

# Wikipedia on the CompTox Chemicals Dashboard: Connecting Resources to Enrich Public Chemical Data

Gabriel Sinclair, Inthirany Thillainadarajah, Brian Meyer, Vicente Samano, Sakuntala Sivasupramaniam, Linda Adams, Egon L. Willighagen, Ann M. Richard, Martin Walker, and Antony J. Williams\*



Cite This: *J. Chem. Inf. Model.* 2022, 62, 4888–4905



Read Online

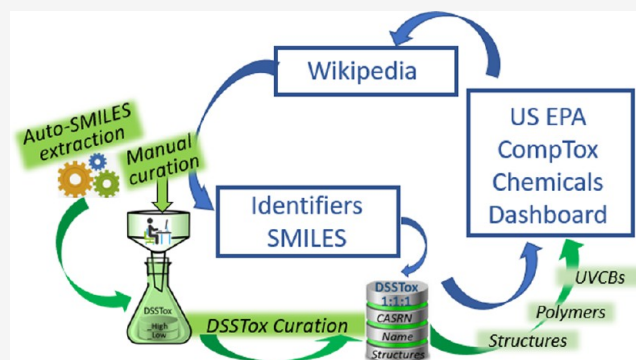
ACCESS |

Metrics & More

Article Recommendations

Supporting Information

**ABSTRACT:** The online encyclopedia Wikipedia aggregates a large amount of data on chemistry, encompassing well over 20,000 individual Wikipedia pages and serves the general public as well as the chemistry community. Many other chemical databases and services utilize these data, and previous projects have focused on methods to index, search, and extract it for review and use. We present a comprehensive effort that combines bulk automated data extraction over tens of thousands of pages, semiautomated data extraction over hundreds of pages, and fine-grained manual extraction of individual lists and compounds of interest. We then correlate these data with the existing contents of the U.S. Environmental Protection Agency's (EPA) Distributed Structure-Searchable Toxicity (DSSTox) database. This was performed with a number of intentions including ensuring as complete a mapping as possible between the Dashboard and Wikipedia so that relevant snippets of the article are loaded for the user to review. Conflicts between Dashboard content and Wikipedia in terms of, for example, identifiers such as chemical registry numbers, names, and InChIs and structure-based collisions such as SMILES were identified and used as the basis of curation of both DSSTox and Wikipedia. This work also allowed us to evaluate available data for sets of chemicals of interest to the Agency, such as synthetic cannabinoids, and expand the content in DSSTox as appropriate. This work also led to improved bidirectional linkage of the detailed chemistry and usage information from Wikipedia with expert-curated structure and identifier data from DSSTox for a new list of nearly 20,000 chemicals. All of this work ultimately enhances the data mappings that allow for the display of the introduction of the Wikipedia article in the community-accessible web-based EPA CompTox Chemicals Dashboard, enhancing the user experience for the thousands of users per day accessing the resource.



## INTRODUCTION

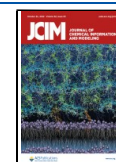
Wikipedia has taken on a role as the primary free access, online encyclopedia for the world. There are many articles in Wikipedia of professional and general interest to the chemistry community. Moreover, for groups in the chemistry community concerned with aggregating chemical data, the work performed by Wikipedia contributors and curators on articles about individual chemicals is invaluable. These individual chemical articles may be very rich in detail, particularly for common household chemicals such as aspirin, or may be very basic “stubs” (<https://en.wikipedia.org/wiki/Wikipedia:Stub>) such as Alizarine Yellow R.<sup>1</sup> Many public databases have established an integration with Wikipedia and display either the ledes (i.e., the first few lines of the article) or snippets of the descriptive text from the articles. This includes ChEBI, with a significant portion of the article embedded in the page;<sup>2</sup> ChemSpider, with a limited selection from the article text;<sup>3</sup> PubChem, with a simple link to the page;<sup>4</sup> and many other online databases. The wide usage of Wikipedia chemistry content, and its adoption

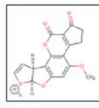
into multiple resources, is indicative of the value that the chemistry user-community recognizes in the work.

The US Environmental Protection Agency's (EPA) CompTox Chemicals Dashboard (hereafter referred to as the Dashboard, available at <https://comptox.epa.gov/dashboard>) is a web-based application that surfaces data and predicted properties of over 900,000 chemicals of interest to EPA's research programs.<sup>5,6</sup> Like other public databases, EPA considers Wikipedia a valuable source of information. Where available, the lede of a Wikipedia chemical article, with a link to the full page, is included to introduce the chemical substance and its uses and relevance to the community (see Figure 1).

Received: July 14, 2022

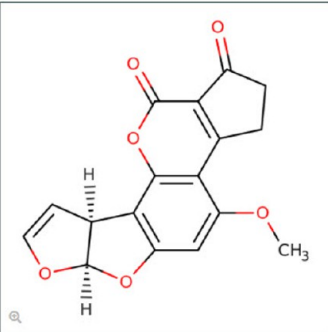
Published: October 10, 2022





**Aflatoxin B1**  
1162-65-8 | DTXSID9020035  
Searched by Approved Name.

**Chemical Details**



**Wikipedia**

Aflatoxin B<sub>1</sub> is an aflatoxin produced by *Aspergillus flavus* and *A. parasiticus*. It is a very potent carcinogen with a TD<sub>50</sub> 3.2 µg/kg/day in rats. This carcinogenic potency varies across species with some, such as rats and monkeys, seemingly much more susceptible than others. Aflatoxin B<sub>1</sub> is a common contaminant in a variety of foods including peanuts, cottonseed meal, corn, and other grains; as well as animal feeds. Aflatoxin B<sub>1</sub>

[Read more](#)

---

**Quality Control Notes**

Rotation (°):

---

**Intrinsic Properties**

Molecular Formula: C<sub>17</sub>H<sub>12</sub>O<sub>6</sub> ▲ MOL FILE 🔍 FIND ALL CHEMICALS

Average Mass: 312.277 g/mol ▲ ISOTOPE MASS DISTRIBUTION

Monoisotopic Mass: 312.063388 g/mol

**Figure 1.** Section of the CompTox Chemicals Dashboard, Chemical Details page for Aflatoxin B1 showing the lede loaded from Wikipedia.

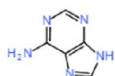
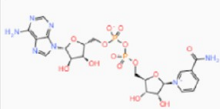
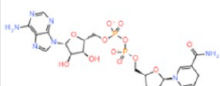
**Wikipedia Chemical Structure Explorer**

Draw a structure

Search mode: Substructure

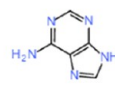
Filter by article name: adenine

Click on a row to see the Wikipedia article

Structure	Article name
	Adenine
	Nicotinamide adenine dinucleotide
	Nicotinamide adenine dinucleotide

**Information**

**Adenine**



MF C<sub>5</sub>H<sub>5</sub>N<sub>5</sub>  
MW 135.13

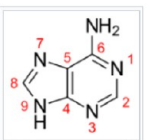
[View in wikipedia](#) [Search similar molecules](#)

*Not to be confused with adenosine.*

**Adenine** (/ˈædɪnin/) (symbol **A** or **Ade**) is a nucleobase (a purine derivative). It is one of the four nucleobases in the nucleic acid of DNA that are represented by the letters G–C–A–T. The three others are guanine, cytosine and thymine. Its derivatives have a variety of roles in biochemistry including cellular respiration, in the form of both the energy-rich adenosine triphosphate (ATP) and the cofactors nicotinamide adenine dinucleotide (NAD), flavin adenine dinucleotide (FAD) and Coenzyme A. It also has functions in protein synthesis and as a chemical component of DNA and RNA.<sup>[2]</sup> The shape of adenine is complementary to either thymine in DNA or uracil in RNA.

The adjacent image shows pure adenine, as an independent molecule. When connected into DNA, a covalent bond is formed between deoxyribose sugar and the bottom left nitrogen (thereby removing the existing hydrogen atom). The remaining structure is called an *adenine residue*, as part of a larger molecule. Adenosine is adenine reacted with ribose, as used in RNA and ATP; deoxyadenosine is adenine attached to deoxyribose, as used to form DNA.

**Structure**

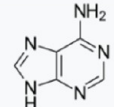
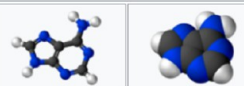


Adenine forms several tautomers, compounds that can be rapidly interconverted and are often considered equivalent. However, in isolated conditions, i.e. in an inert gas matrix and in the gas phase, mainly the 9H-adenine tautomer is found.<sup>[3][4]</sup>

**Biosynthesis**

Purine metabolism involves the formation of adenine and guanine. Both adenine and guanine are derived from the nucleotide inosine monophosphate (IMP), which in turn is synthesized from a pre-existing ribose phosphate through a complex pathway using atoms from the amino acids glycine, glutamine, and aspartic acid, as well as the coenzyme tetrahydrofolate.

**Adenine**

**Names**

Preferred IUPAC name  
9H-Purin-6-amine

Other names  
6-Aminopurine

**Identifiers**

CAS Number	73-24-5 ✓
3D model (JSmol)	<a href="#">Interactive image</a> <a href="#">Interactive image</a>
ChEBI	<a href="#">ChEBI:16708</a> ✓
ChEMBL	<a href="#">ChEMBL226345</a> ✓
ChemSpider	<a href="#">185</a> ✓
DrugBank	<a href="#">DB00173</a> ✓
ECHA TrefCard	<a href="#">100.000.774</a> ✓

**Figure 2.** Wikipedia Chemical Explorer loads chemicals harvested from Wikipedia pages and makes them available for structure, substructure, and similarity searching and loads an embedded page into the application window.

There have been previous efforts to aggregate and curate chemical data on Wikipedia. In December 2007, one of the authors (AJW) initiated a project to curate chemical structures on Wikipedia<sup>7</sup> to link a Wikipedia dataset in ChemSpider<sup>8</sup> to Wikipedia itself. The ultimate goal was to encourage inclusion of a link to the ChemSpider website within chemical articles on Wikipedia, positioned as a “WiChempedia”.<sup>9</sup> The scope

expanded after consultation with Wikipedia user “Walkerma” (MW) when a team of other chemists on the platform engaged to collaboratively curate the data. Part of the activity included a CAS Registry Number (CAS RN) validation aspect of the project<sup>10</sup> that resulted in a collaboration with the Chemical Abstracts Service.<sup>11</sup> This led to the launch of the CAS Common Chemistry data collection and website in 2009,

### This page references all articles with SMILES problems

You can click on any line to open the corresponding article on Wikipedia

#### SMILES with errors (95)

Those SMILES could not be processed by the parser

#### Pages without correct SMILES (68)

List of pages in which all the SMILES have errors. Therefore these pages can not be found in the search page

#### Duplicates (576)

List of pages that contain several SMILES that represent the exact same structure

#### Not found (2103)

List of pages that contain a Chembox or Drugbox but no SMILES field

Last data extraction: 01/10/2022

**Figure 3.** List of different types of SMILES problems indexed on the Wikipedia Chemical Structure Explorer. The largest number of articles with SMILES problems is the class of chemical or drug article pages with other identifying information but without SMILES.

initially providing information for ~8000 chemicals.<sup>12</sup> The site was recently revamped and the dataset expanded to provide information for nearly 500,000 substances from the CAS Registry.<sup>13–15</sup>

Further attempts to harvest chemical data from Wikipedia on an ongoing basis have previously been reported. DBPedia has been extracting structured content from Wikipedia infoboxes for many years,<sup>16</sup> including the Chembox for multiple Wikipedia languages.<sup>17</sup> The Wikipedia Chemical Structure Explorer<sup>18</sup> (available at <http://www.cheminfo.org/wikipedia/>) offers structure and similarity searching of chemicals harvested from Wikipedia (see Figure 2). At the time of writing (10 January 2022), there were 19,204 SMILES structures indexed in the Wikipedia Chemical Structure Explorer across 18,472 English Wikipedia pages. The site provides the ability to search through all chemicals based on exact structure, substructure, and similarity search by drawing a structure in the integrated molecular editor. Searches are conducted in real time: as the structure is drawn, the relevant article name(s) is (are) loaded based on the search option selected, with the article itself alongside.

It is possible to view and download all SMILES that have been harvested from the Wikipedia pages to support the Structure Explorer (available at <http://www.cheminfo.org/wikipedia/smiles.txt>). The downloadable SMILES file includes the article name and the associated SMILES strings. Examination of the file shows that some articles can have multiple associated SMILES. Errors identified in processing the data are also collected in a single page. The errors page includes four types of errors (see Figure 3): pages with SMILES that cannot be processed by the SMILES parser; pages with “incorrect SMILES”; pages with duplicates (distinct SMILES representing the same structure); and pages with other chemical identifier information but without the associated SMILES.<sup>18</sup>

## METHODS

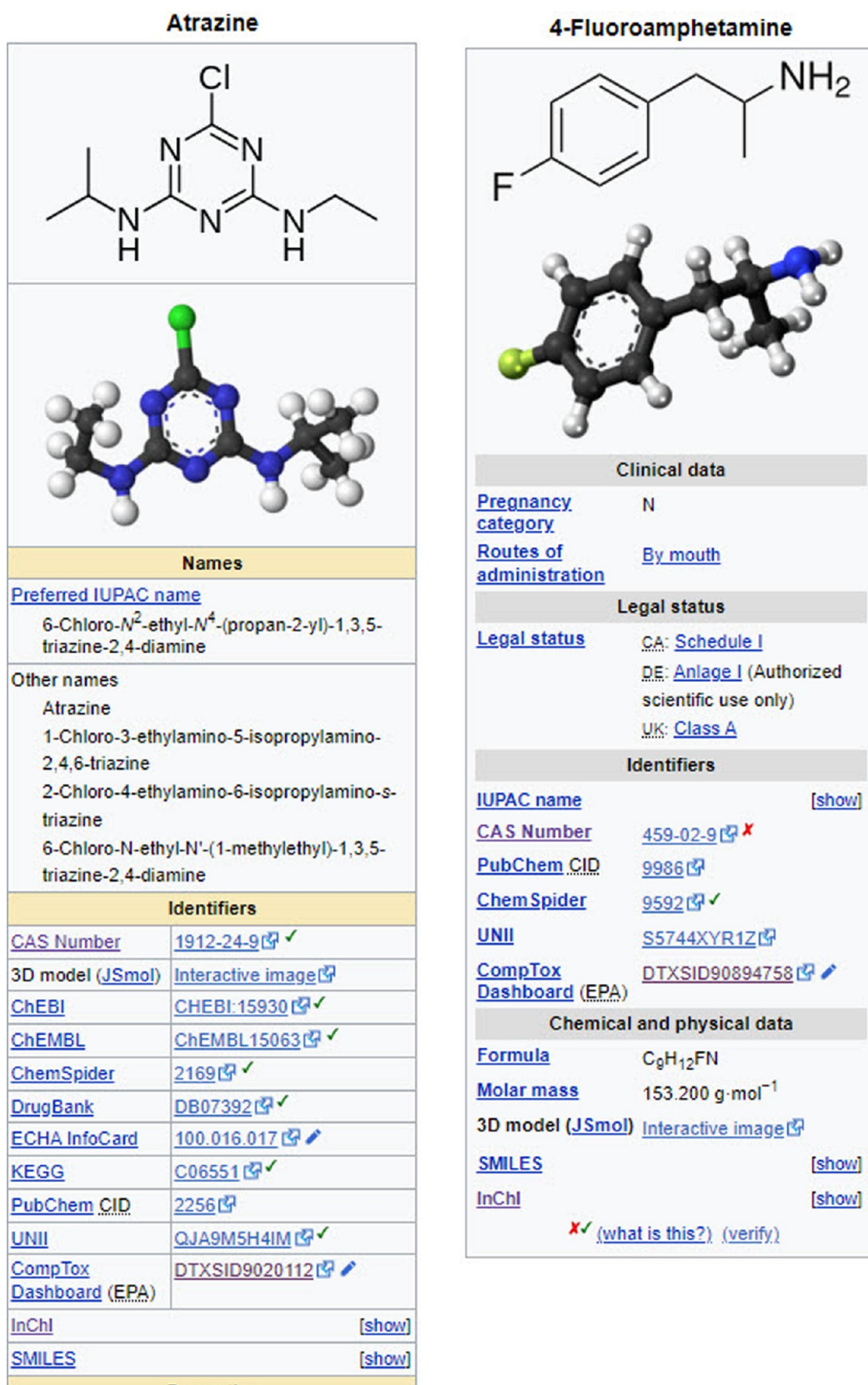
**Wikipedia Chemical Page Data Harvesting.** The harvesting process used by the Wikipedia Chemical Structure Explorer served as a starting point to build a dataset for registration and curation, but different decisions were made in extracting and retaining chemical identifiers; additionally, other chemicals on Wikipedia were added into the dataset by further extraction from indexed pages on the Wikipedia site. Wikipedia editors have assembled a number of category pages listing chemicals,<sup>19</sup> and the structuring of chemistry-related pages

across the online encyclopedia allows navigation of chemical substances by chemical use,<sup>20</sup> chemical substances by CAS RN,<sup>21</sup> and by inclusion on the EPA list of Hazardous Substances.<sup>22</sup> Research and consultation with Wikipedia editors identified a number of chemical list pages, from which additional articles were identified and added to be linked and indexed during this work. Although some of these were discovered by the original search process, others were novel and were deemed to be of interest for linkage.

The automated harvesting of data for this project was based on the process adopted by the Wikipedia Chemical Structure Explorer. Wikipedia pages describing specific chemicals are typically accompanied by a table of summary information (see Figure 4). These tables are formatted using a Wikipedia template code called a “Chembox”<sup>23</sup> or (for pharmaceuticals) “Drugbox”.<sup>24</sup> The MediaWiki API provides an “embedded-in query”,<sup>25</sup> which returns a list of pages containing a specified element. This query was used with the specifications “Template:Chembox” and “Template:Infobox drug” to locate pages describing specific chemicals. The query was restricted to the “main” namespace to exclude unwanted pages, such as drafts or other templates. In total, 12,337 pages embedding the Chembox template and 8332 pages embedding the Drugbox template were identified and harvested for the final data update on 11 January 2022, for a total of 20,669 pages.

To obtain the HTML contents of the located pages, the MediaWiki API was used again with the “parse” action.<sup>26</sup> An initial rough parsing using jsoup<sup>27</sup> was performed to extract the contents of the infoboxes (“Chemboxes” or “Drugboxes”) from each page to reduce storage space. This raw HTML was saved alongside page metadata in JSON format. Across all pages, this covered 12,451 Chemboxes and 8389 Drugboxes. A total of 22 pages were duplicated due to the presence of both a Chembox and a Drugbox on the page. There was little consistency in these pages: in some cases, the information contained within the two infoboxes was duplicative, while in others, the information was contradictory (see Figure 5).

The harvesting of multiple infoboxes on each page is a key difference between this project and the Wikipedia Chemical Structure Explorer. The vast majority of pages harvested in this project contained only single infoboxes (12,237 of the Chembox dataset and 8284 of the Drugbox dataset), but multiple infoboxes were present in a small number of cases. The presence of multiple infoboxes on a page constitutes a small problem in numerical terms, but when it does occur, it



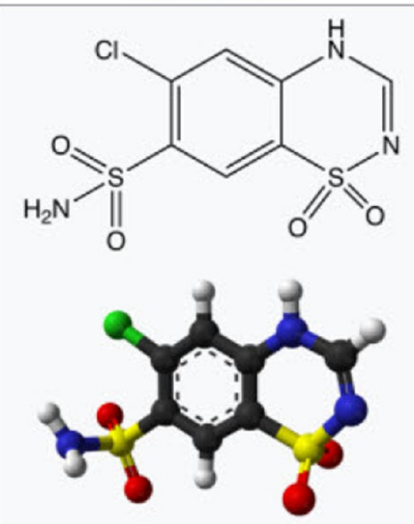
**Figure 4.** Example Chembox (left) for atrazine (DTXSID9020112) and Drugbox (right) for 4-fluoroamphetamine (DTXSID90894758).

can result in data loss or confusion if all data are not carefully harvested (Figure 6).

Among the Chembox dataset, the maximum number of legitimate infoboxes on a page was for the Grubbs catalyst,

with four infoboxes, each representing a distinct chemical entity—in this case, multiple generations of related organoruthenium catalysts. The page for caffeine contains five infoboxes, including both Chemboxes and Drugboxes, but

### Chlorothiazide



Clinical data	
<b>Trade names</b>	Diuril, others
<b>AHFS/Drugs.com</b>	<a href="#">Monograph</a>
<b>MedlinePlus</b>	<a href="#">a682341</a>
<b>Routes of administration</b>	By mouth, IV
<b>ATC code</b>	<a href="#">C03AA04 (WHO)</a>

Legal status	
<b>Legal status</b>	US: <a href="#">R-only</a>

Pharmacokinetic data	
<b>Bioavailability</b>	low
<b>Metabolism</b>	Nil
<b>Elimination half-life</b>	45 to 120 minutes
<b>Excretion</b>	<a href="#">Renal</a>

Identifiers	
<b>IUPAC name</b>	<a href="#">[show]</a>
<b>CAS Number</b>	<a href="#">58-94-6</a> ✓
<b>PubChem CID</b>	<a href="#">2720</a>
<b>IUPHAR/BPS</b>	<a href="#">4835</a>
<b>DrugBank</b>	<a href="#">DB00880</a> ✓
<b>ChemSpider</b>	<a href="#">2619</a> ✓
<b>UNII</b>	<a href="#">77W477J15H</a>
<b>KEGG</b>	<a href="#">D00519</a> ✓
<b>ChEBI</b>	<a href="#">CHEBI:3640</a> ✓

### Chlorothiazide

Names	
Other names	
6-Chloro-2H-1,2,4-benzothiadiazine-7-sulfonamide 1,1-dioxide	

Identifiers	
<b>CAS Number</b>	<a href="#">58-94-6</a> ✓
<b>3D model (JSmol)</b>	<a href="#">Interactive image</a>
<b>ChEBI</b>	<a href="#">CHEBI:3640</a> ✓
<b>ChEMBL</b>	<a href="#">ChEMBL842</a> ✓
<b>ChemSpider</b>	<a href="#">2619</a> ✓
<b>DrugBank</b>	<a href="#">DB00880</a> ✓
<b>ECHA InfoCard</b>	<a href="#">100.000.368</a> ✓
<b>KEGG</b>	<a href="#">D00519</a> ✓
<b>PubChem CID</b>	<a href="#">2720</a>
<b>UNII</b>	<a href="#">77W477J15H</a> ✓
<b>CompTox Dashboard (EPA)</b>	<a href="#">DTXSID0022800</a> ✓
<b>InChI</b>	<a href="#">[show]</a>
<b>SMILES</b>	<a href="#">[show]</a>

Properties	
<b>Melting point</b>	342.5–343 °C (648.5–649.4 °F; 615.6–616.1 K)

Except where otherwise noted, data are given for materials in their [standard state](#) (at 25 °C [77 °F], 100 kPa).

[Infobox references](#)

**Figure 5.** Wikipedia page for chlorothiazide includes two infoboxes, both listing duplicative identifiers including CAS RNs, IUPAC names, and multiple other identifiers (left).

they are incorrectly segregated by format and do not represent distinct entities. This emphasizes one of several problems with the placement of multiple infoboxes on a Wikipedia page: approaching this page naively through an automated mapping system could lead to the false construction of new, unstructured, unidentified chemical entities associated with caffeine in a future dataset.

Among the Drugbox dataset, the legitimate maximum was three infoboxes, for a set of six pages including Sarafotoxin.<sup>28</sup> This page demonstrates another way in which the inclusion of multiple infoboxes on a Wikipedia page may generate errors:

conflation of multiple members of a category, or ignorance of some members of the category, can lead to the treatment of a single category member as synonymous with the category itself. The Wikipedia page Sarafotoxin contains three Drugboxes describing three distinct toxins (Sarafotoxins a, b, and c) of a family. However, the Wikipedia Chemical Structure Explorer harvests only the SMILES structure from the first Drugbox and associates it with the article title Sarafotoxin. It is possible that this could lead a user to interpret this as describing a single, structured chemical entity rather than a family of related structures. The Supporting Information


		<b>Ferumoxytol</b>	
<b>Names</b>		<b>Clinical data</b>	
<b>IUPAC name</b>	iron(II) iron(III) oxide	<b>Trade names</b>	Feraheme, Rienso
<b>Other names</b>	ferrous ferric oxide, ferroso ferric oxide, iron(II,III) oxide, magnetite, black iron oxide, lodestone, rust, iron(II) diiron(III) oxide	<b>AHFS/Drugs.com</b>	Monograph <a href="#">↗</a>
<b>Identifiers</b>		<b>MedlinePlus</b>	a614023 <a href="#">↗</a>
<b>CAS Number</b>	1317-61-9 <a href="#">✓</a>	<b>License data</b>	<a href="#">EU EMA: by INN</a> <a href="#">↗</a> <a href="#">US DailyMed: Ferumoxytol</a> <a href="#">↗</a>
<b>3D model (JSmol)</b>	<a href="#">Interactive image</a> <a href="#">↗</a>	<b>Routes of administration</b>	<a href="#">Intravenous infusion</a>
<b>ChEBI</b>	<a href="#">CHEBI:50821</a> <a href="#">↗</a> <a href="#">✓</a>	<b>ATC code</b>	None
<b>ChEMBL</b>	<a href="#">ChEMBL1201867</a> <a href="#">↗</a> <a href="#">✗</a>	<b>Legal status</b>	
<b>ChemSpider</b>	<a href="#">17215625</a> <a href="#">↗</a> <a href="#">✓</a>	<b>Legal status</b>	<a href="#">US: R-only</a> <a href="#">[23]</a> <a href="#">EU: Rx-only</a> <a href="#">[24]</a>
<b>ECHA InfoCard</b>	<a href="#">100.013.889</a> <a href="#">↗</a> <a href="#">✎</a>	<b>Identifiers</b>	
<b>PubChem CID</b>	<a href="#">16211978</a> <a href="#">↗</a>	<b>IUPAC name</b>	<a href="#">[show]</a>
<b>UNII</b>	<a href="#">XM0M87F357</a> <a href="#">↗</a> <a href="#">✓</a>	<b>CAS Number</b>	<a href="#">1309-38-2</a> <a href="#">↗</a>
<b>CompTox Dashboard (EPA)</b>	<a href="#">DTXSID5029639</a> <a href="#">↗</a> <a href="#">✎</a>	<b>DrugBank</b>	<a href="#">DB06215</a> <a href="#">↗</a>
<b>InChI</b>	<a href="#">[show]</a>	<b>UNII</b>	<a href="#">XM0M87F357</a> <a href="#">↗</a>
<b>SMILES</b>	<a href="#">[show]</a>	<b>KEGG</b>	<a href="#">D04177</a> <a href="#">↗</a>
<b>Properties</b>		<b>ChEBI</b>	<a href="#">CHEBI:46726</a> <a href="#">↗</a>
<b>Chemical formula</b>	Fe <sub>3</sub> O <sub>4</sub> FeO·Fe <sub>2</sub> O <sub>3</sub>	<b>CompTox Dashboard (EPA)</b>	<a href="#">DTXSID5029639</a> <a href="#">↗</a> <a href="#">✎</a>
<b>Molar mass</b>	231.533 g/mol	<b>ECHA InfoCard</b>	<a href="#">100.013.889</a> <a href="#">↗</a> <a href="#">✎</a>
<b>Appearance</b>	solid black powder	<b>Chemical and physical data</b>	
<b>Density</b>	5 g/cm <sup>3</sup>	<b>Formula</b>	Fe <sub>3</sub> O <sub>4</sub>
<b>Melting point</b>	1,597 °C (2,907 °F; 1,870 K)	<b>Molar mass</b>	231.531 g·mol <sup>-1</sup>
<b>Boiling point</b>	2,623 <sup>[1]</sup> °C (4,753 °F; 2,896 K)	<b>3D model (JSmol)</b>	<a href="#">Interactive image</a> <a href="#">↗</a>
<b>Refractive index (n<sub>D</sub>)</b>	2.42 <sup>[2]</sup>	<b>SMILES</b>	<a href="#">[show]</a>
		<b>InChI</b>	<a href="#">[show]</a>

Figure 6. Wikipedia page for Iron (II,III) oxide includes two infoboxes with contradictory information.

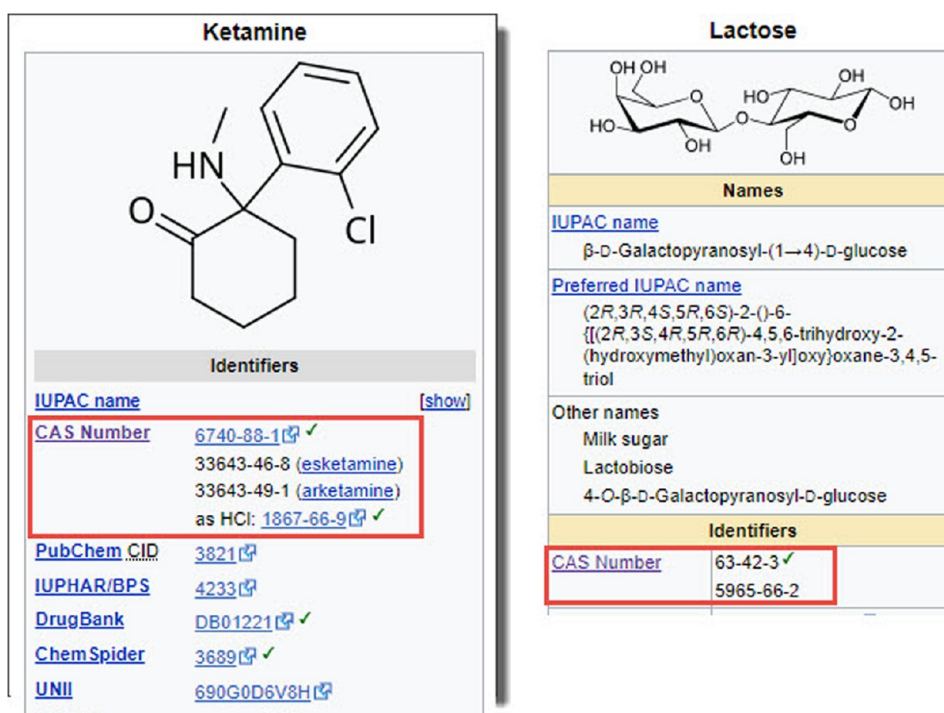
(SI\_file1\_Summary\_Statistics/SI\_file2\_Summary\_Statistics) includes the summary statistics of the Drugboxes and Chemboxes and the type of data included in each of the infoboxes.

After downloading, the page HTML was then parsed, again using jsoup, to extract identifier data from each infobox. The identifiers of interest were the EPA's DSSTox Substance ID (DTXSID), the CAS Registry Number (CAS RN), the InChIKey, and the SMILES string. A combination of Java regular expressions and string parsing was used to validate the format of all identifiers, and this data was saved in JSON format alongside page metadata. In this process, a total of 15

pages with CAS RNs failing the CAS checksum were discovered and corrected in Wikipedia by the author GS.

Individual parsing of infoboxes raises another issue of conflicting data as infoboxes may contain multiple identifiers in any field. This problem is particularly prevalent with SMILES strings, as the Wikipedia Chemical Structure Explorer project has previously noted, as well as with CAS RNs. In some cases, these multiple identifiers may be labeled, often with counterions or chiral stereochemistry, while in others, they are not clarified in any way (see Figure 7).

We also note numerous cases of missing identifiers of all types in Wikipedia. However, in many cases, the absence of an



**Figure 7.** Wikipedia articles may have multiple CAS RNs in the infoboxes. Some have associated definitions (e.g., Ketamine, left) and some have no associated definitions (e.g., Lactose, right).

identifier may be due to legitimate unavailability of that identifier for that chemical. Therefore, it becomes necessary to map the identifiers provided in Wikipedia against an external source—for our purposes, the EPA's Distributed Structure-Searchable Toxicity Database—to assess where information is truly missing or even erroneous.

**Distributed Structure-Searchable Toxicity Database and CompTox Chemicals Dashboard.** The EPA's Distributed Structure-Searchable Toxicity Database (DSSTox) has been publicly accessible since 2004 as a curated collection of chemical structure and identifier data serving the toxicology, environmental chemistry, and computational chemistry communities.<sup>42</sup> The data has been progressively expanded, with a combination of automated and human curation, to surface over 1.2 million chemicals at the latest public release via the Dashboard.<sup>5</sup> The Dashboard is the primary community-facing web application delivering access to the EPA's Center for Computational Toxicology and Exposure (CCTE) data. It provides access to experimental and predicted property data, *in vivo* and *in vitro* toxicology data, and exposure and safety data, as well as real-time prediction capabilities to serve environmental scientists, making it a rich data source for risk assessors.<sup>29</sup> Search capabilities include chemical identifiers (e.g., DTXSID, CAS RN, name); chemical structures; gene or assay (associated with measured bioactivity data); and product use category, as well as batch search.<sup>6</sup>

As explained earlier, the inclusion of a Wikipedia lede (or snippet) is an easy way to provide a basic description for a chemical as well as link to a richer article within Wikipedia itself. Each chemical substance is associated with an EPA DSSTox Substance ID (DTXSID). The DTXSID is an accepted Wikidata identifier (<https://www.wikidata.org/wiki/Property:P3117>) and has enabled linking between Wikipedia and the Dashboard using a simple URL formatted as <https://comptox.epa.gov/dashboard/DTXSIDnumber>. The identifier

can represent an organic or inorganic chemical structure (e.g., fluconazole<sup>30</sup>), a polymer (e.g., polyvinylpyrrolidone<sup>31</sup>), a class of chemicals (e.g., polychlorinated biphenyls (PCBs)),<sup>32</sup> a complex biological drug (e.g., monoclonal antibody drug such as Canakinumab<sup>33</sup>), or a mineral (e.g., Cummingtonite<sup>34</sup>). Thousands of DTXSID identifiers were added to Wikidata based on exact InChIKey match<sup>35</sup> using a script written in Groovy using Bacting extensions.<sup>36,37</sup>

**DSSTox Data Mapping.** The harvested Wikipedia data were mapped against the EPA DSSTox database of identifiers and structure data to provide a more complete picture of the accuracy of the dataset. All mapping was performed automatically using SQL (MySQL v5.1.49) via Java Database Connectivity (JDBC).<sup>38</sup>

To maintain the organization of the data provided by Wikipedia, mapping was conducted "infobox by infobox," presuming that each infobox on a page might represent a distinct chemical. For each infobox, multiple mappings with different identifiers were considered, to allow identification of as many potential discrepancies in the data as possible. These mappings were ordered by priority.

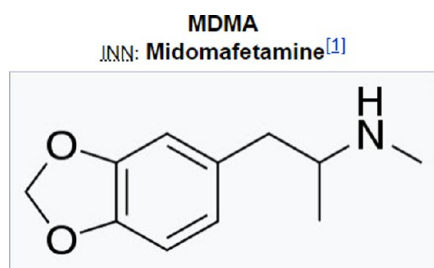
1. DTXSID: If one or more DTXSIDs were provided in the infobox, the associated DSSTox records were added to the mappings for that infobox.
2. Preferred name: If the chemical name given in the title of the infobox matched the preferred name of a DSSTox record, and that record had not been previously mapped by DTXSID, the record was added to the mappings for that infobox.
3. CAS RN: If one or more CAS RNs were provided in the infobox, and those CAS RNs matched DSSTox records, any new such DSSTox records were added to the mappings for that infobox.
4. Other CAS RN: Depreciated CAS RNs matched in DSSTox were also considered.

- Synonym: If the chemical name given in the title of the infobox matched a synonym for a DSSTox record(s), any such new DSSTox records were added to the mappings for that infobox.
- InChIKey: If one or more InChIKeys were provided in the infobox, and those InChIKeys matched structures for DSSTox records, then the mapping between the DSSTox records and the associated Wikipedia articles were added.
- SMILES: If one or more SMILES strings were provided in the infobox, and those SMILES strings matched structures for DSSTox records (after conversion to InChIKey using the Indigo Toolkit<sup>39</sup>), any new such DSSTox records were added to the mappings for that infobox.
- Name-to-Structure: The chemical name given in the title of the infobox was converted to an InChIKey using OPSIN<sup>40,41</sup> and the Indigo Toolkit.<sup>39</sup> If the structure so obtained matched a structure for a DSSTox record, and that record had not been previously mapped, that record was added to the mappings for that infobox.

These are the same automated mapping strategies that underlie manual DSSTox data input to assist the human curation team in locating and resolving conflicts.<sup>42</sup>

Synonyms or other chemical names provided in the Wikipedia data, such as IUPAC systematic names, were not considered due to the likelihood of errors in the presentation across a large dataset: the format of a chemical name cannot be as efficiently checked as a CAS RN or InChIKey, as the latter have fixed formats. Future work may include remapping with the inclusion of additional chemical names.

The use of infobox titles for mapping, while fallible as described (see Figure 8), was considered to add some value to

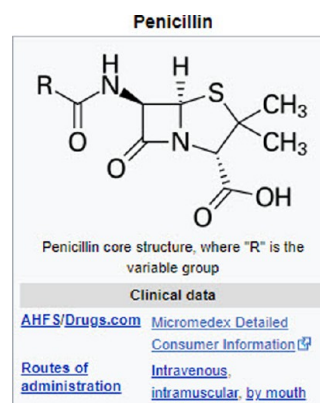


**Figure 8.** Infobox name of MDMA is “MDMA INN: Midomafetamine<sup>1</sup>”: although visually distinct, multiple names are not distinguished within the text field.

the data due to the presence of “degenerate” cases where the infobox title was the only available identifier. As an example, the Wikipedia article for penicillin is clearly of key historical and medical importance. However, no identifying information—DTXSID, CAS RN, InChIKey, or SMILES—is provided beyond the title (see Figure 9). The title alone allowed us to map this chemical class; thus, the inclusion of title-based mapping, while imperfect, is important to describe the dataset appropriately.

#### Wikipedia CAS RN-Compound Linkage Harvesting.

Whereas the primary data associated with this project were obtained using the processes described above, we chose to harvest data via other extraction procedures directly from Wikipedia to see whether it would expand the dataset already identified via the infobox dataset. As an example, Wikipedia



**Figure 9.** Penicillin Wikipedia article discusses a set of chemicals. Penicillin is represented in the CompTox Chemicals Dashboard as DTXSID501017130.

maintains a number of lists of mappings of chemical formulae and compound names to CAS RNs<sup>43</sup> that were relatively easy to harvest. The HTML contents of these list pages were downloaded and parsed using jsoup to obtain a list of JSON objects containing the following information.

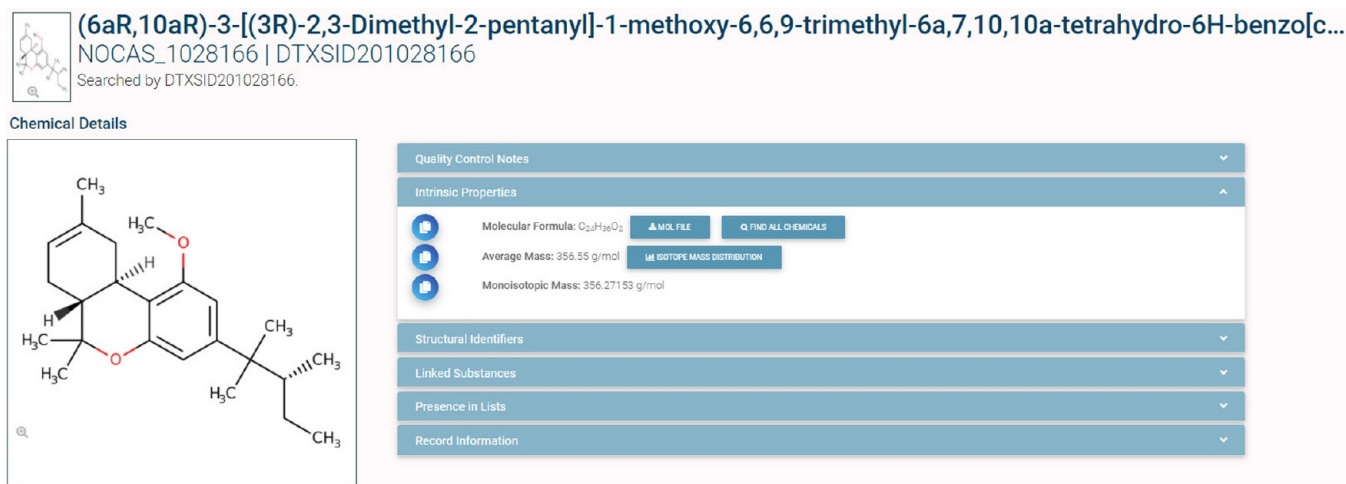
- Chemical formula: The chemical formula string provided by Wikipedia.
- Title: The name of the chemical compound, used as the title of the Wikipedia page for the compound.
- Page exists: Whether an active page with the given title exists in Wikipedia.
- CAS RN: The CAS RN provided by Wikipedia.

From the “List of CAS numbers by chemical compound”,<sup>44</sup> consisting almost entirely of non-carbon-containing inorganics, 1973 records were retrieved. Among these inorganic compounds, 55 were found to exist in Wikipedia and not in DSSTox, whereas 156 were found to exist both in Wikipedia and in DSSTox but without a mapping. The 55 missing from DSSTox were therefore registered and curated. In total, 211 mappings were created for both the new and the existing set. Also, 437 compounds were found to exist in DSSTox and not Wikipedia. This set may become a useful seed set for future Wikipedia articles since the CAS RN, name, SMILES, InChI strings and keys, and structure image can be harvested to provide data for the Chemboxes in new stub articles.

## DISCUSSION

Mapping Wikipedia articles to DSSTox allowed for indepth analysis of the consistency and correctness of information provided in Wikipedia. The first analysis pertains once again to the problem of multiple identifiers. In total, 1082 Drugboxes and 1805 Chemboxes generated more than one DSSTox hit when mapped with the process described above. The maximum among the Drugbox set was for the single infobox on page “Testosterone (medication)”,<sup>45</sup> which mapped to the parent structure of testosterone as well as four ester derivatives commonly used medically. The maxima among the Chembox set were for the single infoboxes on pages Fumiquinazoline and Chlorophyllin, both of which generated seven DSSTox mappings. In the former case, this was due to the presence of multiple identifiers in every field for Fumiquinazolines A–F, as this infobox represented an entire class of natural products. In the latter, confusion was generated within DSSTox, as “Chlorophyllin” was listed as an ambiguous synonym for six





**Figure 10.** Example where replacement of the long systematic name for the chemical, in this case to JWH-359, provides an improved esthetic for the page.

different records within DSSTox, and an additional match was located based on the SMILES structure provided by Wikipedia.

Because of the myriad possible reasons for the appearance of multiple identifiers or multiple DSSTox mappings, further analysis of the presence or correctness of identifiers is restricted to those cases in which a single, unambiguous DSSTox mapping was generated. This encompasses 76.7% of the Drugbox set and 83.9% of the Chembox set.

First, we reconsider the issue of missing identifiers. Some chemicals that appear to be “missing” CAS RNs may in reality have no assigned CAS RN; likewise, chemicals that appear to be missing structures may be mixtures or “unstructurable” substances, which cannot be represented as distinct chemical structures, but may still be of chemical interest. Without mapping, 617 Drugboxes appear to be missing CAS RNs; however, if we exclude compounds for which DSSTox does not provide a CAS RN either, this number drops to 123. In the Chembox set, we see a decrease from 510 to 153 defects. In the Drugbox set, the percentage of naive missing identifiers that resolve to true missing identifiers ranged from 13.6 to 22.3%, while in the Chembox set, it ranged from 30.0 to 48.8%. These sets may serve as seeds for the automated insertion of data into Wikipedia to complete chemical identifying information where possible.

We were also able to analyze the consistency of identifiers in Wikipedia. As previously noted, we discovered and remedied 15 instances of CAS RNs in Wikipedia that failed the required checksum, all of which appeared to be caused by human inattention in data insertion (e.g., substitution of 1492-02-2 for 1492-02-0 for Glybuzole, or transposition of 209349-27-4 for 209394-27-4 for Ladostigil). Bot-driven insertion of CAS Common Chemistry links already occurs in Wikipedia, and it may be fruitful to consider implementation of checksum or URL validation in this process. Similarly, we discovered 37 Drugboxes and 79 Chemboxes containing unparseable SMILES strings in our datasets. This issue has been previously identified by the Wikipedia Chemical Structure Explorer project.

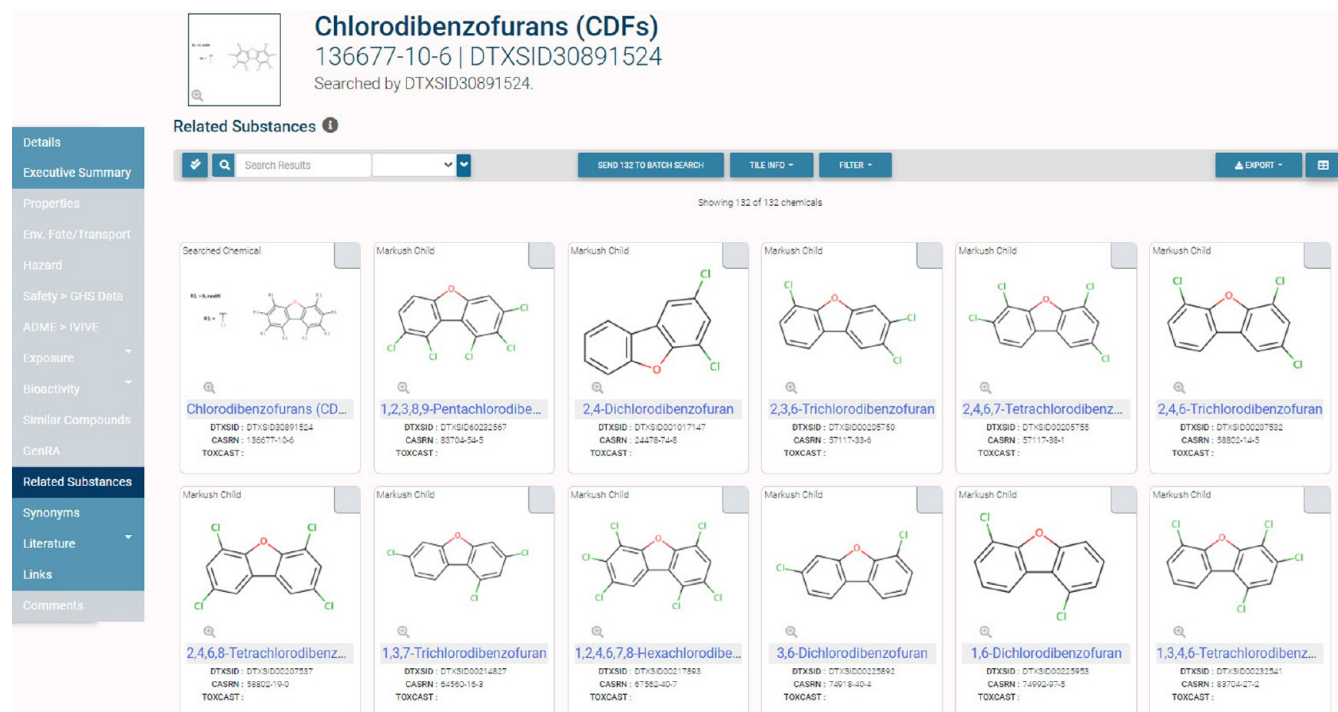
Even where identifiers were correctly formatted and interpretable, we discovered inconsistencies. The problems we considered included disagreement of CAS RNs in Wikipedia and DSSTox; disagreement of InChIKeys with DSSTox structures; disagreement of SMILES with DSSTox

structures; and disagreement of InChIKey and SMILES within Wikipedia. Structural disagreement with DSSTox was the largest problem, with either InChIKey or SMILES disagreeing with DSSTox in at least 10% of both sets; there was structural disagreement within Wikipedia in 4–5% of both sets. CAS RN disagreement was a less significant issue, occurring in roughly 1.5% of both sets.

The primary motivation in characterizing these inconsistencies was to develop a list of mappings without inconsistencies that could be used to automatically load mappings from DSSTox to Wikipedia pages. However, these results will also allow for detailed examination and resolution of errors in both Wikipedia and DSSTox by human curators, work that has already begun. Both the overarching goal of automated loading and certain specific instances of human curation will be discussed in depth below.

**Addition of Synonyms and Preferred Names.** As a result of the curation effort, hundreds of synonyms that were listed across the Wikipedia articles, but not within DSSTox, were curated and added into the DSSTox database. This expanded the coverage and availability of chemical names and synonyms for searching via the Dashboard. Emphasis was made on switching Preferred Names in DSSTox from systematic names to shorter, more appropriate names commonly matching those for the Wikipedia article. For example, the preferred name for JWH-359<sup>46</sup> was originally (6*aR*,10*aR*)-3-[(3*R*)-2,3-Dimethyl-2-pentanyl]-1-methoxy-6,6,9-trimethyl-6*a*,7,10,10*a*-tetrahydro-6*H*-benzo[*c*]chromene, but was edited to the shorter JWH format during this effort. This is beneficial in many cases because systematic names can be very long, occasionally overflowing database field length limits and resulting in errors or truncation, either within the database or on export. Additionally, from a user-experience perspective, these names may spill outside of the Dashboard web page display, as exemplified in Figure 10, rendering information gathering and collation difficult for the user.

**Incorporation of New Wikipedia Links in DSSTox.** Prior to the initiation of this work, 13,454 chemicals were included in the Dashboard Wikipedia list<sup>47</sup> and the resulting linkages in DSSTox were used to ensure that the Wikipedia lede was displayed in the Dashboard interface. However, from the original complete Chembox and Drugbox datasets, 1157



**Figure 11.** Set of related substances associated with polychlorinated dibenzofurans. The Markush representation (Tile 1, top left-hand side) is enumerated using cheminformatics approaches to generate the set of ~130 Markush children.

nonambiguous, nonconflicted mappings were identified where the appropriate linkages had not been made. An import file was generated and these linkages were uploaded, immediately increasing the availability of Wikipedia information via DSSTox by 7%. Among others, the new linkages included a number of common pharmaceuticals such as cimetidine, betadine, and metronidazole, for which easy access to comprehensive data is vital for researchers and consumers alike. A further 105 mappings were added automatically when the dataset was updated prior to publication, and over 4000 more conflicted mappings (as described earlier) were reviewed and resolved by human curators, providing a total of 19,236 Wikipedia article mappings in DSSTox as of 7/10/2022.

**Mapping of Dashboard Chemical Categories.** A number of important chemical categories warrant their own descriptive Wikipedia articles and are also registered as distinct chemical substances on the Dashboard, many via Markush-type structure representations. These include polycyclic aromatic hydrocarbons (PAH, DTXSID3044043),<sup>48</sup> dioxins and dioxin-related compounds (DTXSID201020548),<sup>49</sup> and polychlorinated dibenzofurans (DTXSID30891524).<sup>50</sup> Although the category-level description is of value (as shown in Figure 11 for polychlorinated dibenzofurans), the benefits of the cheminformatics approaches supported by the Dashboard additionally include enumeration of the Markush structural representations to generate a list of definite chemical structures belonging to the category, displayed under the associated Related Substances<sup>51</sup> tab.

This approach has been used to create accessible, searchable summaries of EPA's Toxics Release Inventory (TRI) chemical categories. The EPA TRI program is a resource providing information about toxic chemical releases and pollution prevention activities reported by industrial and federal facilities (<https://www.epa.gov/toxics-release-inventory-tri-program>).

The assembled list of chemicals is available via the Dashboard.<sup>52</sup>

**Investigation of the John W. Huffman Cannabinoid Dataset.** Synthetic cannabinoids are an important part of the effort to build an appropriate library of chemicals to support nontargeted mass spectrometry efforts at the EPA,<sup>53–56</sup> especially with regard to rapid response scenarios where cannabinoid synthesis is a common exposure scenario for responders.<sup>57</sup> The set of synthetic cannabinoid chemicals discovered by John W. Huffman (henceforth JWH) was of particular interest to continue growing the list of synthetic cannabinoids available via the Dashboard.<sup>58</sup> A number of street drugs from the JWH collection are on the List of Schedule 1 drugs<sup>59</sup> and new derivatives continue to emerge over time. Having the library of JWH synthetic cannabinoids registered and available to utilize in the EPA nontargeted workflows would allow for inclusion in our *in silico* mass spectrometry library<sup>60,61</sup> and for the potential identification of known unknowns<sup>61–63</sup> in the synthetic cannabinoid class.

Wikipedia lists 55 chemical pages under the “Category: JWH cannabinoids”<sup>64</sup> out of a total of over 450 cannabinoids that were synthesized as part of a research study. All of these pages were included in the data harvested from Wikipedia, with 53 appearing in the Drugbox dataset and two in the Chembox dataset (JWH-251<sup>65</sup> and JWH-167<sup>66</sup>). Additionally, each of these pages was confirmed to have only a single infobox; however, two pages (JWH-210<sup>67</sup> and JWH-203<sup>68</sup>) included multiple CAS RNs. The page for JWH-210 also contained the CAS RN for JWH-182, and the page for JWH-203 also contained the CAS RN for JWH-204. No structures or other identifiers for JWH-182 or JWH-204 were provided in those pages, and no explanation for these faulty inclusions was given. No other instances of multiple identifiers were observed in the dataset. Of the remaining 53 pages with singular identifiers, 44 were successfully mapped to DSSTox, and nine were left

Allotropes of carbon <span style="float: right;">[hide]</span>	
<b>sp<sup>3</sup> forms</b>	Diamond (cubic) · Lonsdaleite (hexagonal diamond)
<b>sp<sup>2</sup> forms</b>	Graphite · Graphene · Fullerenes, including C <sub>60</sub> (buckminsterfullerene), C <sub>70</sub> , Fullerene whiskers, Nanotubes, Nanobuds, Nanoscrolls · Glassy carbon
<b>sp forms</b>	Linear acetylenic carbon · C <sub>18</sub> (cyclo[18]carbon)
<b>mixed sp<sup>3</sup>/sp<sup>2</sup> forms</b>	Amorphous carbon · Carbon nanofoam · Carbide-derived carbon · Q-carbon
<b>other forms</b>	C <sub>1</sub> (atomic carbon) · C <sub>2</sub> (diatomic carbon) · C <sub>3</sub> (tricarbon)
<b>hypothetical forms</b>	C <sub>3</sub> (cyclopropatriene) · C <sub>5</sub> (benzotrityne) · C <sub>8</sub> (prismane C8) · Chaotite · Haeckelites · Cubic carbon · Metallic carbon · Penta-graphene
<b>related</b>	Activated carbon · Carbon black · Charcoal · Carbon fiber · Aggregated diamond nanorod

Figure 12. List of allotropes of carbon listed in the associated Wikipedia article.

unmapped. Of the 44 mapped, 31 were identified by the DTXSID provided in Wikipedia; two were identified by a match to a DSSTox preferred name; eight were identified by CAS RN; and three were identified by a match to a DSSTox synonym.

A total of 30 of the mapped pages were mapped without conflicts, and 14 had identified conflicts. Of the latter, 12 were missing a DTXSID in Wikipedia, whereas two had structure conflicts. One of the structure conflicts, JWH-359,<sup>69</sup> also lacked a DTXSID in Wikipedia. Additionally, although the InChIKey provided by Wikipedia for JWH-359 matched the structure in DSSTox, the SMILES string provided by Wikipedia disagreed with both the Wikipedia InChIKey and the DSSTox structure. The other structure conflict, JWH-176,<sup>70</sup> differed from the DSSTox structure in both InChIKey and SMILES: the Wikipedia-provided InChIKey and the DSSTox structure described two different stereoisomers, whereas the Wikipedia-provided SMILES was nonisomeric.

This investigation allowed numerous corrections to be made to complete the JWH cannabinoid dataset. An import file with Wikipedia-provided identifiers was generated for the nine unmapped Wikipedia pages that was then curated and those substances added to DSSTox. Additionally, seven Dashboard pages for JWH substances were identified that did not link back appropriately to Wikipedia pages; again, an import file was generated and these links were added.

Considering the ambiguous and conflicted pages, manual curation was required to evaluate and make corrections. Details of the identified conflicts and corrections applied are provided in the Supporting Information (SI\_file3\_JWH\_Cannabinoids).

In a related fashion, a list of fentanyl analogues was harvested from the “List of fentanyl analogues”<sup>71</sup> page and the entire list was harvested, mapped, and missing chemicals were added to DSSTox.

**Where Manual Mapping is Required.** During this work, it became obvious that whereas the process of harvesting data from Drugboxes and Chemboxes is highly beneficial for the purpose of mapping, there are still Wikipedia article pages where this process will fail; manual curation and mapping is still required for such pages. A key example is the allotropes of carbon page<sup>72</sup> that includes Graphite, Graphene, Graphane, Graphyne, Diamond, and a set of fullerenes (as shown in Figure 12).

A number of allotropes of carbon that were not in DSSTox were registered and mapped to the relevant Wikipedia articles. These included C20-Fullerene (DTXSID301319413), C28-Fullerene (DTXSID001319414), C70-Fullerene (DTXSID90151050), C76-Fullerene (DTXSID401319416), C82-Fullerene (DTXSID801319418), C84-Fullerene

(DTXSID701319415), Carbon Black (DTXSID7051216), Carbon nanotubes (DTXSID301020377), Carbon nanofoam (DTXSID801336804), Lonsdaleite (DTXSID101336803), Cyclo(18)carbon (DTXSID801267863), and Graphyne (DTXSID201336806).

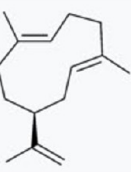
There are many more allotropes of carbon listed on the Wikipedia category page ([https://en.wikipedia.org/wiki/Category:Allotropes\\_of\\_carbon](https://en.wikipedia.org/wiki/Category:Allotropes_of_carbon)<sup>73</sup>) but many of these are hypothetical allotropes and not deemed relevant to include into the Dashboard.

**Multiple Chemicals Mapped to a Single Article.** As discussed above, a number of Wikipedia pages are mapped to various categories and subcategories of chemicals. Some categories are sufficiently expansive and structurally diverse as to not have a set of explicit chemical members listed on the page (e.g., polycyclic aromatic hydrocarbons (PAHs)). The list of PAHs<sup>74</sup> on the dashboard has a defined list of 73 representative members but these cannot be represented by a simple structural representation relating all of them. In other cases, however, the category is amenable to a Markush-type structure representation yielding a limited and finite number of members (e.g., polychlorinated biphenyls). Again, in this case, there may be no specific list of chemicals provided in the article (i.e., no list of distinct CAS RN, names, etc.). On the other hand, there are many category-based articles where there are explicit listings of chemicals included in the article.

The article regarding Germacrene<sup>75</sup> lists four explicit forms of the chemical in the Chembox (see Figure 13)—Germacrene A, B, C, and D. These are all related as isomers based on double bond placement and orientation in the ring as well as explicit stereocenters. Each of the four forms are mapped to the Germacrene article so that on the Dashboard, the four isomeric forms all display the category article description.

Four different nonylphenol substances were mapped to the Nonylphenol Wikipedia article.<sup>76</sup> Even though a much larger collection of nonylphenols in the Dashboard could be mapped, 32 in total,<sup>77</sup> subjective selection was made to map only the four listed in the Chembox in the nonylphenol article.

Cases of multiple chemicals from the Dashboard mapped to a single Wikipedia article are especially prominent for active pharmaceutical ingredients and their salt forms, where it is quite common to have both listed in their Drugbox. An example is Trefentani,<sup>78</sup> where both the neutral chemical and the hydrochloride salt<sup>79</sup> on the Dashboard both show the ledes from the Wikipedia article. Although there are many possibilities to extend such mappings between salts on the Dashboard with articles about the neutral chemical on Wikipedia, this is not being pursued at this time. For example, perfluorooctanesulfonic acid (PFOS) has 11 salts displayed<sup>80</sup> on the Dashboard, but only the neutral PFOS compound is

Germacrene A	
	
Names	
<b>IUPAC name</b>	(1E,5E,8S)-1,5-dimethyl-8-(prop-1-en-2-yl)cyclodeca-1,5-diene
<b>Other names</b>	(-)-Germacrene A (E,E)-Germacra-3,9,11-triene (1E,5E,8S)-1,5-dimethyl-8-(1-methylethenyl)-1,5-cyclodecadiene (S-(E,E))-1,5-dimethyl-8-(1-methylethenyl)-1,5-cyclodecadiene
Identifiers	
<b>CAS Number</b>	28387-44-2 ✓ B: 15423-57-1 C: 34323-15-4 D: 37839-63-7

**Figure 13.** ChemBox associated with Germacrene lists four explicit isomeric forms.

mapped to the Wikipedia article. The general rule that is adhered to is the mapping between the Dashboard content and Wikipedia articles based on the presence of CAS RNs in the Chembox or Drugbox. Other examples of multiple chemicals mapped to a single Wikipedia article include multiple mappings of coke fuel-related materials and Kraton polymers.<sup>81</sup> As explained earlier, one of the advantages of having the lede of the Wikipedia article immediately visible for a particular substance record is that it generally gives some context regarding the nature of the chemical and its applications. As shown in Figure 14, having the lede visible for the four different Kraton-related chemicals visible in the Dashboard certainly benefits the user in terms of having an obscure name (i.e., Kraton Liquid L 1203) be defined as a member of a family of high-performance elastomers.

**Polymers.** Polymers are of interest to the EPA as there are thousands of them in commerce, and enormous quantities of polymers and their related fillers, plasticizers, and colorants find their way into the environment. Over 6000 polymers are listed in the Toxic Substances Control Act (TSCA) inventory. Polymers are a particularly challenging case for the mapping between DSSTox chemicals and Wikipedia for a number of reasons. Many of the chemicals in the organic polymer

category<sup>82</sup> have detailed Wikipedia articles, but will not be found in the Wikipedia Chemical Structure Explorer, as there are no associated SMILES. The alternative may also be true where a polymer does have an explicit SMILES (e.g., for Poly(methyl methacrylate)<sup>83</sup> and Polydimethylsiloxane<sup>84</sup> that specify only a single oligomeric length).

The polymers in Wikipedia may or may not have an associated CAS RN and may have an associated Chembox (e.g., Polyglycerol polyricinoleate<sup>85</sup>) or none at all (e.g., polyallylamine hydrochloride<sup>86</sup>). There are also polymer category articles that may be useful article mappings for some chemicals (e.g., polyorthoester<sup>87</sup> or polydiacetylenes<sup>88</sup>). There are numerous cases where a single polymeric substance on the Dashboard mapped to a Wikipedia article is a co- or terpolymer where the mapping to the individual monomers is feasible. For example, the Wikipedia article regarding Acrylonitrile butadiene styrene (ABS)<sup>89</sup> has three related monomers. Using the capability of the DSSTox chemical registration system to add “successor substances” to a registered substance, these monomers were registered and are viewable in the Dashboard as Related Substances (see Figure 15). This process has been applied to many other co- and terpolymers and continues iteratively to expand the coverage of such mappings.

Polymers and their association with their related monomeric units is of particular value at this time in relation to the research efforts of the Agency to identify the potential impacts of PFAS (Per- and polyfluoroalkyl substances) on human health. Mapping PFAS monomeric units in polymers (e.g., hexafluoropropene) is beneficial in terms of identifying potential substructural moieties that could result from environmental degradation of the fluoropolymer.

## ■ FUTURE WORK

The work outlined in this article has highlighted limitations in the present data model supporting mapping of Dashboard content to Wikipedia articles. For example, two Wikipedia articles cannot be mapped against one chemical as in the case of Adrenaline as a hormone<sup>90</sup> and Epinephrine as a medication,<sup>91</sup> both based on the same chemical structure. Future enhancements to the Dashboard data model could be introduced to support mapping of multiple articles to a single substance. This could be of value to allow, for example, a polymeric substance to be mapped to the polymer article as well as all monomer-related articles.

# Kraton Liquid L 1203

## 191617-93-3 | DTXSID201059436

Searched by DSSTox Substance Id.

### Wikipedia

**Kraton** is the trade name given to a number of high performance elastomers manufactured by Kraton Polymers (NYSE: KRA), and used as synthetic replacements for rubber. Kraton polymers offers many of the properties of natural rubber, such as flexibility, high traction, and sealing abilities, but with increased resistance to heat, weathering, and chemicals. It was first made by the chemical division of the Shell Oil Company in the 1950s, under the technical leadership of Murray Luftglass

...

[Read more](#)

**Figure 14.** Lede associated with the Wikipedia article regarding Kraton polymers is displayed on four separate substances.

## Acrylonitrile Butadiene Styrene

9003-56-9 | DTXSID70858757

Searched by CAS-RN.

The screenshot shows a dashboard for the chemical Acrylonitrile Butadiene Styrene (DTXSID:DTXSID70858757, CASRN:9003-56-9). The interface includes a search bar with filters for DTXSID, CASRN, and TOXCAST. Below the search bar are four panels:

- Searched Chemical:** 3 related chemical structures with this substance.
  - Acrylonitrile Butadiene Styrene (DTXSID:DTXSID70858757, CASRN:9003-56-9, TOXCAST:-)
- Monomer:** Acrylonitrile (DTXSID:DTXSID5020029, CASRN:107-13-1, TOXCAST:10/434). Structure: C=CC#N
- Monomer:** 1,3-Butadiene (DTXSID:DTXSID3020203, CASRN:106-99-0, TOXCAST:-). Structure: C=CC=CC
- Monomer:** Styrene (DTXSID:DTXSID2021284, CASRN:100-42-5, TOXCAST:1/235). Structure: C=Cc1ccccc1

**Figure 15.** Related substance mappings between the monomers of acrylonitrile butadiene styrene and the polymeric form allow for display of the related monomers in the Dashboard.

In its simplest form, the mapped dataset produced as an output of this work will be between substance identifiers (DTXSIDs) and Wikipedia article titles. It is this mapping that allows for the article leads to be loaded, in real time, into the Dashboard interface. This dataset can provide the necessary path to allow for the bulk loading of DTXSIDs into Chemboxes or Drugboxes so that Wikipedia articles can link back to the Dashboard, as is done for other public databases such as PubChem, ChemSpider, and many others. In addition, a rich collection of metadata associated with each DTXSID can be used to check or enhance data in Wikidata, as well as in Wikipedia itself. For example, we have observed that the SMILES strings listed in Wikipedia for certain substances listed in the Chemboxes or Drugboxes may be SMILES with stereochemistry removed rather than the full isomeric forms with inclusive stereochemical details. The Supporting Information provided ([SI\\_file4\\_DSSTox\\_WIKIPEDIA\\_List](#)) includes isomeric SMILES that can be populated for any records missing such entity representations, as well as, additionally, InChI Strings, InChIKeys, molecular formulae, mass, IUPAC Name, etc. It is acknowledged that the cross-referencing between data on Wikipedia, in Wikidata, and in the Dashboard, may require additional manual curation, but we believe that this iterative process will ultimately arrive at a consensus agreement regarding the chemical representation(s) and the associated metadata and benefit the overall community of Wikipedia users.

The effort expended to this point on the curation of the Wikipedia chemicals set will provide a stable dataset for only a very short period of time as chemicals are being continuously added. Iterative edits and curation of the Wikipedia chemicals set will not cease, especially as new chemicals are added on an ongoing basis. The creation of new Wikipedia chemical pages peaked in 2007, with 2140 pages being added across the Chembox and Drugbox datasets in that year; the peak for the Chembox dataset was slightly earlier, in 2006, with 1369 pages added, whereas the peak for the Drugbox dataset also came in 2007, with 999 pages added. Although the growth of the datasets has slowed considerably since, appreciable additions are still being made: the Chembox dataset averaged 31 new pages every month in 2021 (median 32), while the Drugbox

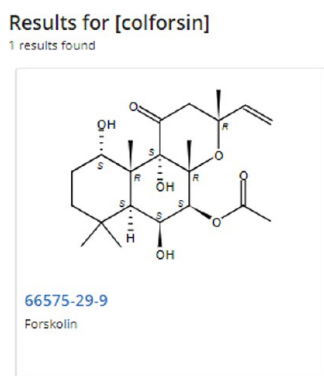
dataset averaged 15 new pages (median 13). The trend in both datasets is constant or increasing, which further indicates the importance of developing systems to regularly extract and synchronize data across sources. Extraction and synchronization are only parts of the challenge, however, as curation and validation of the data will remain an ongoing issue.

This article has referenced Wikidata but has not yet emphasized the strong connection between Wikidata and Wikipedia. Originally, Wikidata was introduced as a solution to reduce duplication of facts among different language Wikipedias. After all, the SMILES or boiling point is the same in the English Wikipedia as in the Dutch, Swedish, or Russian Wikipedia. Sitelinks connect an English Wikipedia page to a Wikidata page, which in turn links to other language-based Wikipedias. This powers the functionality to easily open the Wikipedia page in another language. A second important goal of Wikidata is to store facts that can be shared in any Wikipedia or used in other platforms when called using the Wikidata API. These data are already used in ChemBoxes and, for example, the E number, ECHA InfoCard ID, and DTXSID substance identifier are populated from Wikidata. This requires, obviously, an accurate linkage of Wikipedia with Wikidata, which is complicated when a single Wikipedia page or single ChemBox describes more than one chemical entity. Wikidata starts from the design goal that every Wikidata page represents a single chemical with a single InChI. Several efforts to improve the chemicals in Wikidata resulted in a collection of around 1.3 million chemical entities, linking to many different chemical and biological databases.<sup>92–94</sup>

An example where this approach would be useful in ensuring that data is updated across multiple Wikipedia sites is exemplified in a real-world example in our laboratory. One of our colleagues noted that, in the Dashboard, a search for Forskolol returned the chemical with the preferred name of Colforsin. Forskolol is one of the chemicals of interest in our *in vitro* bioactivity screening studies under the ToxCast project<sup>95,96</sup> and is registered in DSSTox and available on the Dashboard.<sup>97</sup> The presence of both names as synonyms in the Dashboard suggests that they are alternative names for the same chemical. However, in contradiction, two distinct articles existed on Wikipedia with the chemical names of Forskolol and

Colforsin as titles. The chemical structures for the two chemicals were distinct and, to add further confusion, both pages contained the same CAS RN, 66575-29-9. The historical articles are available as Forskolin<sup>98</sup> and Colforsin.<sup>99</sup>

Correct identification of the structure of Forskolin is critical for accurate reporting of our *in vitro* data. Investigative work indicated that the chemical name of Colforsin searched on Common Chemistry<sup>100</sup> returned a chemical named Forskolin and the CAS RN mapped to both Wikipedia articles (see Figure 16).



**Figure 16.** Search in Common Chemistry for Colforsin returns a chemical with the preferred name of Forskolin.

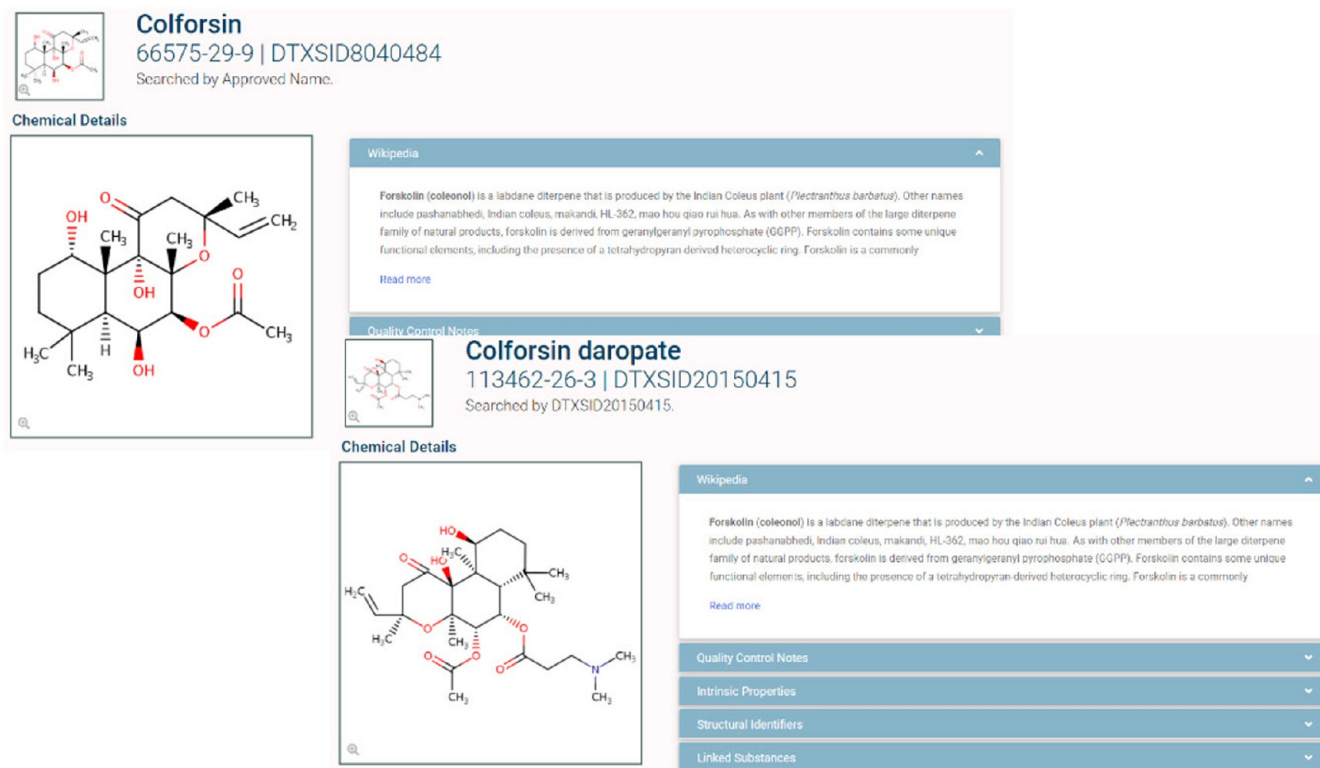
Further iterative inspection and review of the data contained in Common Chemistry, in the Dashboard, on PubChem, and on Wikipedia ultimately showed that the error was in Wikipedia itself. The chemical incorrectly described in the

Wikipedia page titled Colforsin was actually Colforsin daropate (a carboxylic ester of Forskolin) but was annotated with the incorrect name and CAS RN. These errors in Wikipedia translated to having the same Wikipedia article ledes showing against two different chemicals as shown in Figure 17.

The final aspects of this detailed curation effort involved cleaning up the errors in Wikipedia so that the ledes would be correctly displayed in the Dashboard *and* benefit all Wikipedia users directly. One of the authors (GS) joined as an editor and resolved the errors in both articles. The articles were remapped in DSSTox and, ultimately, will refresh in a future Dashboard release. The Wikidata record was also updated,<sup>101</sup> impacting all related Wikipedia pages.

## CONCLUSIONS

The importance of Wikipedia as an encyclopedic source of both the history of and data associated with almost 20,000 chemicals is exemplified by its usage. Wikipedia articles about chemicals are linked to by many popular chemistry websites (e.g. PubChem, ChemSpider, the CompTox Chemicals Dashboard) as the descriptions and rich content are unique relative to the more limited data-centric-only websites that have proliferated in recent years. There have been previous efforts, specifically via the Wikipedia Chemical Structure Explorer, to account for chemical substances on Wikipedia that are represented in Chemboxes and Drugboxes. This approach depends on the harvesting of SMILES strings to assemble the list of chemicals. Our approach has benefited from this historic work but has been extended to ensure that multiple Chemboxes and Drugboxes on a single Wikipedia page can be harvested and extracted. Cross-referencing multiple identifiers harvested from the pages provided a path



**Colforsin**  
66575-29-9 | DTXSID8040484  
Searched by Approved Name.

**Colforsin daropate**  
113462-26-3 | DTXSID20150415  
Searched by DTXSID20150415.

Wikipedia  
Forskolin (coleonol) is a labdane diterpene that is produced by the Indian Coleus plant (*Nectanthus barbatus*). Other names include pashanabedi, Indian coleus, makandi, HL-362, mao hou qiao rui hua. As with other members of the large diterpene family of natural products, forskolin is derived from geranylgeranyl pyrophosphate (GGPP). Forskolin contains some unique functional elements, including the presence of a tetrahydropyran derived heterocyclic ring. Forskolin is a commonly

Wikipedia  
Forskolin (coleonol) is a labdane diterpene that is produced by the Indian Coleus plant (*Nectanthus barbatus*). Other names include pashanabedi, Indian coleus, makandi, HL-362, mao hou qiao rui hua. As with other members of the large diterpene family of natural products, forskolin is derived from geranylgeranyl pyrophosphate (GGPP). Forskolin contains some unique functional elements, including the presence of a tetrahydropyran-derived heterocyclic ring. Forskolin is a commonly

**Figure 17.** Result of a search for the chemical name “forskolin” on the CompTox Chemicals Dashboard returned a chemical Colforsin (top left). Colforsin daropate (bottom right) displays the same lede as Colforsin: the lede is Forskolin in both cases.

to enhance our DSSTox dataset with Wikipedia mappings, as well as identify issues requiring manual curation on DSSTox and highlighting a number of issues on Wikipedia.

Manual curation and mapping have enhanced the DSSTox data collection to include the Supporting Information on polymers, monoclonal antibodies, amphiboles and other minerals, and other substances where SMILES strings cannot be harvested from Wikipedia. This is relevant to the work of the EPA as tens of thousands of chemicals of interest to the Agency cannot be represented as distinct chemical structures but are registered as unique substances in the DSSTox database. This mapping work provides the benefit of incorporating the display of the Wikipedia lede from the article into the Dashboard. Simple examples would be Tall Oil, Asbestos, Tannin, Zylon, and a myriad of other non-structurable examples that would not be harvested and mapped using only the approach of SMILES harvesting.

The ongoing expansion and curation of the DSSTox database to assist the research needs of the EPA continues unabated, and at a much faster rate than Wikipedia. We continue to add chemical substances from regulatory lists, peer-reviewed publications, datasets related to chemicals of emerging concern, and many other relevant sources. Chemicals selected to include in an encyclopedic source such as Wikipedia are likely to be relevant and notable and, as a result, we continue to monitor for updates and add the new substance-article mappings on a regular basis. These mappings, in turn, are incorporated into regular updates of the public version of the Dashboard. The ledes of the article can help to add important context to a particular chemical. As an example, 6PPD<sup>102</sup> may be an obscure chemical, but when it is linked to the Wikipedia article, the displayed lede highlights the utility of the chemical as an antiozonant and antioxidant in rubber tires and describes how the related quinone transformation product has resulted in acute mortality in coho salmon.<sup>103</sup>

## ■ DATA AND SOFTWARE AVAILABILITY

All data scraped from Wikipedia at the time of publication for use in this project is available as the Supporting Information ([SI\\_file5\\_Wikipedia\\_Infobox\\_Data](#)). A list of accepted mappings between Wikipedia and DSSTox records, with Wikipedia page title and summary DSSTox identifier information, is also available as the Supporting Information ([SI\\_file4\\_DSSTox\\_WIKIPEDIA\\_List](#)).

All software used for this project was written in-house in Java. Code used to query and process data from Wikipedia has been made publicly available in a GitHub repository (<https://github.com/MrMSDS/wikipedia-infoboxes>). Code to perform mapping against the DSSTox database has not been made publicly available, as it cannot be executed without credential access, and detailed information on the internal structure of the database inherent in such a code may constitute a security risk. The general workflow of this code is described in detail under [Implementation > DSSTox data mapping](#).

## ■ ASSOCIATED CONTENT

### SI Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jcim.2c00886>.

Spreadsheet summaries of identifier availability and correctness in Wikipedia ([XLSX](#))

Tabular summaries of identifier availability and correctness in Wikipedia; summary statistics of drugboxes and chemboxes (Table S1) ([PDF](#))

Investigation of John W. Huffman cannabinoid dataset ([PDF](#))

Summary of Wikipedia pages linked to DSSTox records ([XLSX](#))

Complete identifier data scraped from Wikipedia Chembox and Drugbox pages ([XLSX](#))

## ■ AUTHOR INFORMATION

### Corresponding Author

**Antony J. Williams** – Center for Computational Toxicology and Exposure, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina 27711, United States; [orcid.org/0000-0002-2668-4821](https://orcid.org/0000-0002-2668-4821); Email: [Williams.antony@epa.gov](mailto:Williams.antony@epa.gov)

### Authors

**Gabriel Sinclair** – ORAU Student Services Contractor to Center for Computational Toxicology and Exposure, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina 27711, United States; [orcid.org/0000-0003-0802-2282](https://orcid.org/0000-0003-0802-2282)

**Inthirany Thillainadarajah** – Senior Environmental Employment Program, US Environmental Protection Agency, Research Triangle Park, North Carolina 27711, United States

**Brian Meyer** – Senior Environmental Employment Program, US Environmental Protection Agency, Research Triangle Park, North Carolina 27711, United States

**Vicente Samano** – Senior Environmental Employment Program, US Environmental Protection Agency, Research Triangle Park, North Carolina 27711, United States

**Sakuntala Sivasupramaniam** – Senior Environmental Employment Program, US Environmental Protection Agency, Research Triangle Park, North Carolina 27711, United States

**Linda Adams** – Center for Computational Toxicology and Exposure, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina 27711, United States

**Egon L. Willighagen** – Department of Bioinformatics—BiGCaT, Maastricht University, 6229 ER Maastricht, The Netherlands; [orcid.org/0000-0001-7542-0286](https://orcid.org/0000-0001-7542-0286)

**Ann M. Richard** – Center for Computational Toxicology and Exposure, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina 27711, United States; [orcid.org/0000-0003-2116-2300](https://orcid.org/0000-0003-2116-2300)

**Martin Walker** – Martin Walker, SUNY Potsdam—Chemistry, Potsdam, New York 13676, United States; [orcid.org/0000-0001-9202-0356](https://orcid.org/0000-0001-9202-0356)

Complete contact information is available at: <https://pubs.acs.org/doi/10.1021/acs.jcim.2c00886>

### Notes

The authors declare no competing financial interest. The views expressed in this manuscript are solely those of the authors and do not represent the policies of the U.S. Environmental Protection Agency. Mention of trade names of commercial products should not be interpreted as an

endorsement by the U.S. Environmental Protection Agency. This work has been internally reviewed at the US EPA and has been approved for publication.

## ACKNOWLEDGMENTS

The authors acknowledge the CompTox Chemicals Dashboard software development team for their dedicated efforts to the development of the Dashboard application as this application provides access to our curation work for the community. The authors appreciate our colleagues Katie Paul-Friedman, Louis Groff, and Mark Strynar for their feedback and comments on the manuscript. The information in this document has been funded wholly or in part by the U.S. Environmental Protection Agency.

## REFERENCES

- (1) Wikipedia Page Regarding Alizarine Yellow R, 2022. [https://en.wikipedia.org/w/index.php?title=Alizarine\\_Yellow\\_R&oldid=1006198457](https://en.wikipedia.org/w/index.php?title=Alizarine_Yellow_R&oldid=1006198457).
- (2) Example of a Wikipedia Article Lede Embedded in a ChEBI Page: Aspirin, 2022. <https://www.ebi.ac.uk/chebi/searchId.do?chebiId=CHEBI:15365>.
- (3) ChemSpider Page for Aspirin, 2022. <http://www.chemspider.com/Chemical-Structure.2157.html>.
- (4) PubChem Page Showing a Wikipedia Link: Aspirin, 2022. <https://pubchem.ncbi.nlm.nih.gov/compound/2244#section=Wikipedia>.
- (5) Williams, A. J.; Grulke, C. M.; Edwards, J.; McEachran, A. D.; Mansouri, K.; Baker, N. C.; Patlewicz, G.; Shah, I.; Wambaugh, J. F.; Judson, R. S.; Richard, A. M. The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. *J. Cheminf.* **2017**, *9*, No. 61.
- (6) Lowe, C. N.; Williams, A. J. Enabling High-Throughput Searches for Multiple Chemical Data Using the U.S.-EPA CompTox Chemicals Dashboard. *J. Chem. Inf. Model.* **2021**, *61*, 565–570.
- (7) ChemConnector Blog Page: Dedicating Christmas Time to the Cause of Curating Wikipedia, 2022. <http://www.chemconnector.com/2008/01/09/dedicating-christmas-time-to-the-cause-of-curating-wikipedia/>.
- (8) Pence, H. E.; Williams, A. J. ChemSpider: An Online Chemical Information Resource. *J. Chem. Educ.* **2010**, *87*, 1123–1124.
- (9) ChemConnector Blog: WiChempedia Blog Posts, 2022. <http://www.chemconnector.com/>.
- (10) Wikipedia Project: Chemical Abstracts Service Registry Number (CASRN) validation, 2022. [https://en.wikipedia.org/wiki/Wikipedia\\_talk:WikiProject\\_Chemistry/CAS\\_validation](https://en.wikipedia.org/wiki/Wikipedia_talk:WikiProject_Chemistry/CAS_validation).
- (11) Announcement of Collaboration between Chemical Abstracts Service and Wikipedia, 2022. [https://en.wikipedia.org/wiki/Wikipedia\\_talk:WikiProject\\_Chemistry/CAS\\_validation#New\\_announcement\\_from\\_CAS](https://en.wikipedia.org/wiki/Wikipedia_talk:WikiProject_Chemistry/CAS_validation#New_announcement_from_CAS).
- (12) Walker, M. A. Wikipedia as a Resource for Chemistry. In *Enhancing Learning with Online Resources, Social Networking, and Digital Libraries*; American Chemical Society, 2010; Vol. 1060, pp 79–92.
- (13) CAS Common Chemistry Website, 2022. <https://commonchemistry.cas.org/>.
- (14) Press Release: CAS Common Chemistry Expands Collection of Publicly Available Chemical Information, 2022. <https://www.cas.org/resources/press-releases/common-chemistry>.
- (15) Jacobs, A.; Williams, D.; Hickey, K.; Patrick, N.; Williams, A. J.; Chalk, S.; McEwen, L.; Willighagen, E.; Walker, M.; Bolton, E.; Sinclair, G.; Sanford, A. CAS Common Chemistry in 2021: Expanding Access to Trusted Chemical Information for the Scientific Community. *J. Chem. Inf. Model.* **2022**, *62*, 2737–2743.
- (16) Bizer, C.; Lehmann, J.; Kobilarov, G.; Auer, S.; Becker, C.; Cyganiak, R.; Hellmann, S. DBpedia—A crystallization point for the Web of Data. *J. Web Semant.* **2009**, *7*, 154–165.
- (17) Gamble, M.; Goble, C.; Klyne, G.; Zhao, J. In *MIM: A Minimum Information Model Vocabulary and Framework for Scientific Linked Data*, 2012 IEEE 8th International Conference on E-Science (E-Science); IEEE, 2012.
- (18) Ertl, P.; Patiny, L.; Sander, T.; Rufener, C.; Zasso, M. Wikipedia Chemical Structure Explorer: substructure and similarity searching of molecules from Wikipedia. *J. Cheminf.* **2015**, *7*, No. 10.
- (19) Wikipedia Category: Chemical Substances Page, 2022. [https://en.wikipedia.org/wiki/Category:Chemical\\_substances](https://en.wikipedia.org/wiki/Category:Chemical_substances).
- (20) Wikipedia Category: Chemical Substances by Use, 2022. [https://en.wikipedia.org/wiki/Category:Chemical\\_substances\\_by\\_use](https://en.wikipedia.org/wiki/Category:Chemical_substances_by_use).
- (21) Wikipedia Category: List of CAS Numbers by Chemical Compound, 2022. [https://en.wikipedia.org/wiki/List\\_of\\_CAS\\_numbers\\_by\\_chemical\\_compound](https://en.wikipedia.org/wiki/List_of_CAS_numbers_by_chemical_compound).
- (22) Wikipedia Category: EPA List of Extremely Hazardous Substances, 2022. [https://en.wikipedia.org/wiki/EPA\\_list\\_of\\_extremely\\_hazardous\\_substances](https://en.wikipedia.org/wiki/EPA_list_of_extremely_hazardous_substances).
- (23) Wikipedia: ChemBox Template, 2022. <https://en.wikipedia.org/wiki/Template:Chembox>.
- (24) Wikipedia: DrugBox Template, 2022. [https://en.wikipedia.org/wiki/Template:Infobox\\_drug](https://en.wikipedia.org/wiki/Template:Infobox_drug).
- (25) MediaWiki API help: Embedded-in Query, 2022. <https://www.mediawiki.org/w/api.php?action=help&modules=query%2Bembeddedin>.
- (26) Mediawiki API: Parsing Wikitext, 2022. [https://www.mediawiki.org/wiki/API:Parsing\\_wikitext](https://www.mediawiki.org/wiki/API:Parsing_wikitext).
- (27) jsoup: Java HTML Parser, 2021. <https://jsoup.org/>.
- (28) Wikipedia Chemical Page: Sarafotxin, 2022. <https://en.wikipedia.org/wiki/Sarafotoxin>.
- (29) Williams, A. J.; Lambert, J. C.; Thayer, K.; Dorne, J.-L.C.M. Sourcing data on chemical properties and hazard data from the US-EPA CompTox Chemicals Dashboard: A practical guide for human risk assessment. *Environ. Int.* **2021**, *154*, No. 106566.
- (30) CompTox Chemicals Dashboard: Fluconazole Chemical Page, 2022. <https://comptox.epa.gov/dashboard/DTXSID3020627>.
- (31) CompTox Chemicals Dashboard: Polyvinylpyrrolidone Chemical Page, 2022. <https://comptox.epa.gov/dashboard/DTXSID0025941>.
- (32) CompTox Chemicals Dashboard: Polychlorinated Biphenyls Related Substances page, 2020. <https://comptox.epa.gov/dashboard/dstoxdb/results?search=DTXSID5024267#related-substances>.
- (33) CompTox Chemicals Dashboard: Canakinumab Chemical Page, 2022. <https://comptox.epa.gov/dashboard/chemical/details/DTXSID601017652>.
- (34) CompTox Chemicals Dashboard: Cummingtonite Chemical Page, 2022. <https://comptox.epa.gov/dashboard/chemical/details/DTXSID201015604>.
- (35) Mapping file of InChIStrings, InChIKeys and DTXSIDs for the EPA CompTox Dashboard, 2022. [https://figshare.com/articles/dataset/Mapping\\_file\\_of\\_InChIStrings\\_InChIKeys\\_and\\_DTXSIDs\\_for\\_the\\_EPA\\_CompTox\\_Dashboard/3578313/1](https://figshare.com/articles/dataset/Mapping_file_of_InChIStrings_InChIKeys_and_DTXSIDs_for_the_EPA_CompTox_Dashboard/3578313/1).
- (36) Willighagen, E. L. Bacting: a next generation, command line version of Bioclipse. *J. Open Source Softw.* **2021**, *6*, No. 2558.
- (37) Willighagen, E. L. Github Repository of Groovy Code Used to Add DTXSID Identifiers to Wikidata, 2022. <https://github.com/egonw/ons-wikidata/blob/master/ExtIdentifiers/comptox.groovy>.
- (38) Java Database Connectivity, 2022. [https://en.wikipedia.org/wiki/Java\\_Database\\_Connectivity](https://en.wikipedia.org/wiki/Java_Database_Connectivity).
- (39) epam Indigo Toolkit, 2021. <https://lifescience.opensource.epam.com/indigo/>.
- (40) Lowe, D. M.; Corbett, P. T.; Murray-Rust, P.; Glen, R.C. Chemical name to structure: OPSIN, an open source solution. *J. Chem. Inf. Model.* **2011**, *51*, 739–753.
- (41) OPSIN Application, 2021. <https://opsin.ch.cam.ac.uk/>.
- (42) Grulke, C. M.; Williams, A. J.; Thillanadarajah, I.; Richard, A. M. EPA's DSSTox database: History of development of a curated chemistry resource supporting computational toxicology research. *Comput. Toxicol.* **2019**, *12*, No. 100096.



- (43) Wikipedia: List of Chemical Compounds, 2022. [https://en.wikipedia.org/wiki/Category:Lists\\_of\\_chemical\\_compounds](https://en.wikipedia.org/wiki/Category:Lists_of_chemical_compounds).
- (44) Wikipedia: List of CAS Numbers by Chemical Compound, 2022. [https://en.wikipedia.org/wiki/List\\_of\\_CAS\\_numbers\\_by\\_chemical\\_compound](https://en.wikipedia.org/wiki/List_of_CAS_numbers_by_chemical_compound).
- (45) Wikipedia Page: Testosterone (medication), 2022. [https://en.wikipedia.org/wiki/Testosterone\\_\(medication\)](https://en.wikipedia.org/wiki/Testosterone_(medication)).
- (46) CompTox Chemicals Dashboard: JWH-359 Chemical Page, 2022. <https://comptox.epa.gov/dashboard/chemical/details/DTXSID201028166>.
- (47) CompTox Chemicals Dashboard: Wikipedia chemicals list, 2022. <https://comptox.epa.gov/dashboard/chemical-lists/WIKIPEDIA>.
- (48) Wikipedia: Polycyclic Aromatic Hydrocarbons Chemical Page, 2022. [https://en.wikipedia.org/wiki/Polycyclic\\_aromatic\\_hydrocarbon](https://en.wikipedia.org/wiki/Polycyclic_aromatic_hydrocarbon).
- (49) Wikipedia: Dioxins and Dioxin like Compounds Chemical Page, 2022. [https://en.wikipedia.org/wiki/Dioxins\\_and\\_dioxin-like\\_compounds](https://en.wikipedia.org/wiki/Dioxins_and_dioxin-like_compounds).
- (50) Wikipedia: Polychlorinated Dibenzofurans Chemical Page, 2022. [https://en.wikipedia.org/wiki/Polychlorinated\\_dibenzofurans](https://en.wikipedia.org/wiki/Polychlorinated_dibenzofurans).
- (51) CompTox Chemicals Dashboard: Chlorodibenzofurans Related Substances Page, 2022. <https://comptox.epa.gov/dashboard/chemical/related-substances/DTXSID30891524>.
- (52) CompTox Chemicals Dashboard: Synthetic Cannabinoids List Page, 2022. <https://comptox.epa.gov/dashboard/chemical-lists/PSYCHOCANNAB>.
- (53) Ulrich, E. M.; Sobus, J. R.; Grulke, C. M.; Richard, A. M.; Newton, S. R.; Strynar, M. J.; Mansouri, K.; Williams, A. J. EPA's non-targeted analysis collaborative trial (ENTACT): genesis, design, and initial findings. *Anal. Bioanal. Chem.* **2019**, *411*, 853–866.
- (54) Sobus, J. R.; Wambaugh, J. F.; Isaacs, K. K.; Williams, A. J.; McEachran, A. D.; Richard, A. M.; Grulke, C. M.; Ulrich, E. M.; Rager, J. E.; Strynar, M. J.; Newton, S. R. Integrating tools for non-targeted analysis research and chemical safety evaluations at the US EPA. *J. Exposure Sci. Environ. Epidemiol.* **2018**, *28*, 411–426.
- (55) Sobus, J. R.; Grossman, J. N.; Chao, A.; Singh, R.; Williams, A. J.; Grulke, C. M.; Richard, A. M.; Newton, S. R.; McEachran, A. D.; Ulrich, E. M. Using prepared mixtures of ToxCast chemicals to evaluate non-targeted analysis (NTA) method performance. *Anal. Bioanal. Chem.* **2019**, *411*, 835–851.
- (56) Newton, S. R.; McMahan, R. L.; Sobus, J. R.; Mansouri, K.; Williams, A. J.; McEachran, A. D.; Strynar, M. J. Suspect screening and non-targeted analysis of drinking water using point-of-use filters. *Environ. Pollut.* **2018**, *234*, 297–306.
- (57) Phillips, A. L.; Williams, A. J.; Sobus, J. R.; Ulrich, E. M.; Gundersen, J.; Langlois-Miller, C.; Newton, S. R. A Framework for Utilizing High-Resolution Mass Spectrometry and Non-targeted Analysis in Rapid Response and Emergency Situations. *Environ. Toxicol. Chem.* **2022**, *41*, 1117–1130.
- (58) CompTox Chemicals Dashboard: Synthetic Cannabinoids List Page, 2022. <https://comptox.epa.gov/dashboard/chemical-lists/PSYCHOCANNAB>.
- (59) Wikipedia: List of Schedule I Drugs Page, 2022. [https://en.wikipedia.org/wiki/List\\_of\\_Schedule\\_I\\_drugs\\_\(US\)](https://en.wikipedia.org/wiki/List_of_Schedule_I_drugs_(US)).
- (60) Chao, A.; Al-Ghoul, H.; McEachran, A. D.; Balabin, I.; Transue, T.; Cathey, T.; Grossman, J. N.; Singh, R. R.; Ulrich, E. M.; Williams, A. J.; Sobus, J. R. In silico MS/MS spectra for identifying unknowns: a critical examination using CFM-ID algorithms and ENTACT mixture samples. *Anal. Bioanal. Chem.* **2020**, *412*, 1303–1315.
- (61) McEachran, A. D.; Balabin, I.; Cathey, T.; Transue, T. R.; Al-Ghoul, H.; Grulke, C. M.; Sobus, J. R.; Williams, A. J. Linking in silico MS/MS spectra with chemistry data to improve identification of unknowns. *Sci. Data* **2019**, *6*, No. 11505.
- (62) Little, J. L.; Williams, A. J.; Pshenichnov, A.; Tkachenko, V. Identification of "known unknowns" utilizing accurate mass data and ChemSpider. *J. Am. Soc. Mass Spectrom.* **2012**, *23*, 179–185.
- (63) McEachran, A. D.; Chao, A.; Al-Ghoul, H.; Lowe, C.; Grulke, C. M.; Sobus, J. R.; Williams, A. J. Revisiting Five Years of CASMI Contests with EPA Identification Tools. *Metabolites* **2020**, *10*, No. 260.
- (64) Wikipedia Category: JWH Cannabinoids, 2022. [https://en.wikipedia.org/wiki/Category:JWH\\_cannabinoids](https://en.wikipedia.org/wiki/Category:JWH_cannabinoids).
- (65) Wikipedia: JWH-251 Chemical Page, 2022. <https://en.wikipedia.org/wiki/JWH-251>.
- (66) Wikipedia: JWH-167 Chemical Page, 2022. <https://en.wikipedia.org/wiki/JWH-167>.
- (67) Wikipedia: JWH-210 Chemical Page, 2022. <https://en.wikipedia.org/wiki/JWH-210>.
- (68) Wikipedia: JWH-203 Chemical Page, 2022. <https://en.wikipedia.org/wiki/JWH-203>.
- (69) Wikipedia: JWH-359 Chemical Page, 2022. <https://en.wikipedia.org/wiki/JWH-359>.
- (70) Wikipedia: JWH-176 Chemical Page, 2022. <https://en.wikipedia.org/wiki/JWH-176>.
- (71) Wikipedia: List of Fentanyl Analogues, 2022. [https://en.wikipedia.org/wiki/List\\_of\\_fentanyl\\_analogues](https://en.wikipedia.org/wiki/List_of_fentanyl_analogues).
- (72) Wikipedia: Allotropes of Carbon, 2022. [https://en.wikipedia.org/wiki/Allotropes\\_of\\_carbon](https://en.wikipedia.org/wiki/Allotropes_of_carbon).
- (73) Wikipedia: Category: Allotropes of Carbon, 2022. [https://en.wikipedia.org/wiki/Category:Allotropes\\_of\\_carbon](https://en.wikipedia.org/wiki/Category:Allotropes_of_carbon).
- (74) CompTox Chemicals Dashboard: Polycyclic Aromatic Hydrocarbons Chemical Page, Related Substances, 2022. <https://comptox.epa.gov/dashboard/chemical/related-substances/DTXSID3044043>.
- (75) Wikipedia: Germacrene Chemical Page, 2022. <https://en.wikipedia.org/wiki/Germacrene>.
- (76) Wikipedia: Nonylphenol Chemical Page, 2022. <https://en.wikipedia.org/wiki/Nonylphenol>.
- (77) CompTox Chemicals Dashboard: Nonylphenol Chemical Page, Related Substances, 2022. <https://comptox.epa.gov/dashboard/chemical/related-substances/DTXSID20891518>.
- (78) Wikipedia: Trefentanil Chemical Page, 2022. <https://en.wikipedia.org/wiki/Trefentanil>.
- (79) CompTox Chemicals Dashboard: Search Results for the Trefentanil Substring, 2022. [https://comptox.epa.gov/dashboard/dsstoxdb/multiple\\_results?input\\_type=synonym\\_substring&inputs=Trefentanil](https://comptox.epa.gov/dashboard/dsstoxdb/multiple_results?input_type=synonym_substring&inputs=Trefentanil).
- (80) CompTox Chemicals Dashboard: List of Perfluorooctanesulfonic Acid Related Salts, 2022. [https://comptox.epa.gov/dashboard/search-results?input\\_type=synonym\\_substring&inputs=perfluorooctanesulfonate](https://comptox.epa.gov/dashboard/search-results?input_type=synonym_substring&inputs=perfluorooctanesulfonate).
- (81) Wikipedia: Kraton Polymers Chemical Page, 2022. [https://en.wikipedia.org/wiki/Kraton\\_\(polymer\)](https://en.wikipedia.org/wiki/Kraton_(polymer)).
- (82) Wikipedia: Category: Organic Polymers, 2022. [https://en.wikipedia.org/wiki/Category:Organic\\_polymers](https://en.wikipedia.org/wiki/Category:Organic_polymers).
- (83) Wikipedia: Poly(Methyl methacrylate), 2022. [https://en.wikipedia.org/w/index.php?title=Poly\(methyl\\_methacrylate\)](https://en.wikipedia.org/w/index.php?title=Poly(methyl_methacrylate)).
- (84) Wikipedia: Polydimethylsiloxane Chemical Page, 2022. <https://en.wikipedia.org/wiki/Polydimethylsiloxane>.
- (85) Wikipedia: Polyglycerol Polyricinoleate Chemical Page, 2022. [https://en.wikipedia.org/wiki/Polyglycerol\\_polyricinoleate](https://en.wikipedia.org/wiki/Polyglycerol_polyricinoleate).
- (86) Wikipedia: Polyallylamine Hydrochloride Chemical Page, 2022. [https://en.wikipedia.org/wiki/Polyallylamine\\_hydrochloride](https://en.wikipedia.org/wiki/Polyallylamine_hydrochloride).
- (87) Wikipedia: Polyorthoester Chemical Page, 2022. <https://en.wikipedia.org/wiki/Polyorthoester>.
- (88) Wikipedia: Polydiacetylenes Chemical Page, 2022. <https://en.wikipedia.org/wiki/Polydiacetylenes>.
- (89) Wikipedia: Acrylonitrile-Butadiene-Styrene Chemical Page, 2022. [https://en.wikipedia.org/wiki/Acrylonitrile\\_butadiene\\_styrene](https://en.wikipedia.org/wiki/Acrylonitrile_butadiene_styrene).
- (90) Wikipedia: Adrenaline Chemical Page, 2022. <https://en.wikipedia.org/wiki/Adrenaline>.
- (91) Wikipedia: Epinephrine as a Medication Chemical Page, 2022. [https://en.wikipedia.org/wiki/Epinephrine\\_\(medication\)](https://en.wikipedia.org/wiki/Epinephrine_(medication)).
- (92) Willighagen, E. L.; Slenter, D.; Mietchen, D.; Evelo, C.; Nielsen, F. *Wikidata and Scholia as a Hub Linking Chemical Knowledge*; Maastricht University, 2018, DOI: 10.6084/m9.figshare.6356027.v1.

(93) Nielsen, F.; Mietchen, D.; Willighagen, E. In *Scholia, Scientometrics and Wikidata*, European Semantic Web Conference; Springer: Cham, 2017; pp 237–539.

(94) Waagmeester, A.; Stupp, G.; Burgstaller-Muehlbacher, S.; Good, B. M.; Griffith, M.; Griffith, O. L.; Hanspers, K.; Hermjakob, H.; Hudson, T. S.; Hybiske, K.; Keating, S. M.; Manske, M.; Mayers, M.; Mietchen, D.; Mitraka, E.; Pico, A. R.; Putman, T.; Riutta, A.; Queral-Rosinach, N.; Schriml, L. M.; Shafee, T.; Slenter, D.; Stephan, R.; Thornton, K.; Tsueng, G.; Tu, R.; Ul-Hasan, S.; Willighagen, E.; Wu, C.; Su, A. I. Science Forum: Wikidata as a knowledge graph for the life sciences. *eLife* **2020**, 9, No. e52614.

(95) Richard, A. M.; Judson, R. S.; Houck, K. A.; Grulke, C. M.; Volarath, P.; Thillainadarajah, I.; Yang, C.; Rathman, J.; Martin, M. T.; Wambaugh, J. F.; Knudsen, T. B.; Kancherla, J.; Mansouri, K.; Patlewicz, G.; Williams, A. J.; Little, S. B.; Crofton, K. M.; Thomas, R. S. ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. *Chem. Res. Toxicol.* **2016**, 29, 1225–1251.

(96) Dix, D. J.; Houck, K. A.; Martin, M. T.; Richard, A. M.; Setzer, R. W.; Kavlock, R. J. The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicol. Sci.* **2007**, 95, 5–12.

(97) Wikipedia: Forskolin Chemical Page, 2022. <https://comptox.epa.gov/dashboard/chemical/details/DTXSID8040484>.

(98) Wikipedia: Forskolin Old Chemical Page, 2022. <https://en.wikipedia.org/w/index.php?title=Forskolin&oldid=1055714133>.

(99) Wikipedia: Colforsin Old Chemical Page, 2022. [https://en.wikipedia.org/w/index.php?title=Colforsin\\_daropate&oldid=1057894525](https://en.wikipedia.org/w/index.php?title=Colforsin_daropate&oldid=1057894525).

(100) CommonChemistry: Colforsin Chemical Page, 2022. <https://commonchemistry.cas.org/results?q=Colforsin>.

(101) Wikidata: Revision History of Colforsin, 2022. <https://www.wikidata.org/w/index.php?title=Q5144763&action=history>.

(102) CompTox Chemicals Dashboard: 6PPD Chemical Page, 2022. <https://comptox.epa.gov/dashboard/chemical/details/DTXSID9025114>.

(103) Tian, Z.; Zhao, H.; Peter, K. T.; Gonzalez, M.; Wetzal, J.; Wu, C.; Hu, X.; Prat, J.; Mudrock, E.; Hettinger, R.; Cortina, A. E.; Biswas, R. G.; Kock, F.V.C.; Soong, R.; Jenne, A.; Du, B.; Hou, F.; He, H.; Lundeen, R.; Gilbreath, A.; Sutton, R.; Scholz, N. L.; Davis, J. W.; Dodd, M. C.; Simpson, A.; McIntyre, J. K.; Kolodziej, E. P. A ubiquitous tire rubber-derived chemical induces acute mortality in coho salmon. *Science* **2021**, 371, 185–189.