	Schollee et al, Similarity of High-Resolution Tandem Mass S...
Mammal	DOI:10.1002/bms.1200050604	S68 | HSDBTPS | Transformation Products Extracted from HS...
Plant	DOI:10.5281/zenodo.3827487	USEPA/Office of Pesticides and Toxic Substances; Simazine: ...

Figure 3: An example of various environmental transformations constructed according to the proposed Transformations template[44] (using Name and PubChem CID), taking a subset of transformations from NORMAN-SLE datasets (REFTPS[37], HSDBTPS[21], SLUPESTTPS[38], EAWAGTPS[33] and SWISSPEST19[31]). Image created in RStudio (Version 1.2.5042).



Predecessor_Name	Predecessor_SMILES	Successor_Name	Successor_SMILES	Transformation
Carbamazepine	C1=CC=C2C(=C1)C=CC3=CC=CC=C3N2C(=O)N	Carbamazepine-10,11-epoxide	C1=CC=C2C(=C1)C3C(O3)C4=CC=CC=C4N2C(=O)N	Epoxidation of 1,2-disubstituted alkene / Human Phase I
Acrolein	C=CC=O	Acrylic Acid	C=CC(=O)O	
Furan	C1=COC=C1	(E)-2-Butenedial	C(=C\C=O)\C=O	Oxidation / Human Phase I
Benzene	C1=CC=CC=C1	Phenol	C1=CC=C(C=C1)O	Hydroxylation of aromatic carbon / Human Phase I
Nicotinamide	C1=CC(=CN=C1)C(=O)N	MNAM	C[N+]1=CC=CC(=C1)C(=O)N	

Biosystem	Enzyme	Gene_ID	Protein_ID	Reference_ID	Reference_Description
Human	CYP3A4|CYP2C8			DOI:10.1186/s13321-018-0324-5	Brown, C.M. et al. (2008) Cytochromes P450: A Structure-Bas...
	Aldehyde dehydrogenase 1A1	216	P00352	DOI:10.1111/j.1365-2125.2006.02690.x	Data from ChEMBL - IDs (pred,succ,enzyme): CHEMBL721|C...
Human	CYP2E1			PMID:20043645	S73 | METXBIODB | Metabolite Reaction Database from BioT...
Human	CYP2E1			DOI:10.1186/s13321-018-0324-5	Brown, C.M. et al. (2008); Cytochromes P450: A Structure-Ba...
	Nicotinamide N-methyltransferase	4837	P40261	DOI:10.1124/dmd.112.049734	Data from ChEMBL - ChEMBL IDs (pred,succ,enzyme): CHEM...

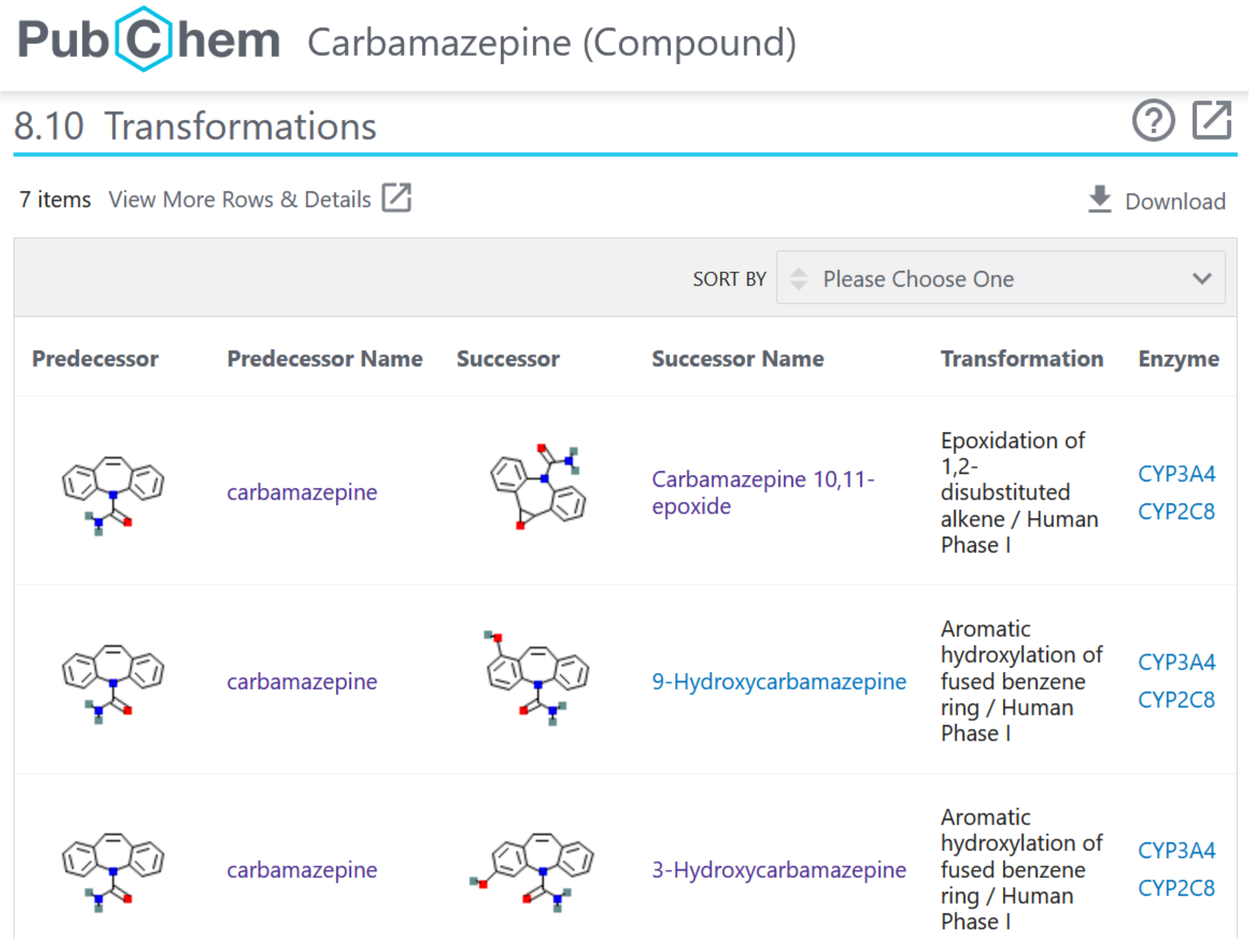
Figure 4: An example of biological transformations constructed according to the proposed Transformations template[44] (using Name and SMILES), taking a subset of transformations from NORMAN-SLE dataset MetXBioDB[35] (from BioTransformer[36]) and the ChEMBL[43] datasets on PubChem; both datasets have some degree of enzyme, gene and/or protein information available.

If available, a brief description of the transformation is useful and can be provided in the "***Transformation***" field (top panel, Figure 3 and Figure 4). Short, informative descriptions are preferred; the current entries have been either extracted automatically from existing datasets or entered manually. In the future, it may be possible to provide some guidance via an ontology as the public dataset grows to improve the machine readability. Similarly, if information on the biosystem is available (*i.e.,* where the transformation takes place), this can be included in the ***Biosystem*** column (see Figure 3 and Figure 4 for examples).

For datasets with biological information, this can be provided (optionally) in the ***Enzyme***, ***Gene_ID*** and ***Protein_ID*** columns. At this stage the template allows flexible input (see Figure 4 for examples) but recommend ***Enzyme*** are provided as either: Enzyme Commission (EC) number,[45–47] such as "EC 2.3.2.23"; gene symbol, such as "CYP1A1"; or as enzyme names, such as "Aryl hydrocarbon hydroxylase". The ***Gene_ID*** is expected to be an NCBI Gene[48] ID, such as "1543". The ***Protein_ID*** is expected to be either an NCBI Protein[49] accession, such as "NP_059488.2" or an UniProt identifier,[50] such as "P08684". If multiple entries for ***Enzyme***, ***Gene_ID*** and ***Protein_ID*** are provided, they should be separated by a "pipe" symbol ("|") or provided as new rows.

Finally, the ***Reference_ID*** and ***Reference_Description*** columns provide the opportunity to credit the original sources of the information. ***Reference_ID*** entries should be either PubMed identifiers[51] (PMIDs) or Digital Object Identifiers[52] (DOIs), preceded with "PMID:" or "DOI:", respectively, for easy recognition, and separated by a "pipe" ("|") if multiple IDs exist (they can be mixed – for example, "PMID:33929905|DOI:10.1186/s13321-018-0324-5"). The ***Reference_Description*** can be used to provide a free text form of the reference, to describe the data source (if no PMID / DOI available) or to describe evidence of the transformation. Only ***Reference_ID*** can be processed automatically. Again, see Figure 3 and Figure 4 and the Transformations template[44] for examples.

So far, about 6000 Transformations have been processed using these templates, from nine different sources (many of these being composite data from several sources themselves, including ChEMBL[43], MetXBioDB[35] and REFTPS[37]). The Transformations are being integrated into current computational mass spectrometry workflows (such as patRoon[53] and as documented in Krier *et al.*[22]) and are openly available for all. The summarized files are likewise available for comprehensive efforts such as BioTransformer[36] to add this new data to their training set (MetXBioDB[35] is the library behind BioTransformer) and likewise improve predictions. Overall, FAIR transformations data will greatly support exposomics, and discussions to extend these templates into fields with formal ontologies and/or other formats such as mzTab[54,55] in the future are welcomed. As demonstrated in Figure 5 and Figure 6, one can see the benefits of arranging data in FAIR templates. Figure 5 is an example of a resulting Transformation entry in PubChem, while Figure 6 can be created automatically in CDK Depict using simple code in R to create annotated reaction SMILES from the fields shown in Figure 3 only.



PubChem Carbamazepine (Compound)

8.10 Transformations

7 items View More Rows & Details

Download

SORT BY Please Choose One

Predecessor	Predecessor Name	Successor	Successor Name	Transformation	Enzyme
	carbamazepine		Carbamazepine 10,11-epoxide	Epoxidation of 1,2-disubstituted alkene / Human Phase I	CYP3A4 CYP2C8
	carbamazepine		9-Hydroxycarbamazepine	Aromatic hydroxylation of fused benzene ring / Human Phase I	CYP3A4 CYP2C8
	carbamazepine		3-Hydroxycarbamazepine	Aromatic hydroxylation of fused benzene ring / Human Phase I	CYP3A4 CYP2C8

Figure 5: Example "Transformations" table in PubChem for Carbamazepine, demonstrating possible display options (including hyperlinking) for FAIR Transformations. Source: https://pubchem.ncbi.nlm.nih.gov/compound/2554#section=Transformations.

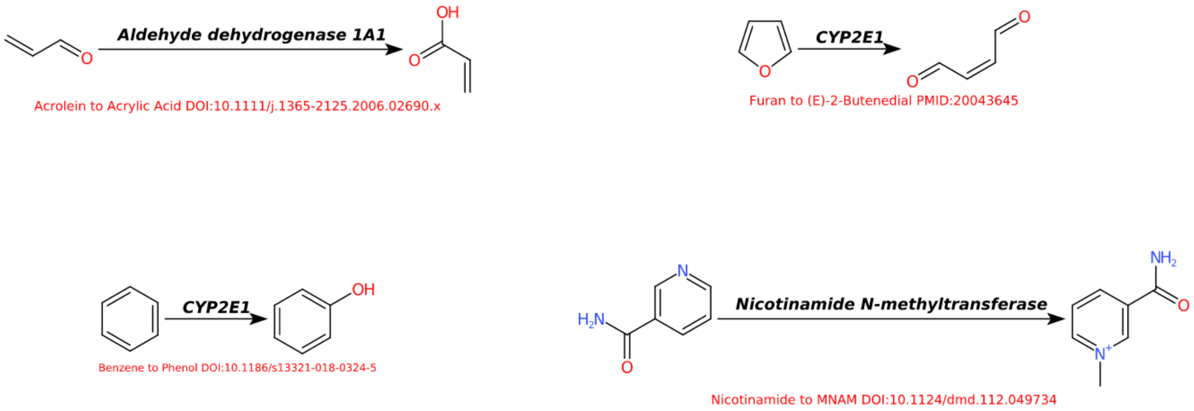

Aldehyde dehydrogenase 1A1
O
OH
O
Acrolein to Acrylic Acid DOI:10.1111/j.1365-2125.2006.02690.x
CYP2E1
O
O
O
Furan to (E)-2-Butenedial PMID:20043645
CYP2E1
OH
Benzene to Phenol DOI:10.1186/s13321-018-0324-5
N
H2N
O
Nicotinamide N-methyltransferase
NH2
O
N+
Nicotinamide to MNAM DOI:10.1124/dmd.112.049734

Figure 6: Example reactions corresponding with the last four rows of Figure 3, automatically created and depicted with CDK Depict[56] (https://www.simolecule.com/cdkdepict/depict.html) directly from template content shown in Figure 3 (SMILES, Name, Enzyme and Reference_ID fields).

Closing

Exposomics is a data-driven science, and vast quantities of information will be needed for it to be successful. By making the output of exposomics research available in a more machine-readable way, we can accelerate our progress and rise to the challenge. The templates provided here are a means to make primary outputs FAIR (***Findable, Accessible, Interoperable***, ***Reusable***). When authors provide this content as Supplementary Information, it can be readily accessed and utilized, ideally without human intervention. When the journal interlinks these Supplementary Material files with the article DOI and associated metadata, other resources can rapidly find and integrate this content and provide enhanced services for the entire community. Improving the ***FAIR***ness of Supplementary Material greatly decreases the effort to combine and aggregate information between papers and improves the correctness of the information over text-mining based approaches. It also greatly enhances the visibility of the individual works and research outputs. As a young scientific discipline, the exposome should learn from its closely related 'elder' disciplines. Genomic approaches gained incredible traction due to the widely encouraged and eventually mandated sharing of information. Let us take these lessons to heart and advance together as a field. We need to share information – and lots of it – to help make sense of the exposome. The use of these facile, ready-to-use templates will help advance exposomics by contributing vital information to complete the exposomics "puzzle".

Acknowledgements and Funding

We gratefully acknowledge discussions with the entire PubChem team, especially Jian (Jeff) Zhang and Tiejun Cheng for their joint work on the transformations, as well as Ben Shoemaker, Paul Thiessen, Siqian He, and Asta Gindulyte. We also gratefully acknowledge discussions with Egon Willighagen and the editorial team at the Journal of Cheminformatics (surrounding the lead-up article to this article), and many collaborators who have worked on depositions within PubChem and the NORMAN-SLE. Special mentions go to Frank Menger (SLU, Sweden) and Lidia Belova (University of Antwerp, Belgium), for testing and depositing data using earlier versions of the transformations template (SLUPESTTPS and UACCSCEC, respectively). We are also grateful to Anca Baesu (McGill University, Canada) and Parviel Chirsir (University of Luxembourg), as well as Noelia Ramirez and colleagues (URV, Tarragona, Spain) for their testing and contributions using the existing templates (REFTPS and THSTPS, respectively).

EEB is funded by the Intramural Research Program of the National Library of Medicine, National Institutes of Health; ELS acknowledges funding support from the Luxembourg National Research Fund (FNR) for project A18/BM/12341006.

Supplementary Materials

The chemical structure data submission template and transformations template are provided as Supplementary Material and are also available online[18,44,57].

All Transformations mentioned in this article are openly available on the NORMAN-SLE and PubChem.

References

1. Wild CP. Complementing the genome with an “exposome”: the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol Biomarkers Prev*. 2005;14(8):1847-1850. doi:10.1158/1055-9965.EPI-05-0456

2. Miller GW. Exposome: a new field, a new journal. *Exposome*. 2021;1(1). doi:10.1093/exposome/osab001

3. GlobeNewswire, Inc. Genomics Market to Reach USD 94.66 Billion by 2028; Increasing Genomics’ Application & Rising Government Investments to Amplify Market Growth: Says Fortune Business Insights^{TM}. Accessed September 5, 2021. https://www.globenewswire.com/news-release/2021/05/20/2233128/0/en/Genomics-Market-to-Reach-USD-94-66-Billion-by-2028-Increasing-Genomics-Application-Rising-Government-Investments-to-Amplify-Market-Growth-Says-Fortune-Business-Insights.html

4. P&S Intelligence. Global Genomics Market to Reach $68 Billion by 2030: P&S Intelligence. Accessed September 5, 2021. https://www.prnewswire.com/news-releases/global-genomics-market-to-reach-68-billion-by-2030-ps-intelligence-301125318.html

5. Vermeulen R, Schymanski EL, Barabási AL, Miller GW. The exposome and health: Where chemistry meets biology. *Science*. 2020;367(6476):392. doi:10.1126/science.aay3164

6. Miller GW, Jones DP. The nature of nurture: refining the definition of the exposome. *Toxicol Sci*. 2014;137(1):1-2. doi:10.1093/toxsci/kft251

7. Kim S, Chen J, Cheng T, et al. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Research*. 2019;47(D1):D1102-D1109. doi:10.1093/nar/gky1033

8. Davis AP, Grondin CJ, Johnson RJ, et al. Comparative Toxicogenomics Database (CTD): update 2021. *Nucleic Acids Research*. 2021;49(D1):D1138-D1143. doi:10.1093/nar/gkaa891

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

10. Schymanski EL, Baker NC, Williams AJ, et al. Connecting environmental exposure and neurodegeneration using cheminformatics and high resolution mass spectrometry: potential and challenges. *Environ Sci: Processes Impacts*. 2019;21(9):1426-1445. doi:10.1039/C9EM00068B

11. Schymanski EL, Bolton EE. FAIR chemical structures in the Journal of Cheminformatics. *J Cheminform*. 2021;13(1):50. doi:10.1186/s13321-021-00520-4

12. Baker NC, Schymanski EL, Williams AJ. Literature Neurotoxicants: Excel Macro File. *FigShare*. doi:10.23645/epacomptox.7334603

13. Peter Suber. Open Access Overview (definition, introduction). Accessed July 3, 2021. http://legacy.earlham.edu/~peters/fos/overview.htm

14. GO FAIR. FAIR Principles. Published 2021. Accessed March 23, 2021. https://www.go-fair.org/fair-principles/

15. Wilkinson MD, Dumontier M, Aalbersberg IjJ, et al. Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*. 2016;3(1):1-9. doi:10.1038/sdata.2016.18

16. Heller S, McNaught A, Stein S, Tchekhovskoi D, Pletnev I. InChI - the worldwide chemical structure identifier standard. *Journal of Cheminformatics*. 2013;5(1):7. doi:10.1186/1758-2946-5-7

17. NCBI/NLM/NIH. PubChem Documentation: Substance SDF Submission. Published 2021. Accessed March 23, 2021. https://pubchem.ncbi.nlm.nih.gov/upload/docs/examples/substance_submission.sdf

18. NCBI/NLM/NIH. Chemical Structure Data Template (CSV). Published 2021. Accessed May 9, 2021. https://ftp.ncbi.nlm.nih.gov/pubchem/Other/Submissions/Chemical_Structure_Data_Template.csv

19. NCBI/NLM/NIH. PubChem Identifier Exchange. Published 2021. Accessed March 23, 2021. https://pubchem.ncbi.nlm.nih.gov/idexchange/idexchange.cgi

20. United States Environmental Protection Agency. CompTox Batch Search. Published 2021. Accessed March 23, 2021. https://comptox.epa.gov/dashboard/dsstoxdb/batch_search

21. LCSB-ECI, Krier, Jessy, Schymanski, Emma, et al. S68 | HSDBTPS | Transformation Products Extracted from HSDB Content in PubChem. Published online June 11, 2020. doi:10.5281/ZENODO.3827487

22. Krier J, Singh RR, Kondić T, et al. Discovering pesticides and their TPs in Luxembourg waters using open cheminformatics approaches. *Environment International*. 2022;158:106885. doi:10.1016/j.envint.2021.106885

23. Montáns FJ, Chinesta F, Gómez-Bombarelli R, Kutz JN. Data-driven modeling and learning in science and engineering. *Comptes Rendus Mécanique*. 2019;347(11):845-855. doi:10.1016/j.crme.2019.11.009

24. Clauset A, Larremore DB, Sinatra R. Data-driven predictions in the science of science. *Science*. 2017;355(6324):477-480. doi:10.1126/science.aal4217

25. Liu K, Lee C, Singer G, et al. Enzyme-Based Chemical Identification for Metabolomics. *FASEB j*. 2021;35(S1):fasebj.2021.35.S1.04277. doi:10.1096/fasebj.2021.35.S1.04277

26. Ross DH, Seguin RP, Krinsky AM, Xu L. *High-Throughput Measurement and Machine Learning-Based Prediction of Collision Cross Sections for Drugs and Drug Metabolites*. Bioinformatics; 2021. doi:10.1101/2021.05.13.443945

27. Schymanski EL, Kondić T, Neumann S, Thiessen PA, Zhang J, Bolton EE. Empowering large chemical knowledge bases for exposomics: PubChemLite meets MetFrag. *J Cheminform*. 2021;13(1):19. doi:10.1186/s13321-021-00489-0

28. NORMAN Network. NORMAN Suspect List Exchange. NORMAN Suspect List Exchange. Accessed June 9, 2019. https://www.norman-network.com/nds/SLE/

29. NORMAN Network. NORMAN Suspect List Exchange on Zenodo. NORMAN Suspect List Exchange: Zenodo Community. Accessed June 9, 2019. https://zenodo.org/communities/norman-sle/

30. NORMAN Network, NCBI/NLM/NIH. NORMAN SLE Classification Browser. Accessed May 7, 2020. https://pubchem.ncbi.nlm.nih.gov/classification/#hid=101

31. Kiefer, Karin, Müller, Adrian, Singer, Heinz, Hollender, Juliane. S60 | SWISSPEST19 | Swiss Pesticides and Metabolites from Kiefer et al 2019. Published online November 17, 2019. http://doi.org/10.5281/zenodo.3544760

32. Kiefer K, Müller A, Singer H, Hollender J. New relevant pesticide transformation products in groundwater detected using target and suspect screening for agricultural and urban micropollutants with LC-HRMS. *Water Research*. 2019;165:114972. doi:10.1016/j.watres.2019.114972

33. Schollee, Jennifer, Schymanski, Emma. S66 | EAWAGTPS | Parent-Transformation Product Pairs from Eawag. Published online April 23, 2020. doi:10.5281/ZENODO.3754448

34. Schollée JE, Schymanski EL, Stravs MA, Gulde R, Thomaidis NS, Hollender J. Similarity of High-Resolution Tandem Mass Spectrometry Spectra of Structurally Related Micropollutants and Transformation Products. *J Am Soc Mass Spectrom*. 2017;28(12):2692-2704. doi:10.1007/s13361-017-1797-6

35. Djoumbou-Feunang, Yannick, Schymanski, Emma, Zhang, Jeff, Wishart, David S. S73 | METXBIODB | Metabolite Reaction Database from BioTransformer. Published online November 5, 2020. doi:10.5281/ZENODO.4056560

36. Djoumbou-Feunang Y, Fiamoncini J, Gil-de-la-Fuente A, Greiner R, Manach C, Wishart DS. BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *J Cheminform*. 2019;11(1):2. doi:10.1186/s13321-018-0324-5

37. Schymanski, Emma. S74 | REFTPS | Transformation Products and Reactions from Literature. Published online December 12, 2020. doi:10.5281/ZENODO.4318838

38. Menger, Frank, Boström, Gustaf. S78 | SLUPESTTPS | Pesticides and TPs from SLU, Sweden. Published online May 10, 2021. doi:10.5281/ZENODO.4687924

39. Menger F, Boström G, Jonsson O, et al. Identification of Pesticide Transformation Products in Surface Water Using Suspect Screening Combined with National Monitoring Data. *Environ Sci Technol*. 2021;55(15):10343-10353. doi:10.1021/acs.est.1c00466

40. Belova L, Caballero-Casero N, van Nuijs ALN, Covaci A. Ion Mobility-High-Resolution Mass Spectrometry (IM-HRMS) for the Analysis of Contaminants of Emerging Concern (CECs): Database Compilation and Application to Urine Samples. *Anal Chem*. 2021;93(16):6428-6436. doi:10.1021/acs.analchem.1c00142

41. Belova, Lidia, Caballero-Casero, Noelia, van Nuijs, Alexander L. N., Covaci, Adrian. S79 | UACCSCEC | Collision Cross Section (CCS) Library from UAntwerp. Published online May 10, 2021. doi:10.5281/ZENODO.4704648

42. Merino C, Vinaixa M, Ramirez N. S81 | THSTPS | Thirdhand Smoke Specific Metabolites. Published online September 2, 2021. doi:10.5281/ZENODO.5394629

43. Gaulton A, Hersey A, Nowotka M, et al. The ChEMBL database in 2017. *Nucleic Acids Res*. 2017;45(D1):D945-D954. doi:10.1093/nar/gkw1074

44. NCBI/NLM/NIH. Transformations Data Template (CSV). Published 2021. Accessed May 25, 2021. https://ftp.ncbi.nlm.nih.gov/pubchem/Other/Submissions/Transformations_Template.csv

45. McDonald AG, Boyce S, Tipton KF. ExplorEnz: the primary source of the IUBMB enzyme list. *Nucleic Acids Research*. 2009;37(Database):D593-D597. doi:10.1093/nar/gkn582

46. Bairoch A. The ENZYME database in 2000. *Nucleic Acids Res*. 2000;28(1):304-305. doi:10.1093/nar/28.1.304

47. Chang A, Jeske L, Ulbrich S, et al. BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic Acids Research*. 2021;49(D1):D498-D508. doi:10.1093/nar/gkaa1025

48. Brown GR, Hem V, Katz KS, et al. Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res*. 2015;43(Database issue):D36-42. doi:10.1093/nar/gku1055

49. Benson DA, Cavanaugh M, Clark K, et al. GenBank. *Nucleic Acids Res*. 2018;46(D1):D41-D47. doi:10.1093/nar/gkx1094

50. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res*. 2017;45(D1):D158-D169. doi:10.1093/nar/gkw1099

51. Sayers EW, Beck J, Bolton EE, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2021;49(D1):D10-D17. doi:10.1093/nar/gkaa892

52. International DOI Foundation. Frequently Asked Questions about the DOI® System. Accessed September 7, 2021. https://www.doi.org/faq.html

53. Helmus R, ter Laak TL, van Wezel AP, de Voogt P, Schymanski EL. patRoon: open source software platform for environmental mass spectrometry based non-target screening. *J Cheminform*. 2021;13(1):1. doi:10.1186/s13321-020-00477-w

54. Griss J, Jones AR, Sachsenberg T, et al. The mzTab Data Exchange Format: Communicating Mass-spectrometry-based Proteomics and Metabolomics Experimental Results to a Wider Audience. *Molecular & Cellular Proteomics*. 2014;13(10):2765-2775. doi:10.1074/mcp.O113.036681

55. Hoffmann N, Rein J, Sachsenberg T, et al. mzTab-M: A Data Standard for Sharing Quantitative Results in Mass Spectrometry Metabolomics. *Anal Chem*. 2019;91(5):3302-3310. doi:10.1021/acs.analchem.8b04310

56. Mayfield J. CDK Depict Web Interface. Accessed October 30, 2018. http://simolecule.com/cdkdepict/depict.html

57. NCBI/NLM/NIH. PubChem Submissions Template Folder. Published 2021. Accessed May 25, 2021. https://ftp.ncbi.nlm.nih.gov/pubchem/Other/Submissions/