PubChemLite plus Collision Cross Section (CCS) values for enhanced interpretation of non-target environmental data

Anjana Elapavalore[a], Dylan H. Ross[b,c], Valentin Grouès[a], Dagny Aurich[a], Allison M. Krinsky[b], Sunghwan Kim[d], Paul A. Thiessen[d], Jian Zhang[d], James N. Dodds[e], Erin S. Baker[e], Evan E. Bolton[d°], Libin Xu[b°], Emma L. Schymanski[a°]

[a] Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 6 Avenue du Swing, 4367, Belvaux, Luxembourg. ORCIDs AE: 0000-0002-0295-6618; VG: 0000-0001-6501-0806 DA: 0000-0001-8823-0596; ELS: 0000-0001-6868-8145.

[b] Department of Medicinal Chemistry, University of Washington, Seattle, Washington 98195, United States. ORCIDs DHR: 0009-0005-2943-2282; AMK: NA; LX: 0000-0003-1021-5200.

[c] Current Address: Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, USA.

[d] National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), Bethesda, MD, 20894, USA. ORCIDs SK: 0000-0001-9828-2074; PAT: 0000-0002-1992-2086; JZ: 0000-0002-6192-4632; EEB: 0000-0002-5959-6190.

[e] Department of Chemistry, University of North Carolina, Chapel Hill, North Carolina 27599, USA. ORCIDs JND: 0000-0002-9702-2294 ESB: 0000-0001-5246-2213

°Contact: EEB: bolton@ncbi.nlm.nih.gov, LX: libinxu@uw.edu and ELS: emma.schymanski@uni.lu

Abstract

Finding relevant chemicals in the vast (known) chemical space is a major challenge for environmental and exposomics studies leveraging non-target high resolution mass spectrometry (NT-HRMS) methods. Chemical databases now contain hundreds of millions of chemicals, yet many are not relevant. This article details an extensive collaborative, open science effort to provide a dynamic collection of chemicals for environmental, metabolomics and exposomics research, along with supporting information about their relevance to assist researchers in the interpretation of candidate hits. The PubChemLite for Exposomics collection is compiled from ten annotation categories within PubChem, enhanced with patent, literature and annotation counts, predicted partition coefficient (logP) values, as well as predicted collision cross section (CCS) values using CCSbase. Monthly versions are archived on Zenodo under a CC-BY license, supporting reproducible research, and a new interface has been developed, including the chemical stripes on patent and literature data, for researchers to browse the collection. This article further describes how PubChemLite can support researchers in environmental/exposomics studies, describes efforts to increase the availability of experimental CCS values, and explores known limitations and potential for future developments. The data and code behind these efforts are openly available. PubChemLite content can be explored at https://pubchemlite.lcsb.uni.lu.

Keywords: non-target screening; identification; PubChemLite; exposomics; ion mobility; collision cross section; PubChem

Synopsis: PubChemLite empowers environmental non-target screening data interpretation by combining annotation content, patent, literature, logP and predicted collision cross section data.

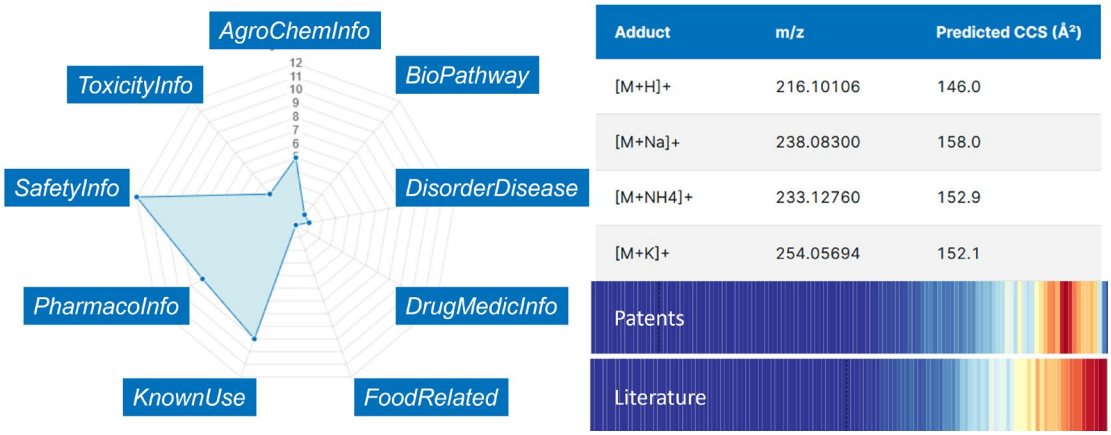

Table of Contents Graphic

Introduction

Environmental and exposomics researchers are faced with the daunting task of determining which chemicals, among other factors, may be either potentially detrimental or beneficial in the context of human and/or environmental health. Non-target screening (NTS) methods leveraging high resolution mass spectrometry (HRMS) approaches are now commonly used to explore complex samples due to high sensitivity and selectivity plus improved availability of HRMS instruments[1,2]. Ion mobility spectrometry (IMS), which separates molecules based on their size and shape, is also increasingly accessible. The calculated collision cross section (CCS) values serve as an additional parameter to support identification in NTS[3,4,5]. Nonetheless, the identification and - importantly - interpretation of features detected during NTS is still challenging, while integration of IMS/CCS into workflows remains poor, hindering the broader adoption of NTS[1]. Identification in NTS primarily relies on mass spectral libraries, suspect lists (lists of hundreds or thousands of chemicals that may occur in the samples) and chemical databases, as recently reviewed elsewhere[1,2].

A diverse array of compound databases, often openly accessible, serve as primary sources of candidates for identification in NTS. These range from hundreds of thousands to just over a million entries (*e.g.,* HMDB[6] with 220,945 metabolites and CompTox[7] with 1,218,248 chemicals as of 20 Nov. 2024), through to hundreds of millions of entries (*e.g.,* ChemSpider[8], PubChem[9] and the Chemical Abstract Services (CAS) Registry[10], with 129, 119 and 219 million chemicals as of 20 Nov. 2024, respectively). Since the CAS Registry is licensed and ChemSpider introduced limitations to their application programming interface (API) in 2018, PubChem has become the *de facto* standard large chemical database for open science-based NTS methods. While PubChem, with >1000 sources, integrates the contents of many of the smaller openly available databases, PubChem also includes tens of millions of entries that are neither likely to be found in the environment, nor pertinent to the exposome. This hinders both the performance and efficiency of NTS. Additionally, many of the chemicals in other potential sources for NTS identification efforts, such as the Global Chemical Inventory (350,000 chemicals)[11] and various lists

from European regulators contributing to the NORMAN Suspect List Exchange (NORMAN-SLE)[12] include large proportions of chemicals that have very little supporting evidence about their existence and relevance, which makes interpretation of potential hits in NTS very challenging.

To mitigate these challenges, a subset of PubChem called PubChemLite was developed specifically to streamline NTS identification and interpretation[13]. PubChemLite has been integrated into existing HR-MS workflows, such as patRoon[14] and MetFrag[15]. Although PubChemLite is familiar to many researchers already, the original 2021 article[13] was primarily technical. This article explains PubChemLite for an environmental/exposomics audience and details extensions since the original publication, including the development of an open experimental CCS pipeline in PubChem, integration of predicted CCS values in PubChemLite to support IMS, and finally a new web interface (https://pubchemlite.lcsb.uni.lu/).

Materials and Methods

Building PubChemLite

The full technical details of PubChemLite are published elsewhere[13]. Briefly, PubChemLite is derived from major categories relevant to environmental/exposomics applications appearing in the PubChem Table of Contents (TOC) pages of the PubChem database[16]. The ten categories currently used to compile PubChemLite (see Figure 1) are: Agrochemical Information (AgroChemInfo), Associated Disorders and Diseases (DisorderDisease), Drug and Medication Information (DrugMedicInfo), Food Additives and Ingredients (FoodRelated), Identification (Identification), Interactions and Pathways – Pathways subset (BioPathway), Pharmacology and Biochemistry (PharmacoInfo), Safety and Hazards (SafetyInfo), Toxicity (ToxicityInfo), Use and Manufacturing (KnownUse). These categories have remained consistent since the original publication, except for the "Biomolecular Interactions and Pathways" category, which was renamed by PubChem to "Interactions and Pathways" in 2022, then limited to the Pathways subset in May 2023 (see Results and Discussion). The input files and code for the PubChemLite build system are available on the Environmental Cheminformatics (ECI) GitLab repository (see Data Availability Statement).

Any compound with one or more of the selected annotation categories is included. The matching compounds (represented by PubChem Compound IDentifiers, CIDs) are aggregated by the first block of the InChIKey into a primary entry and related CIDs, where the primary entry is the neutral or "parent" form. Entries such as mixtures, disconnected substances, or those causing errors, such as some transition metals, are excluded (see the build system code for details[13]). Chemical information (SMILES, InChI, InChIKey, formula, mass), patent and literature (PubMed) counts plus predicted XlogP values are retrieved in bulk using PubChem APIs. The chemical identifiers, mass and XlogP values correspond to the "parent" (primary entry), while the annotation, patent and literature counts are aggregated across all related CIDs. Importantly, the presence of an entry in PubChemLite means that at least one of these annotation categories is available for each CID, with the resulting information publicly available on PubChem to help interpret the relevance of the candidate, see Figure 1. Since the chemical content of PubChem changes daily and annotation content weekly, PubChemLite is built and evaluated early Friday mornings, following the weekly PubChem update cycle. New versions are currently released publicly (typically last Friday of the month) on Zenodo (DOI: 10.5281/zenodo.5995885 redirects to the latest version).

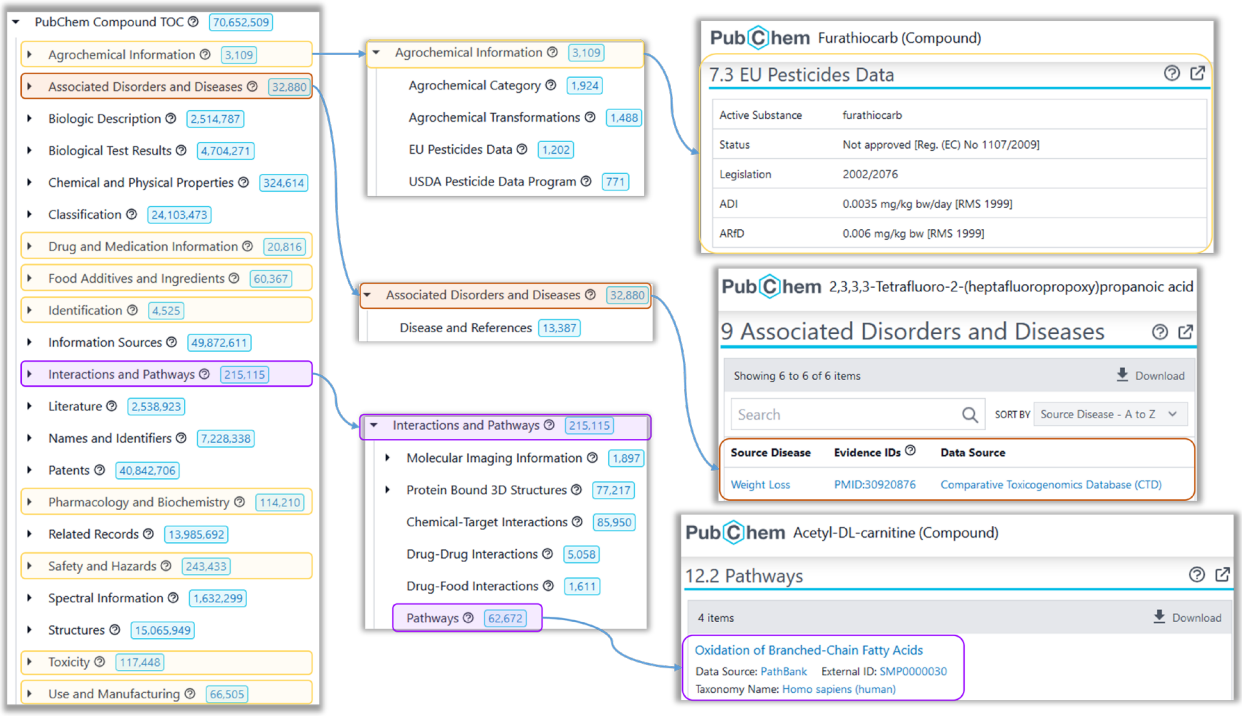

Figure 1: PubChemLite categories in the PubChem Table of Contents (TOC), selected subcategories and associated annotation examples. Yellow shading denotes "environmental" categories (example CID 47759), red the "exposomics" (example CID 114481) and purple the "metabolomics" sections (example CID 1). For high resolution live images, please click the embedded hyperlinks.

Adding CCS Values

Although several methods to predict CCS values are available, few of these are suitable for a large database like PubChem, or even PubChemLite. The calculation time of quantum methods (*e.g.,* MobCal[17] and ISiCLE[18]) is prohibitive. Furthermore, since ISiCLE only predicts values for C, H, N, O, P and S-containing compounds, ~40% of PubChemLite would remain uncovered. Machine learning (ML) prediction methods developed on experimental CCS datasets are much faster, with options including AllCCS[19], CCSbase[20], SigmaCCS[21], DeepCCS[22]. and CCS Predictor 2.0[23]. Since initial calculations with CCSbase were promising, this became the method of choice to establish pipelines for PubChemLite. Calculations are performed using cs3db (the code base behind CCSbase[20]), which has been re-trained with the updated experimental database in CCSbase using the same methodology[20], modified to run on PubChem internal systems. Following the monthly PubChemLite release, CCS value calculations (by PubChem) are triggered on the second day of the following month. The resulting file is transferred by FTP, merged into the PubChemLite files, quality-controlled, then released to Zenodo (DOI: 10.5281/zenodo.4081056 redirects to the latest version).

To ensure that ML models have better coverage of environmentally relevant compounds to improve their predictions, part of this work involved establishing a pipeline to integrate experimental CCS values into PubChem. Currently PubChem contains CCS values from the Baker Lab[24,25], CCSbase[20] and four collections via the NORMAN-SLE[12]: S50 CCSCOMPEND[26,27], S61 UJICCSLIB[28,29], S79 UACCSCEC[3,30] and S116 REFCCS[31]. These values are displayed on individual records in PubChem and navigable in the PubChem Classification Browser via the CCSbase, Baker Lab, NORMAN-SLE and the Aggregated CCS trees (see Figure 2).

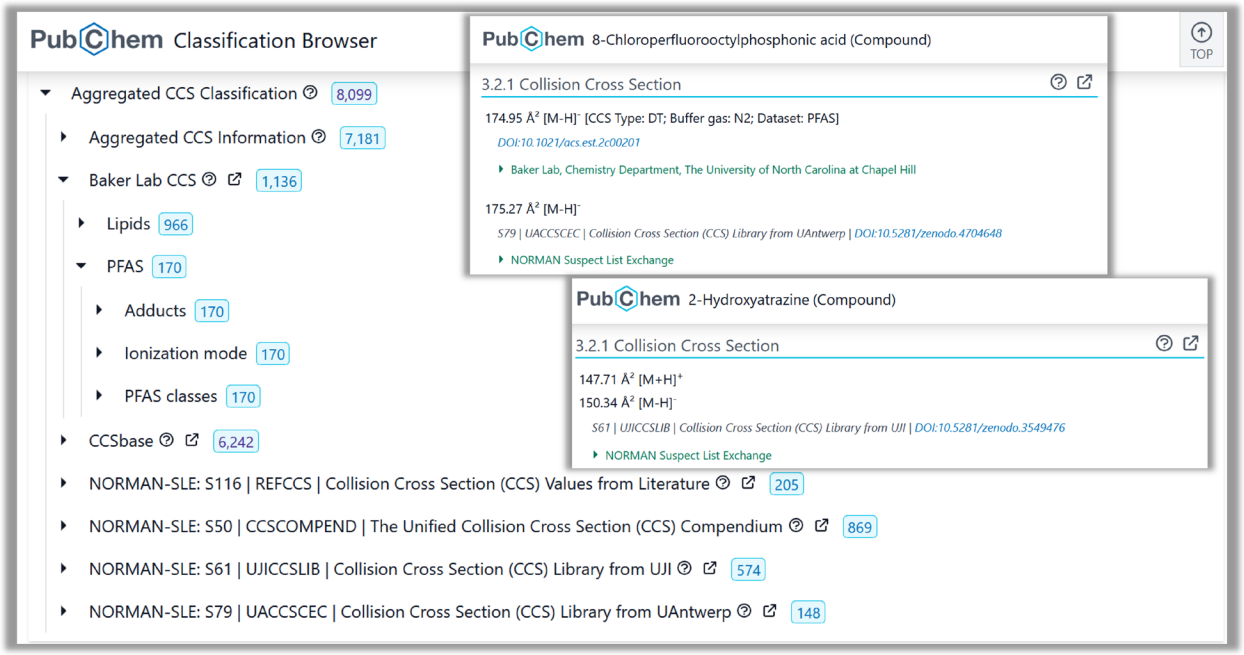

Figure 2: Aggregated Collision Cross Section (CCS) Classification Tree in PubChem. Inset: Experimental CCS values in individual PubChem compound records for Cl-PFOPA (CID 138395139) and the transformation product 2-hydroxyatrazine (CID 135398733). For high resolution live images, please click the embedded hyperlinks.

This data can be retrieved from PubChem (code available on the ECI GitLab). The resulting compiled dataset is available on Zenodo (DOI: 10.5281/zenodo.6800138 redirects to the latest version).

PubChemLite Web Interface

The PubChemLite web interface is developed as a plugin for the ELIXIR-Luxembourg Data Catalog[32,33]. It is developed in Python, CSS, HTML and Javascript, using RDKit[34,35] for structure depiction. For full details, see the PubChemLite-web code on GitLab.

The PubChemLite interface is based almost entirely on the information available in the archived PubChemLite-CCSbase CSV files (see Figure 3), except that additional synonyms (excluded from PubChemLite files for efficiency) are retrieved from the PubChem FTP site for better searchability. Additionally, records are enhanced with visualizations of the annotation categories and tables of the CCS and associated adduct mass values. Finally, the chemical stripes[36,37,38] are included where available for both literature and patents (see Figure 4). The original chemical stripes R version was rewritten in Python for integration in PubChemLite-web, with code available in both repositories[38,39].

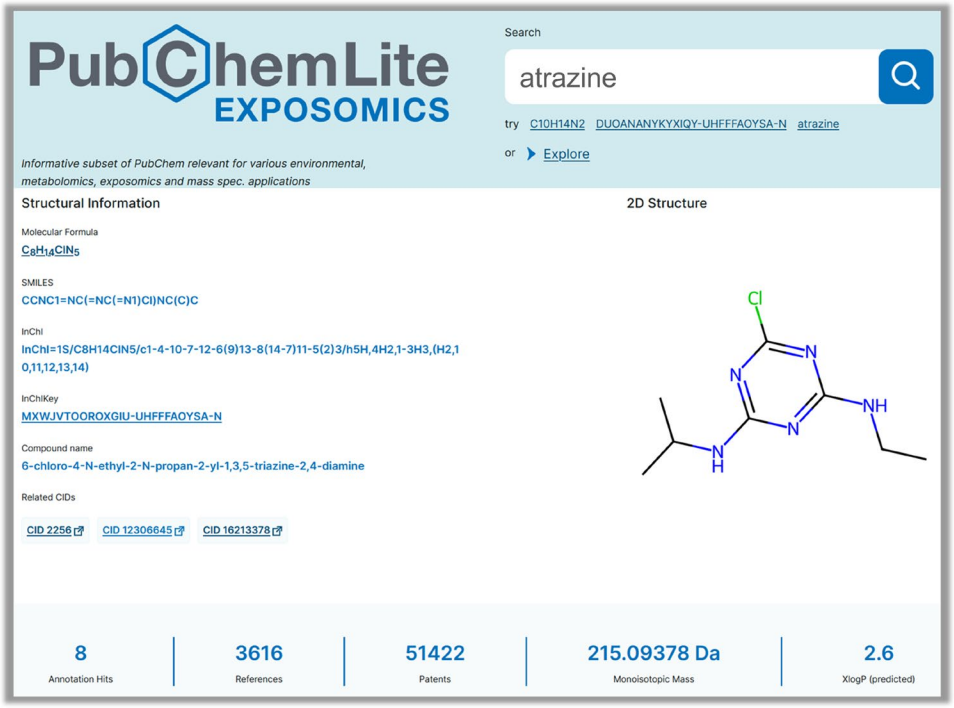

Figure 3: PubChemLite web interface (composite image), compound view of Atrazine. For high resolution live images, please click the embedded hyperlinks.

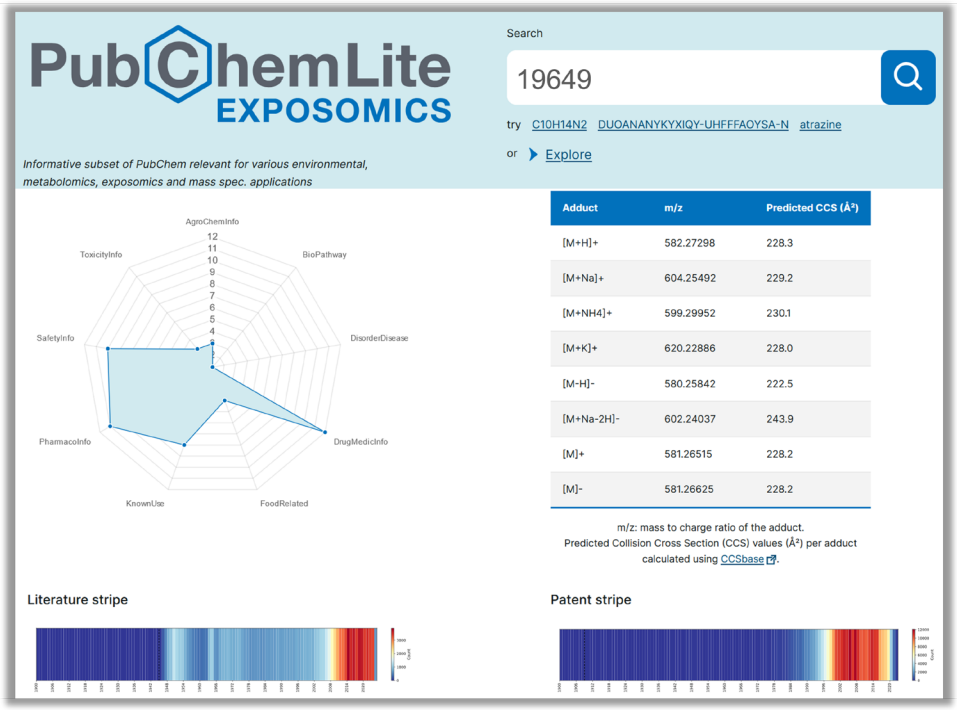

Figure 4: PubChemLite web interface (composite image), view of additional data including annotations, CCS values, patent and literature stripes for Streptomycin. For high resolution live images, please click the embedded hyperlinks.

Results and Discussion

PubChemLite Over Time

The performance of PubChemLite is monitored weekly with every build using the evaluation dataset of 977 compounds established in the original publication[13]. The ranking performance has been quite stable over the three-year period, with median rank=1 of 794 (81.3%), 1-2 of 917 (93.9%), 1-5 of 960 (98.3%) and 12 (1.2%) failures (compounds absent from PubChemLite due to lack of corresponding annotation). The respective ranges [min;max] are rank=1 [788;800], 1-2 [909;922], 1-5 [955;963] and [10;15] failures. The current performance is close to the median values: rank=1 of 797, 1-2 of 916, 1-5 of 960 and 11 failures and slightly better than the original publication (794, 912, 954, respectively, and 15 failures; October 2020 version). The distribution of annotation content included in PubChemLite, including the total number of entries between Feb. 2022 and Nov. 2024, is shown in Figure 5.

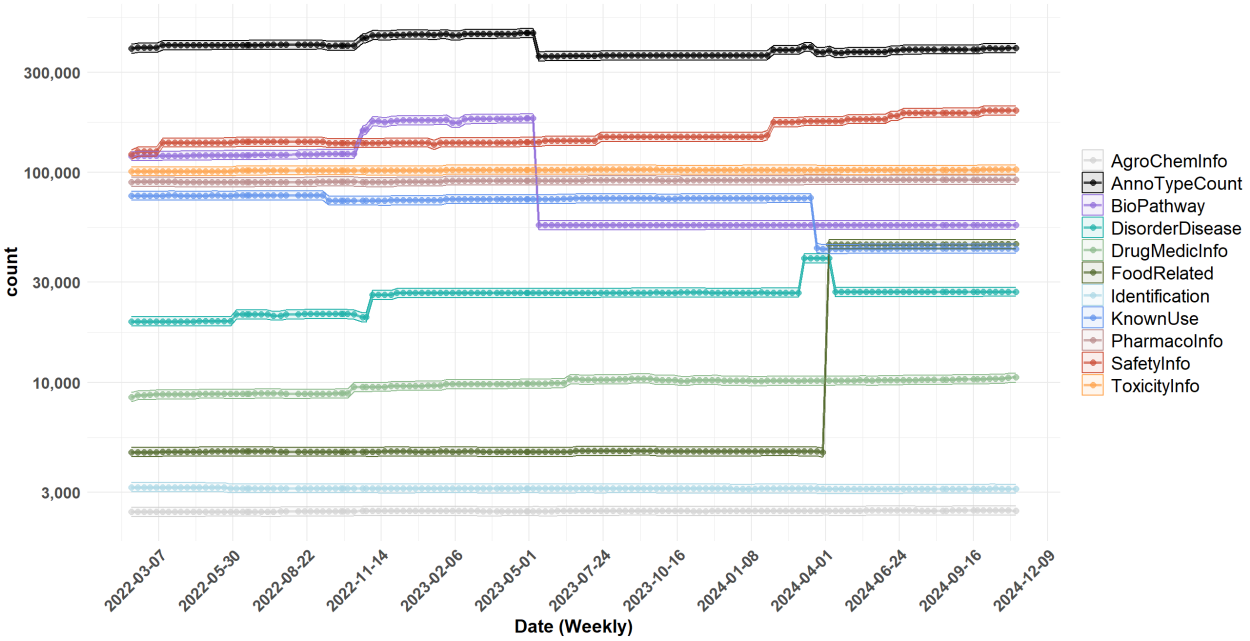

Figure 5: PubChemLite annotation content (total and by category) between 4 Feb. 2022 and 3 Nov. 2024.

Overall, despite PubChem increasing in content dramatically over that time, PubChemLite has remained generally stable at ~380,000 entries, growing slowly over time. The systematic increase of the BioPathway category (purple, Figure 5) starting in October 2022 introduced a number of irrelevant candidates in preliminary results of non-target studies[40] and caused continual expansion; this content was switched to the "Pathways" subcategory rather than the entire "Interactions and Pathways" heading in May 2023 (see Figure 1), improving performance and interpretability of candidate hits. The dramatic increase in FoodRelated information was due to the integration of FooDB[41] into PubChem annotation content; despite the large increase in that category, the overall candidate numbers remained stable, indicating that many of these candidates already had other annotation content in PubChemLite. The increase and then decrease of content in the DiseaseDisorder category in March-April 2024 was due to an update of one data source that suddenly introduced many low-quality candidates - upon contacting the contributor they checked their data and identified several issues; the fixes were processed by 12 April (see Figure 5). Overall, the continuous monitoring and use of PubChemLite in various NTS studies helps ensure relevance and usefulness for the community.

CCS Values

The experimental CCS data in PubChem currently (5 Nov. 2024) includes a total of 22,192 experimental CCS values corresponding to 8099 unique compounds (CIDs). The contributions include 1554 CCS values for 1136 CIDs from the Baker Lab[24,25]; 17,187 CCS values for 6242 CIDs from CCSbase[20]; and 3451 CCS values corresponding to 869, 574, 148 and 205 CIDs from the NORMAN-SLE[12] collections S50 CCSCOMPEND[26,27], S61 UJICCSLIB[28,29], S79 UACCSCEC[3,30] and S116 REFCCS[31], respectively. Information is available for 98 adducts, where the most common adducts are $[M+H]^+$ (7545 CCS values, 4278 CIDs), $[M-H]^-$ (4279 CCS values, 2064 CIDs), $[M+Na]^+$ (4064 CCS values, 2831 CIDs), $[M+K]^+$ (1179 CCS values, 1140 CIDs) and $[M+H-H_2O]^+$ (1154 CCS values, 1113 CIDs).

Predicted CCS values calculated with CCSbase are available for all but 12 CIDs in PubChemLite, since these could not be parsed with RDKit (the toolkit used in cs3db), despite being compatible with the toolkits used in PubChem (OpenEye) and MetFrag (CDK). Toolkit compatibility has been explored elsewhere recently[42]. Nevertheless, 12 failures out of 389,779 CIDs corresponds to only ~0.003% of the dataset. In contrast, 11,373 (~2.9%) XlogP values are missing from the same file (values cannot be calculated with the XlogP model[43] used by PubChem). The CCSbase model currently in use has been trained on a slightly different set to the public CCSbase website and the numbers integrated in PubChemLite may differ slightly. Since a major motivation to improve the availability of open experimental CCS values was to increase the amount of relevant data for predictive systems such as CCSbase, the pipelines presented in this article have been designed to allow upgrades of the CCSbase model once they are ready. The predicted CCS values in PubChemLite have already been applied in NTS studies[44].

Future Perspectives

PubChemLite has been used in several NTS applications; user feedback helps ensure the relevance for environmental screening. Collaborative research activities have already identified less relevant content (as described above), but also poor coverage (possible lack of annotation content) for compounds in sediments. The chemical stripe integration helps researchers visualize and interpret the chemical history over time[36], while the annotation summary pages display basic information available simply, providing direct access to PubChem for the full content. The predicted CCS values will help improve NTS workflows for IMS, while the growth of the experimental CCS dataset will help predictive models such as CCSbase improve accuracy and relevance over time. PubChemLite is openly available (https://pubchemlite.lcsb.uni.lu) - feedback is welcome (see contact page).

Declarations

Data Availability Statement

PubChemLite is compiled weekly from openly available files on the PubChem FTP site and is archived monthly on Zenodo (DOI: 10.5281/zenodo.5995885). CCS values are added using open cs3db code and the PubChemLite-CCS files are archived on Zenodo at DOI: 10.5281/zenodo.4081056. The Zenodo links redirect to the latest version. The code for the PubChemLite build system, inputs, chemical stripes and interface are available on the Environmental Cheminformatics (ECI) GitLab. All are available under open licenses, see individual resources for details.

Acknowledgements

The authors acknowledge the earlier efforts of Todor Kondic (now at LDNS) to parts of this work, and Christine Gallampois (Umea University) for her insights on the sediment NTS, as well as the Environmental Cheminformatics, Bioinformatics Core, Xu lab, BakerLab and PubChem team members and other colleagues and collaborators who contributed to this work indirectly via other collaborative and scientific activities and discussions.

Funding

AE, DA and ELS acknowledge funding support from the Luxembourg National Research Fund (FNR) for project A18/BM/12341006 (AE, DA, ELS), the University of Luxembourg Institute for Advanced Studies (IAS) for the Audacity project "LuxTIME" (DA, ELS) and the European Union Research and Innovation program Horizon Europe for PARC, Grant No. 101057014 (AE). The work of SK, PAT, JZ and EEB was supported by the National Center for Biotechnology Information of the National Library of Medicine (NLM), National Institutes of Health. JND and ESB would like to acknowledge funding support from the National Institute of Environmental Health Sciences (P42 ES027704) and National Institute of General Medical Sciences (R01 GM141277 and RM1 GM145416). LX acknowledges financial support from the National Institute of Environmental Health Sciences, National Institutes of Health (R01 ES031927).

Author Contributions

AE: Data curation, Methodology, Software (PubChemLite build, evaluation), Validation, Writing original draft preparation (joint), Writing review and editing. DHR: Methodology, Software (CCSbase, cs3db), Validation, Writing review and editing. VG: Methodology, Software (PubChemLite-web), Visualization, Writing review and editing. DA: Methodology, Software (chemical stripes), Visualization, Writing review and editing. AMK: Methodology, Software (CCSbase). SK: Methodology, Software (PubChem-CCS interface), Validation, Writing review and editing. PAT: Methodology, Software (PubChemLite, PubChem-CCS interface, experimental CCS), Validation, Writing review and editing. JZ: Data curation, Methodology, Software (PubChemLite, PubChem-CCS interface, experimental CCS), Validation, Writing review and editing. JND: Data curation, Supervision, Writing review and editing. ESB: Data curation, Project administration, Resources, Supervision, Writing review and editing. EEB: Conceptualization, Data curation, Methodology, Project administration, Resources, Software (PubChemLite), Supervision, Validation, Writing review and editing. LX: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Software (CCSbase), Supervision, Writing review and editing. ELS: Conceptualization, Data curation, Funding acquisition, Methodology, Project administration, Resources, Software (PubChemLite, evaluation, experimental CCS), Supervision, Validation, Visualization, Writing original draft preparation (joint), Writing review and editing.

Conflicts of Interest

The authors have no competing financial interests to declare.

Supporting Information

See Data Availability Statement.

References

(1) Hollender, J.; Schymanski, E. L.; Ahrens, L.; Alygizakis, N.; Béen, F.; Bijlsma, L.; Brunner, A. M.; Celma, A.; Fildier, A.; Fu, Q.; Gago-Ferrero, P.; Gil-Solsona, R.; Haglund, P.; Hansen, M.; Kaserzon, S.; Kruve, A.; Lamoree, M.; Margoum, C.; Meijer, J.; Merel, S.; Rauert, C.; Rostkowski, P.; Samanipour, S.; Schulze, B.; Schulze, T.; Singh, R. R.; Slobodnik, J.; Steininger-Mairinger, T.; Thomaidis, N. S.; Togola, A.; Vorkamp, K.; Vulliet, E.; Zhu, L.; Krauss, M. NORMAN Guidance on Suspect and Non-Target Screening in Environmental Monitoring. *Environmental Sciences Europe* **2023**, *35* (1), 75. https://doi.org/10.1186/s12302-023-00779-4.

(2) Lai, Y.; Koelmel, J. P.; Walker, D. I.; Price, E. J.; Papazian, S.; Manz, K. E.; Castilla-Fernández, D.; Bowden, J. A.; Nikiforov, V.; David, A.; Bessonneau, V.; Amer, B.; Seethapathy, S.; Hu, X.; Lin, E. Z.; Jbebli, A.; McNeil, B. R.; Barupal, D.; Cerasa, M.; Xie, H.; Kalia, V.; Nandakumar, R.; Singh, R.; Tian, Z.; Gao, P.; Zhao, Y.; Froment, J.; Rostkowski, P.; Dubey, S.; Coufalíková, K.; Seličová, H.; Hecht, H.; Liu, S.; Udhani, H. H.; Restituito, S.; Tchou-Wong, K.-M.; Lu, K.; Martin, J. W.; Warth, B.; Godri Pollitt, K. J.; Klánová, J.; Fiehn, O.; Metz, T. O.; Pennell, K. D.; Jones, D. P.; Miller, G. W. High-Resolution Mass Spectrometry for Human Exposomics: Expanding Chemical Space Coverage. *Environmental Science & Technology* **2024**, *58* (29), 12784–12822. https://doi.org/10.1021/acs.est.4c01156.

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

(4) Celma, A.; Bade, R.; Sancho, J. V.; Hernandez, F.; Humphries, M.; Bijlsma, L. Prediction of Retention Time and Collision Cross Section (CCSH+, CCSH-, and CCSNa+) of Emerging Contaminants Using Multiple Adaptive Regression Splines. *Journal of Chemical Information and Modeling* **2022**, *62* (22), 5425–5434. https://doi.org/10.1021/acs.jcim.2c00847.

(5) Song, X.-C.; Dreolin, N.; Canellas, E.; Goshawk, J.; Nerin, C. Prediction of Collision Cross-Section Values for Extractables and Leachables from Plastic Products. *Environmental Science & Technology* **2022**, *56* (13), 9463–9473. https://doi.org/10.1021/acs.est.2c02853.

(6) Wishart, D. S.; Guo, A.; Oler, E.; Wang, F.; Anjum, A.; Peters, H.; Dizon, R.; Sayeeda, Z.; Tian, S.; Lee, B. L.; Berjanskii, M.; Mah, R.; Yamamoto, M.; Jovel, J.; Torres-Calzada, C.; Hiebert-Giesbrecht, M.; Lui, V. W.; Varshavi, D.; Varshavi, D.; Allen, D.; Arndt, D.; Khetarpal, N.; Sivakumaran, A.; Harford, K.; Sanford, S.; Yee, K.; Cao, X.; Budinski, Z.; Liigand, J.; Zhang, L.; Zheng, J.; Mandal, R.; Karu, N.; Dambrova, M.; Schiöth, H. B.; Greiner, R.; Gautam, V. HMDB 5.0: The Human Metabolome Database for 2022. *Nucleic Acids Research* **2022**, *50* (D1), D622–D631. https://doi.org/10.1093/nar/gkab1062.

(7) 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. *Journal of Cheminformatics* **2017**, *9* (1), 61. https://doi.org/10.1186/s13321-017-0247-6.

(8) Pence, H. E.; Williams, A. ChemSpider: An Online Chemical Information Resource. *Journal of Chemical Education* **2010**, *87* (11), 1123–1124. https://doi.org/10.1021/ed100697w.

(9) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2023 Update. *Nucleic Acids Research* **2023**, *51*, gkac956. https://doi.org/10.1093/nar/gkac956.

(10) American Chemical Society. CAS REGISTRY - The CAS Substance Collection, 2024. https://www.cas.org/cas-data/cas-registry (accessed 2024-08-03).

(11) Wang, Z.; Walker, G. W.; Muir, D. C. G.; Nagatani-Yoshida, K. Toward a Global Understanding of Chemical Pollution: A First Comprehensive Analysis of National and Regional Chemical Inventories. *Environmental Science & Technology* **2020**, *54* (5), 2575–2584. https://doi.org/10.1021/acs.est.9b06379.

(12) Mohammed Taha, H.; Aalizadeh, R.; Alygizakis, N.; Antignac, J.-P.; Arp, H. P. H.; Bade, R.; Baker, N.; Belova, L.; Bijlsma, L.; Bolton, E. E.; Brack, W.; Celma, A.; Chen, W.-L.; Cheng, T.; Chirsir, P.; Čirka, Ľ.; D'Agostino, L. A.; Djoumbou Feunang, Y.; Dulio, V.; Fischer, S.; Gago-Ferrero, P.; Galani, A.; Geueke, B.; Głowacka, N.; Glüge, J.; Groh, K.; Grosse, S.; Haglund, P.; Hakkinen, P. J.; Hale, S. E.; Hernandez, F.; Janssen, E. M.-L.; Jonkers, T.; Kiefer, K.; Kirchner, M.; Koschorreck, J.; Krauss, M.; Krier, J.; Lamoree, M. H.; Letzel, M.; Letzel, T.; Li, Q.; Little, J.; Liu, Y.; Lunderberg, D. M.; Martin, J. W.; McEachran, A. D.; McLean, J. A.; Meier, C.; Meijer, J.; Menger, F.; Merino, C.; Muncke, J.; Muschket, M.; Neumann, M.; Neveu, V.; Ng, K.; Oberacher, H.; O'Brien, J.; Oswald, P.; Oswaldova, M.; Picache, J. A.; Postigo, C.; Ramirez, N.; Reemtsma, T.; Renaud, J.; Rostkowski, P.; Rüdel, H.; Salek, R. M.; Samanipour, S.; Scheringer, M.; Schliebner, I.; Schulz, W.; Schulze, T.; Sengl, M.; Shoemaker, B. A.; Sims, K.; Singer, H.; Singh, R. R.; Sumarah, M.; Thiessen, P. A.; Thomas, K. V.; Torres, S.; Trier, X.; Wezel, A. P. van; Vermeulen, R. C. H.; Vlaanderen, J. J.; Ohe, P. C. von der; Wang, Z.; Williams, A. J.; Willighagen, E. L.; Wishart, D. S.; Zhang, J.; Thomaidis, N. S.; Hollender, J.; Slobodnik, J.; Schymanski, E. L. The NORMAN Suspect List Exchange (NORMAN-SLE): Facilitating European and Worldwide Collaboration on Suspect Screening in High Resolution Mass Spectrometry. *Environmental Sciences Europe* **2022**, *34* (1), 104. https://doi.org/10.1186/s12302-022-00680-6.

(13) Schymanski, E. L.; Kondić, T.; Neumann, S.; Thiessen, P. A.; Zhang, J.; Bolton, E. E. Empowering Large Chemical Knowledge Bases for Exposomics: PubChemLite Meets MetFrag. *Journal of Cheminformatics* **2021**, *13* (1), 19. https://doi.org/10.1186/s13321-021-00489-0.

(14) Helmus, R.; Laak, T. L. ter; Wezel, A. P. van; Voogt, P. de; Schymanski, E. L. patRoon: Open Source Software Platform for Environmental Mass Spectrometry Based Non-Target Screening. *Journal of Cheminformatics* **2021**, *13* (1), 1. https://doi.org/10.1186/s13321-020-00477-w.

(15) Ruttkies, C.; Schymanski, E. L.; Wolf, S.; Hollender, J.; Neumann, S. MetFrag Relaunched: Incorporating Strategies Beyond in Silico Fragmentation. *Journal of Cheminformatics* **2016**, *8* (1), 3. https://doi.org/10.1186/s13321-016-0115-9.

(16) NCBI/NLM/NIH. PubChem Table of Contents Classification Browser, 2024. https://pubchem.ncbi.nlm.nih.gov/classification/#hid=72 (accessed 2024-07-08).

(17) Ieritano, C.; Hopkins, W. S. Assessing Collision Cross Section Calculations Using MobCal-MPI with a Variety of Commonly Used Computational Methods. *Materials Today Communications* **2021**, *27*, 102226. https://doi.org/10.1016/j.mtcomm.2021.102226.

(18) Colby, S. M.; Thomas, D. G.; Nuñez, J. R.; Baxter, D. J.; Glaesemann, K. R.; Brown, J. M.; Pirrung, M. A.; Govind, N.; Teeguarden, J. G.; Metz, T. O.; Renslow, R. S. ISiCLE: A Quantum Chemistry Pipeline for

Establishing in Silico Collision Cross Section Libraries. *Analytical Chemistry* **2019**, *91* (7), 4346–4356. https://doi.org/10.1021/acs.analchem.8b04567.

(19) Zhou, Z.; Luo, M.; Chen, X.; Yin, Y.; Xiong, X.; Wang, R.; Zhu, Z.-J. Ion Mobility Collision Cross-Section Atlas for Known and Unknown Metabolite Annotation in Untargeted Metabolomics. *Nature Communications* **2020**, *11* (1), 4334. https://doi.org/10.1038/s41467-020-18171-8.

(20) Ross, D. H.; Cho, J. H.; Xu, L. Breaking Down Structural Diversity for Comprehensive Prediction of Ion-Neutral Collision Cross Sections. *Analytical Chemistry* **2020**, *92* (6), 4548–4557. https://doi.org/10.1021/acs.analchem.9b05772.

(21) Guo, R.; Zhang, Y.; Liao, Y.; Yang, Q.; Xie, T.; Fan, X.; Lin, Z.; Chen, Y.; Lu, H.; Zhang, Z. Highly Accurate and Large-Scale Collision Cross Sections Prediction with Graph Neural Networks. *Communications Chemistry* **2023**, *6* (1), 1–10. https://doi.org/10.1038/s42004-023-00939-w.

(22) Plante, P.-L.; Francovic-Fontaine, É.; May, J. C.; McLean, J. A.; Baker, E. S.; Laviolette, F.; Marchand, M.; Corbeil, J. Predicting Ion Mobility Collision Cross-Sections Using a Deep Neural Network: DeepCCS. *Analytical Chemistry* **2019**, *91* (8), 5191–5199. https://doi.org/10.1021/acs.analchem.8b05821.

(23) Rainey, M. A.; Watson, C. A.; Asef, C. K.; Foster, M. R.; Baker, E. S.; Fernández, F. M. CCS Predictor 2.0: An Open-Source Jupyter Notebook Tool for Filtering Out False Positives in Metabolomics. *Analytical Chemistry* **2022**, *94* (50), 17456–17466. https://doi.org/10.1021/acs.analchem.2c03491.

(24) Kirkwood, K. I.; Christopher, M. W.; Burgess, J. L.; Littau, S. R.; Foster, K.; Richey, K.; Pratt, B. S.; Shulman, N.; Tamura, K.; MacCoss, M. J.; MacLean, B. X.; Baker, E. S. Development and Application of Multidimensional Lipid Libraries to Investigate Lipidomic Dysregulation Related to Smoke Inhalation Injury Severity. *Journal of Proteome Research* **2022**, *21* (1), 232–242. https://doi.org/10.1021/acs.jproteome.1c00820.

(25) Foster, M.; Rainey, M.; Watson, C.; Dodds, J. N.; Kirkwood, K. I.; Fernández, F. M.; Baker, E. S. Uncovering PFAS and Other Xenobiotics in the Dark Metabolome Using Ion Mobility Spectrometry, Mass Defect Analysis, and Machine Learning. *Environmental Science & Technology* **2022**, *56* (12), 9133–9143. https://doi.org/10.1021/acs.est.2c00201.

(26) Picache, J. A.; Rose, B. S.; Balinski, A.; Leaptrot, K. L.; Sherrod, S. D.; May, J. C.; McLean, J. A. Collision Cross Section Compendium to Annotate and Predict Multi-Omic Compound Identities. *Chemical Science* **2019**, *10* (4), 983–993. https://doi.org/10.1039/C8SC04396E.

(27) Picache, J.; McLean, J. S50 CCSCOMPEND The Unified Collision Cross Section (CCS) Compendium, 2019. https://doi.org/10.5281/zenodo.2658162.

(28) Celma, A.; Sancho, J. V.; Schymanski, E. L.; Fabregat-Safont, D.; Ibáñez, M.; Goshawk, J.; Barknowitz, G.; Hernández, F.; Bijlsma, L. Improving Target and Suspect Screening High-Resolution Mass Spectrometry Workflows in Environmental Analysis by Ion Mobility Separation. *Environmental Science & Technology* **2020**, *54* (23), 15120–15131. https://doi.org/10.1021/acs.est.0c05713.

(29) Celma, A.; Fabregat-Safont, D.; Ibàñez, M.; Bijlsma, L.; Hernandez, F.; Sancho, J. V. S61 UJICCSLIB Collision Cross Section (CCS) Library from UJI, 2019. https://doi.org/10.5281/ZENODO.3549476.

(30) Belova, L.; Caballero-Casero, N.; Nuijs, A. L. N. van; Covaci, A. S79 UACCSCEC Collision Cross Section (CCS) Library from UAntwerp, 2021. https://doi.org/10.5281/ZENODO.4704648.

(31) Muller, H.; Palm, E.; Schymanski, E. S116 REFCCS Collision Cross Section (CCS) Values from Literature, 2024. https://doi.org/10.5281/ZENODO.10932895.

(32) Grouès, V.; Rocca-Serra, P.; Ded, V. Elixir-Luxembourg/Data-Catalog, 2023. https://github.com/elixir-luxembourg/data-catalog (accessed 2024-08-04).

(33) Welter, D.; Rocca-Serra, P.; Grouès, V.; Sallam, N.; Ancien, F.; Shabani, A.; Asariardakani, S.; Alper, P.; Ghosh, S.; Burdett, T.; Sansone, S.-A.; Gu, W.; Satagopam, V. The Translational Data Catalog - Discoverable Biomedical Datasets. *Scientific Data* **2023**, *10* (1), 470. https://doi.org/10.1038/s41597-023-02258-0.

(34) Greg Landrum. RDKit: Open-Source Cheminformatics Software, 2024. https://www.rdkit.org/ (accessed 2024-08-04).

(35) Greg Landrum; Paolo Tosco; Brian Kelley; Ricardo Rodriguez; David Cosgrove; Riccardo Vianello; sriniker; gedeck; Gareth Jones; NadineSchneider; Eisuke Kawashima; Dan Nealschneider; Andrew Dalke; Matt Swain; Brian Cole; Samo Turk; Aleksandr Savelev; Alain Vaucher; Maciej Wójcikowski; Ichiru Take; Vincent F. Scalfani; Daniel Probst; Kazuya Ujihara; Rachel Walker; Axel Pahl; guillaume godin; tadhurst-cdd; Juuso Lehtivarjo; François Bérenger; Jonathan Bisson. Rdkit/Rdkit: 2024_03_5 (Q1 2024) Release, 2024. https://doi.org/10.5281/ZENODO.591637.

(36) Aurich, D.; Schymanski, E. L.; De Jesus Matias, F.; Thiessen, P. A.; Pang, J. Revealing Chemical Trends: Insights from Data-Driven Visualization and Patent Analysis in Exposomics Research. *Environmental Science & Technology Letters* **2024**, *11* (10), 1046–1052. https://doi.org/10.1021/acs.estlett.4c00560.

(37) Arp, H. P. H.; Aurich, D.; Schymanski, E. L.; Sims, K.; Hale, S. E. Avoiding the Next Silent Spring: Our Chemical Past, Present, and Future. *Environmental Science & Technology* **2023**, *57* (16), 6355–6359. https://doi.org/10.1021/acs.est.3c01735.

(38) Aurich, D. Uniluxembourg / LCSB / Environmental Cheminformatics / Chemicalstripes · GitLab. *GitLab*, 2024. https://gitlab.com/uniluxembourg/lcsb/eci/chemicalstripes (accessed 2024-08-04).

(39) Grouès, V. Uniluxembourg / LCSB / Environmental Cheminformatics / PubChemLite-Web · GitLab. *GitLab*, 2024. https://gitlab.com/uniluxembourg/lcsb/eci/pubchemlite-web (accessed 2024-08-04).

(40) Talavera Andújar, B.; Mary, A.; Venegas, C.; Cheng, T.; Zaslavsky, L.; Bolton, E. E.; Heneka, M. T.; Schymanski, E. L. Can Small Molecules Provide Clues on Disease Progression in Cerebrospinal Fluid from Mild Cognitive Impairment and Alzheimer's Disease Patients? *Environmental Science & Technology* **2024**, *58*, 4181–4192. https://doi.org/10.1021/acs.est.3c10490.

(41) WishartLab. FooDB, 2024. https://foodb.ca/ (accessed 2024-11-06).

(42) Barnabas, S. J.; Böhme, T.; Boyer, S. K.; Irmer, M.; Ruttkies, C.; Wetherbee, I.; Kondić, T.; Schymanski, E. L.; Weber, L. Extraction of Chemical Structures from Literature and Patent Documents Using Open Access Chemistry Toolkits: A Case Study with PFAS. *Digital Discovery* **2022**, *1* (4), 490–501. https://doi.org/10.1039/D2DD00019A.

(43) Cheng, T.; Zhao, Y.; Li, X.; Lin, F.; Xu, Y.; Zhang, X.; Li, Y.; Wang, R.; Lai, L. Computation of Octanol−Water Partition Coefficients by Guiding an Additive Model with Knowledge. *Journal of Chemical Information and Modeling* **2007**, *47* (6), 2140–2148. https://doi.org/10.1021/ci700257y.

(44) Menger, F.; Celma, A.; Schymanski, E. L.; Lai, F. Y.; Bijlsma, L.; Wiberg, K.; Hernández, F.; Sancho, J. V.; Ahrens, L. Enhancing Spectral Quality in Complex Environmental Matrices: Supporting Suspect and Non-Target Screening in Zebra Mussels with Ion Mobility. *Environment International* **2022**, *170*, 107585. https://doi.org/10.1016/j.envint.2022.107585.