www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

COMMENT

Navigating *in vitro* bioactivity data by investigating available resources using model compounds

Received: 8 February 2018
Accepted: 6 March 2019
Published online: 29 April 2019

Sten Ilmjärv[1,2], Fiona Augsburger[1], Jerven Tjalling Bolleman[3], Robin Liechti[2], Alan James Bridge[3], Jenny Sandström[4], Vincent Jaquet[1], Ioannis Xenarios[2,5,6] & Karl-Heinz Krause[1]

The number of chemical compounds and associated experimental data in public databases is growing, but presently there is no simple way to access these data in a quick and synoptic manner. Instead, data are fragmented across different resources and interested parties need to invest invaluable time and effort to navigate these systems.

Both the CAS Registry^{SM} and PubChem[1] contain more than 90 million compounds, with new compounds added daily. Most of these compounds are missing toxicological characterization, due in part to the limited capacity of current methods to assess a compound's bioactivity in a living system. High-throughput and scalable *in vitro* test systems aim to bridge that gap. In combination with structural information and known molecular properties, these high-throughput data will allow researchers to describe toxicity pathways more comprehensively. However, the increasing amounts of new data presents its own set of challenges.

Anomalies in metadata records and the inadequate use of ontologies are hindering for the data to be FAIR[2]. Even after a compound has been published in a scientific document, the diversity of compound synonyms and identifiers, and lack of precise metadata and annotations, can lead to false conclusions and difficulties identifying the compound correctly[3]. To improve the reproducibility of experimental results and to test new hypotheses (e.g. development of predictive computational models), availability and accessibility of raw data are crucial. Using a set of four arbitrarily chosen model compounds (*aspirin*, *rosiglitazone*, *valproic acid*, and *tamoxifen*; Table 1), we investigated data access and consistency within publicly available online resources (Table 2). We observed that modest adoption of semantic web technologies and poor annotations of experimental metadata represent a major obstacle for high-quality data integration and reusability. We argue that this could be substantially improved by annotating compound-related experimental data with standardized ontologies. Also, new and existing resources should adapt to accommodate ontology-based data representation on their platforms and compounds should always be accompanied with a unique structural identifier that helps later discoverability and reduces mistakes.

Abbreviations:
CASRN - Chemical Abstract Service Registry Number;
ChEBI - Chemical Entities of Biological Interest;
FAIR - Findable, Accessible, Interoperable and Reusable;
InChI - International Chemical Identifier;
InChIKey - International Chemical Identifier Key;
IUPAC - International Union of Pure and Applied Chemistry;
SPARQL - SPARQL Protocol and RDF Query Language;
SMILES - Simplified Molecular Input Line Entry System;
RESTFul API - Representational State Transfer Application Programming Interface;
RDF - Resource Description Framework.

[1]Department of Pathology and Immunology, Medical School, University of Geneva, Geneva, Switzerland. [2]Vital-IT Group, SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland. [3]Swiss-Prot group, SIB Swiss Institute of Bioinformatics, Medical School, Geneva, Switzerland. [4]SCAHT Swiss Centre for Applied Human Toxicology, Basel, Switzerland. [5]Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland. [6]Departement of Biochemistry and Chemistry, University of Geneva, Geneva, Switzerland. Correspondence and requests for materials should be addressed to S.I. (email: sten.ilmjarv@unige.ch) or K.-H.K. (email: karl-heinz.krause@unige.ch)

Aspirin	**ChEBI ID**	CHEBI:15365
	PubChem CID	CID2244
	InChIKey	BSYNRYMUTXBXSQ-UHFFFAOYSA-N
	InChI	InChI = 1S/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(H,11,12)
	SMILES	CC(=O)Oc1ccccc1C(O)=O
	ChEBI URI	http://purl.obolibrary.org/obo/CHEBI_15365
Rosiglitazone	**ChEBI ID**	CHEBI:50122
	PubChem CID	CID77999
	InChIKey	YASAKCUCGLMORW-UHFFFAOYSA-N
	InChI	InChI = 1S/C18H19N3O3S/c1-21(16-4-2-3-9-19-16)10-11-24-14-7-5-13(6-8-14)12-15-17(22)20-18(23)25-15/h2-9,15H,10-12H2,1H3,(H,20,22,23)
	SMILES	CN(CCOc1ccc(CC2SC=O)NC2=O)cc1)c1ccccn1
	ChEBI URI	http://purl.obolibrary.org/obo/CHEBI_50122
Valproic acid	**ChEBI ID**	CHEBI:39867
	PubChem CID	CID3121
	InChIKey	NIJJYAXOARWZEE-UHFFFAOYSA-N
	InChI	InChI = 1S/C8H16O2/c1-3-5-7(6-4-2)8(9)10/h7H,3-6H2,1-2H3,(H,9,10)
	SMILES	CCCC(CCC)C(O)=O
	ChEBI URI	http://purl.obolibrary.org/obo/CHEBI_39867
Tamoxifen	**ChEBI ID**	CHEBI:41774
	PubChem CID	CID2733526
	InChIKey	NKANXQFJJICGDU-QPLCGJKRSA-N
	InChI	InChI = 1S/C26H29NO/c1-4-25(21-11-7-5-8-12-21)26(22-13-9-6-10-14-22)23-15-17-24(18-16-23)28-20-19-27(2)3/h5-18H,4,19-20H2,1-3H3/b26-25-
	SMILES	CC\C(c1ccccc1)=C(/c1ccccc1)c1ccc(OCCN(C)C)cc1
	ChEBI URI	http://purl.obolibrary.org/obo/CHEBI_41774

Table 1. Table of model compounds used in the study and their identifiers including unified resource identifier (URI).

Identifying Data in Compound-Specific Resources

A chemical compound can be referenced with many identifiers, such as a trade name, a generic name, a systematic IUPAC name, a registry number (e.g. CASRN), or a unique database identifier and its structure-derived representations, i.e. structural identifiers: InChI, InChIKey and SMILES. Any of the above can potentially be used to search for a compound within an online resource, but researchers need to be careful about the variability between resources. For example, the compound *rosiglitazone* has 157 depositor-supplied synonyms in PubChem, but only two synonyms in ChEBI. Predictably, the PubChem depositor-supplied synonym for *rosiglitazone* termed *Gaudil* failed to recognize the compound in ChEBI.

Structural identifiers, intuitively, should be the most unique identifiers of a compound, but disparity between the resources still exists. Among eleven resources that reported SMILES (BindingDB, ChEBI, ChEMBL, ChemIDPlus, ChemSpider, CompTox, CTD, DrugBank, HMDB, HSDB, PubChem, T3DB and ZINC15), we found 8 different SMILES for *rosiglitazone* and *tamoxifen*, 5 for *aspirin* and 3 for *valproic acid*. A single InChIKeys was observed for *aspirin*, *valproic acid* and *tamoxifen* but three different ones for *rosiglitazone*. IUPAC systematic names were only reported in ChEBI, ChemSpider, CompTox, DrugBank, HMDB, PubChem and T3DB and demonstrated the largest variability: 3 different names for *aspirin*, 4 for *rosiglitazone*, 1 for *valproic acid* and 5 for *tamoxifen*. UniChem[4] provides a cross-referencing service connecting 39 individual database identifiers of various resources using InChIKeys but this service is only useful when one already knows the compound's database identifier or the InChIKey. Currently, it cannot be used with other structural identifiers or compound names.

InChIKey was the most unique identifier among the various databases, possibly because InChI is derived from a single algorithm, whereas several proprietary and open-source algorithms exist for SMILES, whose implementations differ from one another[5]. Although widely used, we did not look at CASRN because the accuracy of CASRN in the public domain is not absolute and reliable information can only be accessed by paid services provided by the Chemical Abstract Service (CAS)[6].

Identification of Compound Data in Omics Databases

The identity of chemical compounds reported in omics experiments can be ambiguous since compounds are often mentioned by name without the accompanying structure representations[3]. We investigated this issue by searching a series of omics data resources using structural identifiers of the compounds in Table 1 as reported in ChEBI, using web-based free-text searches (ArrayExpress, ExpressionAtlas, BioSamples, GEO and PRIDE). We were able to retrieve data for all model compounds from at least four resources using compound names.

Database	Number of compounds	Last checked	Data type	Link
ArrayExpress[12]	—	—	Raw	ebi.ac.uk/arrayexpress
BindingDB[13]	717,572	04-2019	Curated	bindingdb.org
BioSamples[14]	—	—	Raw	ebi.ac.uk/biosamples
ChEBI[15]	55,660	04-2019	Curated	ebi.ac.uk/chebi
ChEMBL[16]	1,879,206	04-2019	Curated	ebi.ac.uk/chembl
ChemIDPlus[17]	421,602	04-2019	Curated	chem.nlm.nih.gov/chemidplus
ChemSpider[18]	~71 million	04-2019	Curated	chemspider.com
CompTox[19]	~870,000	04-2019	Raw/Curated	comptox.epa.gov/dashboard
CTD[20]	15,913	04-2019	Curated	ctdbase.org
DrugBank[21]	11,926	04-2019	Curated	drugbank.ca
ExpressionAtlas[22]	—	—	Raw/Curated	ebi.ac.uk/gxa/home
GEO[23]	—	—	Raw	ncbi.nlm.nih.gov/geo/
HMDB[24]	114,100	04-2019	Curated	hmdb.ca
HSDB[25]	6,016	04-2019	Curated	toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?HSDB
PRIDE[26]	—	—	Raw	ebi.ac.uk/pride/archive/
PubChem[1]	>97,400,000	04-2019	Raw/Curated	pubchem.ncbi.nlm.nih.gov
T3DB[27]	3,678	04-2019	Curated	t3db.ca
UniProt[28]	—	—	Curated	uniprot.org
ZINC[29]	>100,000,000	04-2019	Curated	zinc15.docking.org

Table 2. A list of resources used in the study, their categorization and the number of estimated compounds in these resources at the time of the study.

In addition, the IUPAC systematic names of *aspirin*, *rosiglitazone* and *valproic acid* retrieved datasets from ArrayExpress, BioSamples and GEO. Interestingly, in BioSamples, we were able to retrieve datasets *for valproic acid* also with SMILES. These datasets, however, actually corresponded to the sodium salt of *valproic acid*, which has a slightly different SMILES representation in ChEBI compared to *valproic acid*. Confusingly, these samples were not retrieved when the compound name was used instead.

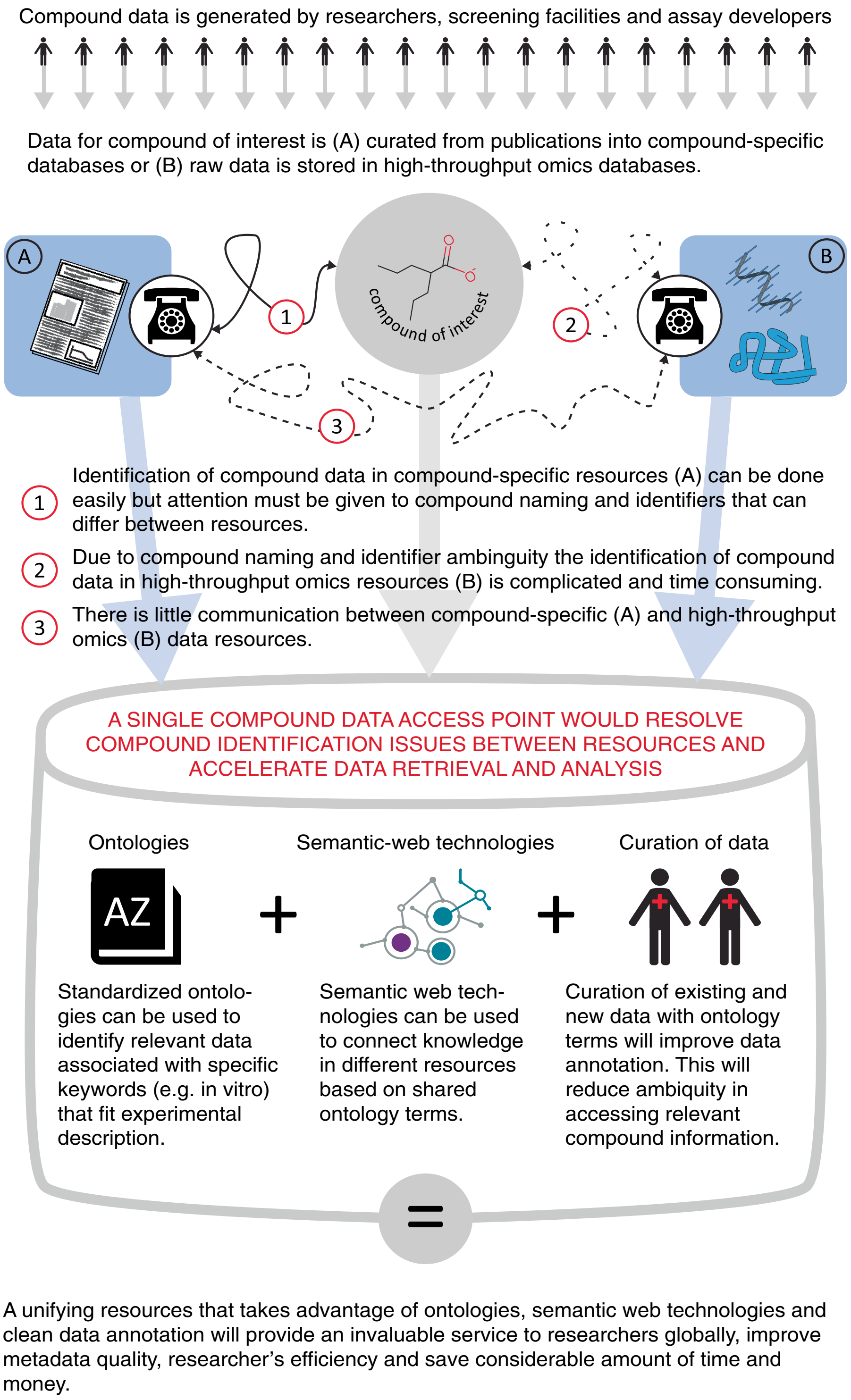
This highlights that, at present, the best way to identify compound-related data from omics resources is with compound names, which requires researchers to exhaust all compound synonyms. To understand this variability between annotations in sample labels, we retrieved the name, synonyms and structural identifiers for each of our model compounds from the ChEMBL public SPARQL endpoint. These were used to identify samples and labels in the BioSamples database through its public SPARQL endpoint. For *rosiglitazone* and *tamoxifen*, only the samples with the respective name was found in any of the sample labels. For aspirin, samples were found using *aspirin*, *asparin*, *asprin*, *levius* and *measurin*. Surprisingly, the compound name *acetylsalicylic acid* was not found in any of the sample labels. *Valproic acid* retrieved results also for *valproate*, *depakote* and *44089*. The latter is a synonym of *valproic acid* in ChEMBL but none of the associated samples were actually associated to *valproic acid*. Of note, all the samples retrieved were unique, i.e. alternative compound labels were not used to annotate the same sample.

Identification of *in vitro* Compound Data

One approach to identify *in vitro* data in public resources is to browse the study descriptions for references of *in vitro* experiment related keywords like "*in vitro*", "cell-line" or specific cell-line names (e.g. "HeLa"). ChEMBL provides a web-based search, which allows one to retrieve data on compounds associated with specific cell-lines or *in vitro* assays. Because this approach is not scalable, most public resources also provide access through bulk data downloads, or programmatically through RESTful API or RDF technologies.

In a RESTful query, the data request is constructed into a single URL which is simple to use and platform independent. Out of the 19 resources in our study, 10 provided free access to their RESTful API. The DrugBank API can be accessed for a fee. Data in RDF compatible formats can be supplied as a bulk download, or through public SPARQL endpoints, which facilitate querying the service provider directly, thus always retrieving the most up-to-date data. In our study, only BioSamples, ChEMBL, ExpressionAtlas and UniProt provided a public SPARQL endpoint. Acquiring data using a SPARQL endpoint can be slower compared to RESTful data access, since the latter is better optimized for specific, recurrent query requests. In contrast, SPARQL queries have the benefit of being customizable, providing flexibility that caters to the researchers' unique needs. Also, since RDF is an inherent part of the "linked data" concept, it can be used to find relationships between datasets in different resources. This is useful for data integration purposes, such as connecting a compound's effect in one resource to its physicochemical properties in another.

Ontology terms can be used to directly associate and retrieve samples with keywords related to *in vitro* experiments. Using BioSamples' public SPARQL endpoint as our target database, we found samples for all our model compounds using ChEBI universal reference identifiers (URI) (Table 1). We were also able to find data for our sample compounds retrieved with ChEBI ontology terms, that had been annotated with the molarity unit term (http://purl.obolibrary.org/obo/UO_0000061, Units of measurement ontology, UO[7]) and the cell-line ontology term (http://www.ebi.ac.uk/efo/EFO_0000322, Experimental Factor Ontology, EFO[8]), both indicators of *in vitro*



Compound data is generated by researchers, screening facilities and assay developers

Data for compound of interest is (A) curated from publications into compound-specific databases or (B) raw data is stored in high-throughput omics databases.

1 Identification of compound data in compound-specific resources (A) can be done easily but attention must be given to compound naming and identifiers that can differ between resources.

2 Due to compound naming and identifier ambinguity the identification of compound data in high-throughput omics resources (B) is complicated and time consuming.

3 There is little communication between compound-specific (A) and high-throughput omics (B) data resources.

A SINGLE COMPOUND DATA ACCESS POINT WOULD RESOLVE COMPOUND IDENTIFICATION ISSUES BETWEEN RESOURCES AND ACCELERATE DATA RETRIEVAL AND ANALYSIS

Standardized ontologies can be used to identify relevant data associated with specific keywords (e.g. in vitro) that fit experimental description.

Semantic web technologies can be used to connect knowledge in different resources based on shared ontology terms.

Curation of existing and new data with ontology terms will improve data annotation. This will reduce ambiquity in accessing relevant compound information.

A unifying resources that takes advantage of ontologies, semantic web technologies and clean data annotation will provide an invaluable service to researchers globally, improve metadata quality, researcher's efficiency and save considerable amount of time and money.

Fig. 1 A graphics illustrating the problems of integrating knowledge between compound of interest and different types of data resources. The problems can be solved with integrated approaches using ontologies, semantic-web technologies and better annotation of the data.

assays. Using the latter, we were able to identify several examples of *rosiglitazone* and *tamoxifen* samples and a single example for *valproic acid*. With the exception of these few examples, we observed that most data for our compounds had been deposited without associated ontology terms. Nevertheless, we are confident that further uptake of the ontologies and improved annotations will be a powerful feature in future search strategies leading to increased data integration capabilities.

Final Thoughts

There already exists a substantial corpus of resources that contain data on a large number of chemical compounds. These data and their sources are diverse and they need to be integrated in order to attain a complete understanding on a compound (Fig. 1). Accessing published data with correct compound information is essential. The problems encountered in accessing data on our model compounds, demonstrate, that using the results from publications stored in public resources and cross-referencing them with omics data still requires substantial investigative capacity. Efforts similar to SourceData[9], that allows to annotate already published figures in existing publication, and RepositiveIO (https://repositive.io/), that makes improving metadata a crowd-sourced task, could provide a potential remedy. Would the efforts necessary for general accession to *in vitro* compound data be worth the money and time? Considering the success of UniProt which incorporates extensively curated and trustworthy protein data, the answer is yes. Indeed an analysis published in the EMBL-EBI value report[10] estimated 46% increase in research efficiency for scientists accessing information relevant to their research question. With around 400,000 unique visitors per month, the reported estimation had an enormous cost-effect benefit for the researcher community. The interest in chemical compounds is even bigger: PubChem alone receives about 1 million unique users per month[11]. This highlights the need for an improved resource that would enhance the efficiency and speed of accessing raw and analyzed compound data in a reliable, simplified and intuitive manner. It would allow researchers to focus on data analysis and its interpretation instead of collection and curation.

References

1. Kim, S. *et al*. PubChem Substance and Compound databases. *Nucleic Acids Res* **44**, D1202–1213, https://doi.org/10.1093/nar/gkv951 (2016).
2. Goncalves, R. S. & Musen, M. A. The variable quality of metadata about biological samples used in biomedical experiments. *Sci Data* **6**, 190021, https://doi.org/10.1038/sdata.2019.21 (2019).
3. Murray-Rust, P., Mitchell, J. B. & Rzepa, H. S. Communication and re-use of chemical information in bioscience. *BMC Bioinformatics* **6**, 180, https://doi.org/10.1186/1471-2105-6-180 (2005).
4. Chambers, J. *et al*. UniChem: a unified chemical structure cross-referencing and identifier tracking system. *J Cheminform* **5**, 3, https://doi.org/10.1186/1758-2946-5-3 (2013).
5. Heller, S. R., McNaught, A., Pletnev, I., Stein, S. & Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J Cheminform* **7**, 23, https://doi.org/10.1186/s13321-015-0068-4 (2015).
6. Williams, A. J. *et al*. Open PHACTS: semantic interoperability for drug discovery. *Drug Discov Today* **17**, 1188–1198, https://doi.org/10.1016/j.drudis.2012.05.016 (2012).
7. Gkoutos, G. V., Schofield, P. N. & Hoehndorf, R. The Units Ontology: a tool for integrating units of measurement in science. *Database (Oxford)* **2012**, bas033, https://doi.org/10.1093/database/bas033 (2012).
8. Malone, J. *et al*. Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* **26**, 1112–1118, https://doi.org/10.1093/bioinformatics/btq099 (2010).
9. Liechti, R. *et al*. SourceData: a semantic platform for curating and searching figures. *Nat Methods* **14**, 1021–1022, https://doi.org/10.1038/nmeth.4471 (2017).
10. Beagrie, N. H. J. The Value and Impact of the European Bioinformatics Institute. Full Report, Charles Beagrie Ltd, https://beagrie.com/static/resource/EBI-impact-report.pdf (2016).
11. Kim, S., Thiessen, P. A., Bolton, E. E. & Bryant, S. H. PUG-SOAP and PUG-REST: web services for programmatic access to chemical information in PubChem. *Nucleic Acids Res* **43**, W605–611, https://doi.org/10.1093/nar/gkv396 (2015).
12. Kolesnikov, N. *et al*. ArrayExpress update–simplifying data submissions. *Nucleic Acids Res* **43**, D1113–1116, https://doi.org/10.1093/nar/gku1057 (2015).
13. Gilson, M. K. *et al*. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res* **44**, D1045–1053, https://doi.org/10.1093/nar/gkv1072 (2016).
14. Faulconbridge, A. *et al*. Updates to BioSamples database at European Bioinformatics Institute. *Nucleic Acids Res* **42**, D50–52, https://doi.org/10.1093/nar/gkt1081 (2014).
15. Degtyarenko, K. *et al*. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res* **36**, D344–350, https://doi.org/10.1093/nar/gkm791 (2008).
16. Bento, A. P. *et al*. The ChEMBL bioactivity database: an update. *Nucleic Acids Res* **42**, D1083–1090, https://doi.org/10.1093/nar/gkt1031 (2014).
17. Tomasulo, P. ChemIDplus-super source for chemical and drug information. *Med Ref Serv Q* **21**, 53–59, https://doi.org/10.1300/J115v21n01_04 (2002).
18. Pence, H. E. & Williams, A. ChemSpider: An Online Chemical Information Resource. *Journal of Chemical Education* **87**, 1123–1124, https://doi.org/10.1021/ed100697w (2010).
19. Williams, A. J. *et al*. The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. *J Cheminform* **9**, 61, https://doi.org/10.1186/s13321-017-0247-6 (2017).
20. Davis, A. P. *et al*. The Comparative Toxicogenomics Database: update 2017. *Nucleic Acids Res* **45**, D972–D978, https://doi.org/10.1093/nar/gkw838 (2017).
21. Law, V. *et al*. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res* **42**, D1091–1097, https://doi.org/10.1093/nar/gkt1068 (2014).
22. Petryszak, R. *et al*. Expression Atlas update–an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Res* **44**, D746–752, https://doi.org/10.1093/nar/gkv1045 (2016).
23. Barrett, T. *et al*. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res* **41**, D991–995, https://doi.org/10.1093/nar/gks1193 (2013).
24. Wishart, D. S. *et al*. HMDB 3.0–The Human Metabolome Database in 2013. *Nucleic Acids Res* **41**, D801–807, https://doi.org/10.1093/nar/gks1065 (2013).
25. Fonger, G. C., Hakkinen, P., Jordan, S. & Publicker, S. The National Library of Medicine's (NLM) Hazardous Substances Data Bank (HSDB): background, recent enhancements and future plans. *Toxicology* **325**, 209–216, https://doi.org/10.1016/j.tox.2014.09.003 (2014).
26. Vizcaino, J. A. *et al*. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res* **44**, D447–456, https://doi.org/10.1093/nar/gkv1145 (2016).
27. Wishart, D. *et al*. T3DB: the toxic exposome database. *Nucleic Acids Res* **43**, D928–934, https://doi.org/10.1093/nar/gku1004 (2015).
28. The UniProt, C. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **45**, D158–D169, https://doi.org/10.1093/nar/gkw1099 (2017).
29. Sterling, T. & Irwin, J. J. ZINC 15–Ligand Discovery for Everyone. *J Chem Inf Model* **55**, 2324–2337, https://doi.org/10.1021/acs.jcim.5b00559 (2015).

Acknowledgements

S.I. was supported by the Swiss Centre of Applied Human Toxicology (SCAHT).

Additional Information

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.