Environmental Science & Technology Letters

pubs.acs.org/journal/estlcu

Letter

PubChemLite Plus Collision Cross Section (CCS) Values for Enhanced Interpretation of Nontarget Environmental Data

Published as part of Environmental Science & Technology Letters special issue "Non-Targeted Analysis of the Environment".

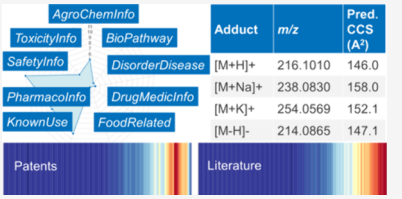
Anjana Elapavalore, Dylan H. Ross, Valentin Grouès, Dagny Aurich, Allison M. Krinsky, Sunghwan Kim, Paul A. Thiessen, Jian Zhang, James N. Dodds, Erin S. Baker, Evan E. Bolton,* Libin Xu,* and Emma L. Schymanski*

Cite This: https://doi.org/10.1021/acs.estlett.4c01003

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Finding relevant chemicals in the vast (known) chemical space is a major challenge for environmental and exposomics studies leveraging nontarget 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 historical trends of patent and literature data, for researchers to browse the collection. This article details how PubChemLite can support researchers in environmental and 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 can be browsed at https://pubchemlite.lcsb.uni.lu.



KEYWORDS: *nontarget screening, identification, PubChemLite, exposomics, ion mobility, collision cross section, PubChem*

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 environmental health. Nontarget 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 increasingly accessible, with the calculated collision cross section (CCS) values serving as an additional parameter to support identification in NTS.[3−5] Nonetheless, the identification and - importantly - interpretation of features detected during NTS is still challenging, hindering the broader adoption of NTS.[1] Identification in NTS primarily relies on mass spectral libraries, relatively small suspect lists and large chemical databases, as recently reviewed elsewhere.[1,2] While the integration of IMS/CCS into workflows generally remains relatively poor, several methods to predict CCS values are now available to alleviate this situation, including quantum (*e.g.,* MobCal[6] and ISiCLE[7]) and machine learning (ML) methods (*e.g.,* AllCCS,[8] CCSbase,[9] SigmaCCS,[10] DeepCCS,[11] and CCS Predictor 2.0[12]).

Compound databases commonly used in NTS include (numbers as of Dec. 2024) HMDB[13] (220,945 metabolites), CompTox[14] (1,218,248 chemicals), PubChem[15] (119 million compounds), and ChemSpider[16] (129 million structures). The Chemical Abstract Services (CAS) Registry[17] (219 million chemicals) is licensed and thus unavailable to open workflows. Since ChemSpider introduced programmatic access limitations 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

Received: November 21, 2024
Revised: December 31, 2024
Accepted: January 2, 2025

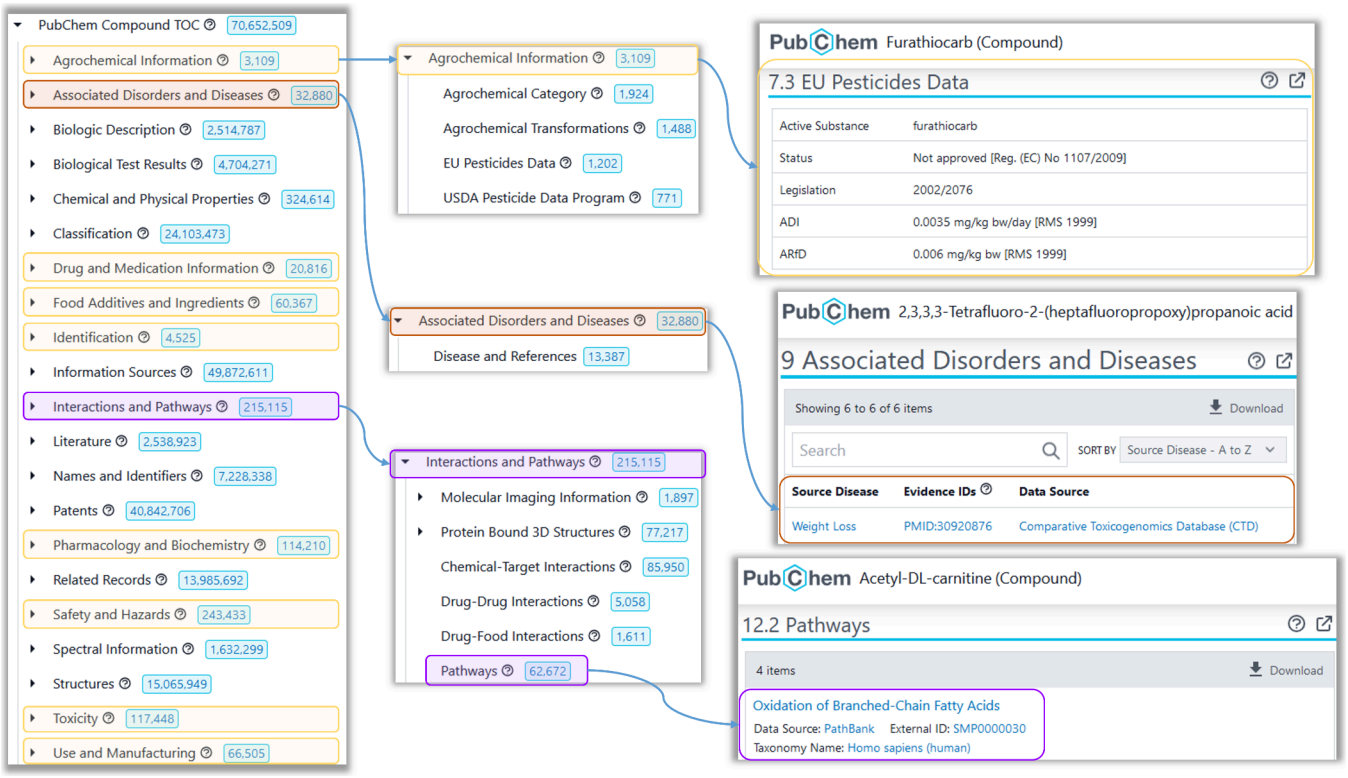


Figure 1. PubChemLite categories in the PubChem Table of Contents (TOC) (https://pubchem.ncbi.nlm.nih.gov/classification/#hid=72), selected subcategories, and associated annotation examples. Yellow shading denotes "environmental" categories (example CID 47759) (https://pubchem.ncbi.nlm.nih.gov/compound/47759#section=EU-Pesticides-Data), red the "exposomics" (example CID 114481) (https://pubchem.ncbi.nlm.nih.gov/compound/114481#section=Associated-Disorders-and-Diseases) and purple the "metabolomics" sections (example CID 1) (https://pubchem.ncbi.nlm.nih.gov/compound/1#section=Pathways). For high resolution live images, please click the embedded hyperlinks. Logo image from GitLab.[28]

be found in the environment, nor pertinent to the exposome. This hinders both the performance and efficiency of NTS. Additionally, many other potential sources for NTS identification efforts, such as the Global Chemical Inventory (350,000 chemicals)[18] and various lists from European regulators contributing to the NORMAN Suspect List Exchange (NORMAN-SLE)[19] 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.[20] PubChemLite has been integrated into existing HR-MS workflows, such as patRoon[21] and MetFrag.[22] Although PubChemLite is familiar to many researchers already, the original 2021 article[20] was primarily technical. This article explains PubChemLite for an environmental/exposomics audience, describes the integration of predicted CCS values in PubChemLite to support IMS, introduces the development of an open experimental CCS pipeline in PubChem to provide more experimental data for future CCS predictions, and presents a new web interface (https://pubchemlite.lcsb.uni.lu/) to browse the contents of PubChemLite.

METHODS AND MATERIALS

Building PubChemLite. The full technical details of PubChemLite are published elsewhere.[20] Briefly, PubChemLite is derived from major categories relevant to environmental/exposomics applications appearing in the PubChem Table of Contents (TOC) (https://pubchem.ncbi.nlm.nih.gov/classification/#hid=72) pages.[23] 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), and 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 (https://gitlab.com/uniluxembourg/lcsb/eci/pubchemlite-input)[24] and code for the PubChemLite build system (https://gitlab.com/uniluxembourg/lcsb/eci/pclbuild)[25] are available on the Environmental Cheminformatics (ECI) GitLab (https://gitlab.com/uniluxembourg/lcsb/eci/)[26] 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 InChIKey into a primary entry and related CIDs, where the primary entry is the neutral or "parent" form. Entries such as mixtures and disconnected substances are

https://doi.org/10.1021/acs.estlett.4c01003
Environ. Sci. Technol. Lett. XXXX, XXX, XXX−XXX

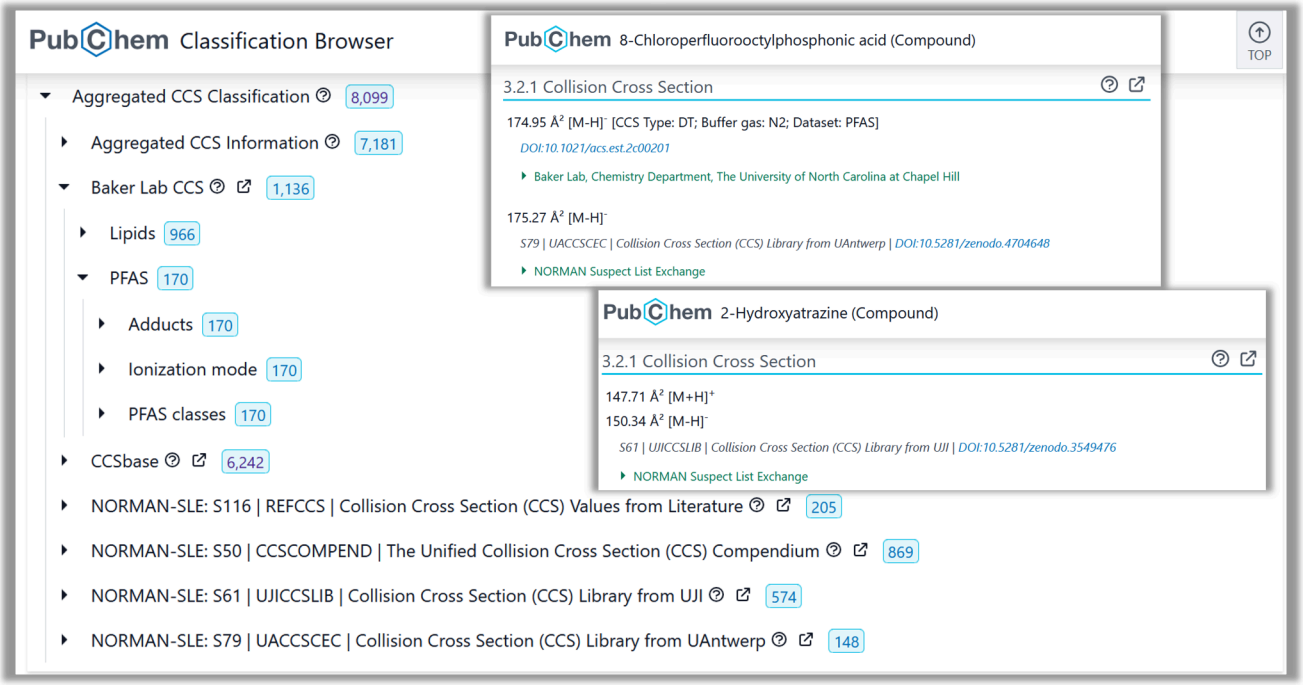


Figure 2. Aggregated Collision Cross Section (CCS) Classification Tree (https://pubchem.ncbi.nlm.nih.gov/classification/#hid=106) in PubChem. Inset: Experimental CCS values in individual PubChem compound records for Cl-PFOPA (CID 138395139) (https://pubchem.ncbi.nlm.nih.gov/compound/138395139#section=Collision-Cross-Section) and the transformation product 2-hydroxyatrazine (CID 135398733) (https://pubchem.ncbi.nlm.nih.gov/compound/135398733#section=Collision-Cross-Section). For high resolution live images, please click the embedded hyperlinks. Logo image from GitLab.[28]

excluded (see the original publication[20] for details). Chemical information (SMILES, InChI, InChIKey, formula, mass), patent, and literature (PubMed) counts plus predicted XlogP values are retrieved in bulk. 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. PubChemLite is built and evaluated each week, with one public release per month.[27]

Adding Predicted CCS Values to PubChemLite. Although quantum CCS prediction methods such as ISiCLE[7] are generally considered more accurate than ML models, the calculation times are prohibitive for the PubChemLite pipeline. Furthermore, since ISiCLE only predicts values for C, H, N, O, P, and S-containing compounds, CCS values would be missing for ∼40% of PubChemLite. In contrast, the current version of the ML method CCSbase[9] runs for all but 12 entries (0.003%) of PubChemLite and completes in 3650 s. Calculations are performed and released publicly once a month using the cs3db (https://github.com/dylanhross/c3sdb/)[29] model (the code behind CCSbase) trained on the data sets listed in Table S1 of the Supporting Information (SI) as published in Ross et al.[9]

Both versions (PubChemLite with and without CCS values) are integrated into MetFrag[22,30] each month (see SI, Section S2).

Adding Experimental CCS Values to PubChem. To ensure that ML models have better coverage of environmentally relevant compounds to improve their predictions in the future, part of this work involved establishing a pipeline to integrate experimental CCS values into PubChem. Currently PubChem contains CCS values from the Baker Lab,[31,32] CCSbase[9] and four collections via the NORMAN-SLE:[19] S50 CCSCOMPEND,[33,34] S61 UJICCSLIB,[35,36] S79 UACCSCEC,[3,37] and S116 REFCCS.[38] These values are displayed on individual records in PubChem and navigable in the PubChem Classification Browser via the CCSbase (https://pubchem.ncbi.nlm.nih.gov/classification/#hid=104),[39] Baker Lab (https://pubchem.ncbi.nlm.nih.gov/classification/#hid=124),[40] NORMAN-SLE (https://pubchem.ncbi.nlm.nih.gov/classification/#hid=101)[41] and the Aggregated CCS (https://pubchem.ncbi.nlm.nih.gov/classification/#hid=106)[42] trees (see Figure 2). All experimental CCS values are retrieved from PubChem (code available on GitLab (https://gitlab.com/uniluxembourg/lcsb/eci/pubchem/-/blob/master/annotations/CCS/CCS_retrieval)[43]) and archived on Zenodo[44] after each update.

PubChemLite Web Interface. The PubChemLite web interface is developed as a plugin for the ELIXIR-Luxembourg Data Catalog (https://github.com/elixir-luxembourg/data-catalog).[45,46] It is developed in Python, CSS, HTML and Javascript, using RDKit[47,48] for structure depiction. For full details, see the PubChemLite-web (https://gitlab.com/uniluxembourg/lcsb/eci/pubchemlite-web) code on GitLab.[49] The information in the archived PubChemLite-CCSbase CSV files (see Figure 3) is enhanced with additional synonyms from PubChem Downloads[50] for improved searchability, visualizations of the annotation categories, and tables of the CCS and associated adduct mass values. Finally, historical literature and patent trends are included using the "chemical stripes"[51,52] where available (see Figure 4). The original R version[53] was rewritten in Python for integration in PubChemLite-web

https://doi.org/10.1021/acs.estlett.4c01003
Environ. Sci. Technol. Lett. XXXX, XXX, XXX−XXX

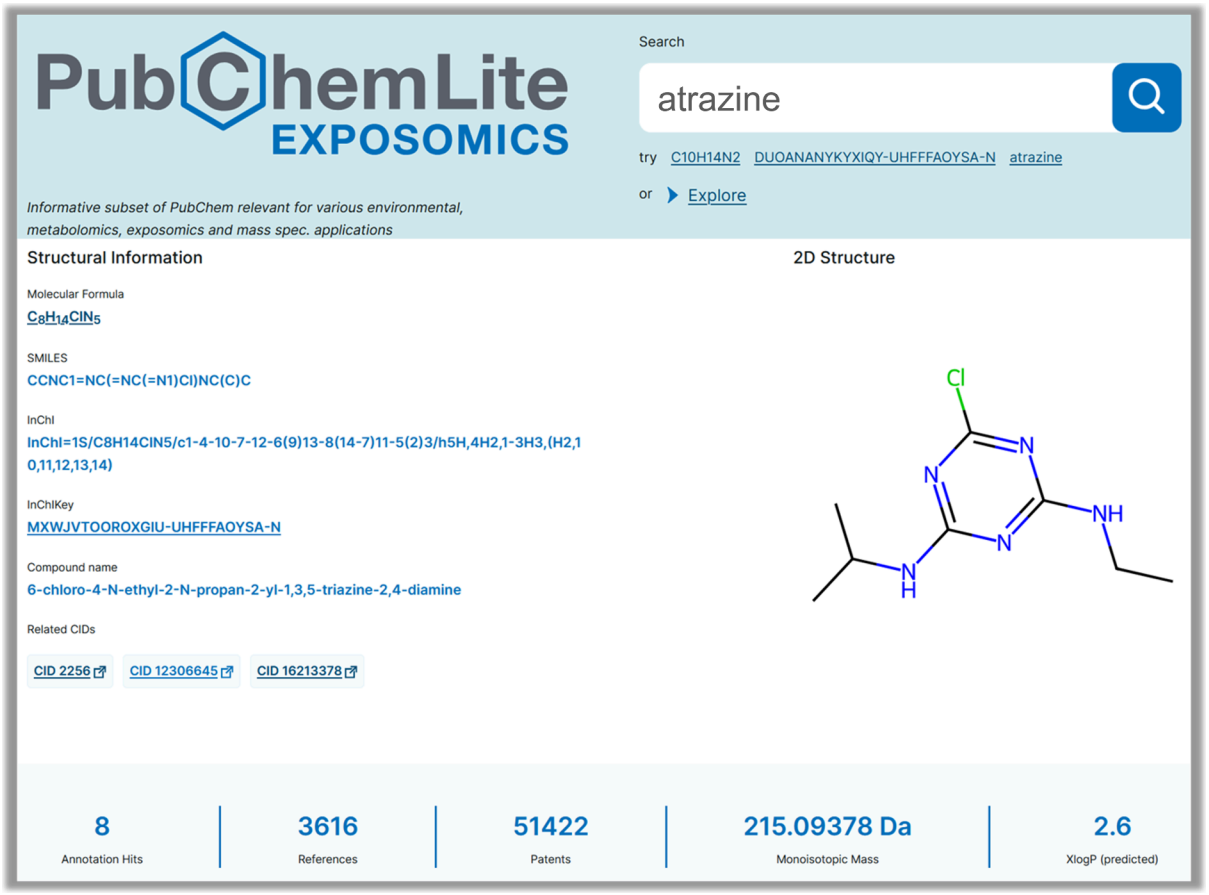


Figure 3. PubChemLite web interface (composite image), compound view of Atrazine (https://pubchemlite.lcsb.uni.lu/e/compound/2256). For high resolution live images, please click the embedded hyperlink. Logo image from GitLab.[28]

(https://gitlab.com/uniluxembourg/lcsb/eci/pubchemlite-web),[49] with code available in both repositories.[49,53]

RESULTS AND DISCUSSION

PubChemLite Over Time. The performance of PubChemLite is monitored weekly with every build using the evaluation data set of 977 compounds established in the original publication.[20] 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). Further details are given in the SI, Section S3, Table S3. 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.

Overall, despite PubChem increasing in content dramatically over that time, PubChemLite has remained generally stable at ∼400,000 entries. The systematic increase of the BioPathway category (purple, Figure 5) starting in October 2022 introduced a number of irrelevant candidates in preliminary NTS results[54] and was alleviated by switching to the "Pathways" subcategory 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[55] 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, which was fixed 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.

Incorporating CCS Values into Candidate Selection with PubChemLite. The application of PubChemLite with MetFrag using the recommended scoring terms (MetFrag score, MoNA Exact Match, AnnoTypeCount, PubMed_Count, Patent_Count) is described in Section S2 of the SI using an example that highlights how all information could be considered when interpreting candidate results. In this case, the data scores rank one candidate first (carbaryl), whereas experimental data clearly point to desethylterbutylazine as the correct structure. As shown in Table S2, experimental CCS values would distinguish all three candidates (accounting for 1% error), but the predicted values (considering 3% error)[9] would not (see Section S2, SI for more details). Many cases in the evaluation set are similar, where the predicted values for most candidates are within a 3% prediction error. However, in some cases, some candidates can be eliminated based on predicted CCS values, such as the example of Acemetacin (see Figure S6). These observations match the results of Menger et al. applying PubChemLite-CCS in NTS of mussels.[56] That study also highlighted discrepancies in predicted CCS values for some environmentally relevant compounds, especially per- and polyfluorinated substances (PFAS). This motivated the

https://doi.org/10.1021/acs.estlett.4c01003
Environ. Sci. Technol. Lett. XXXX, XXX, XXX−XXX

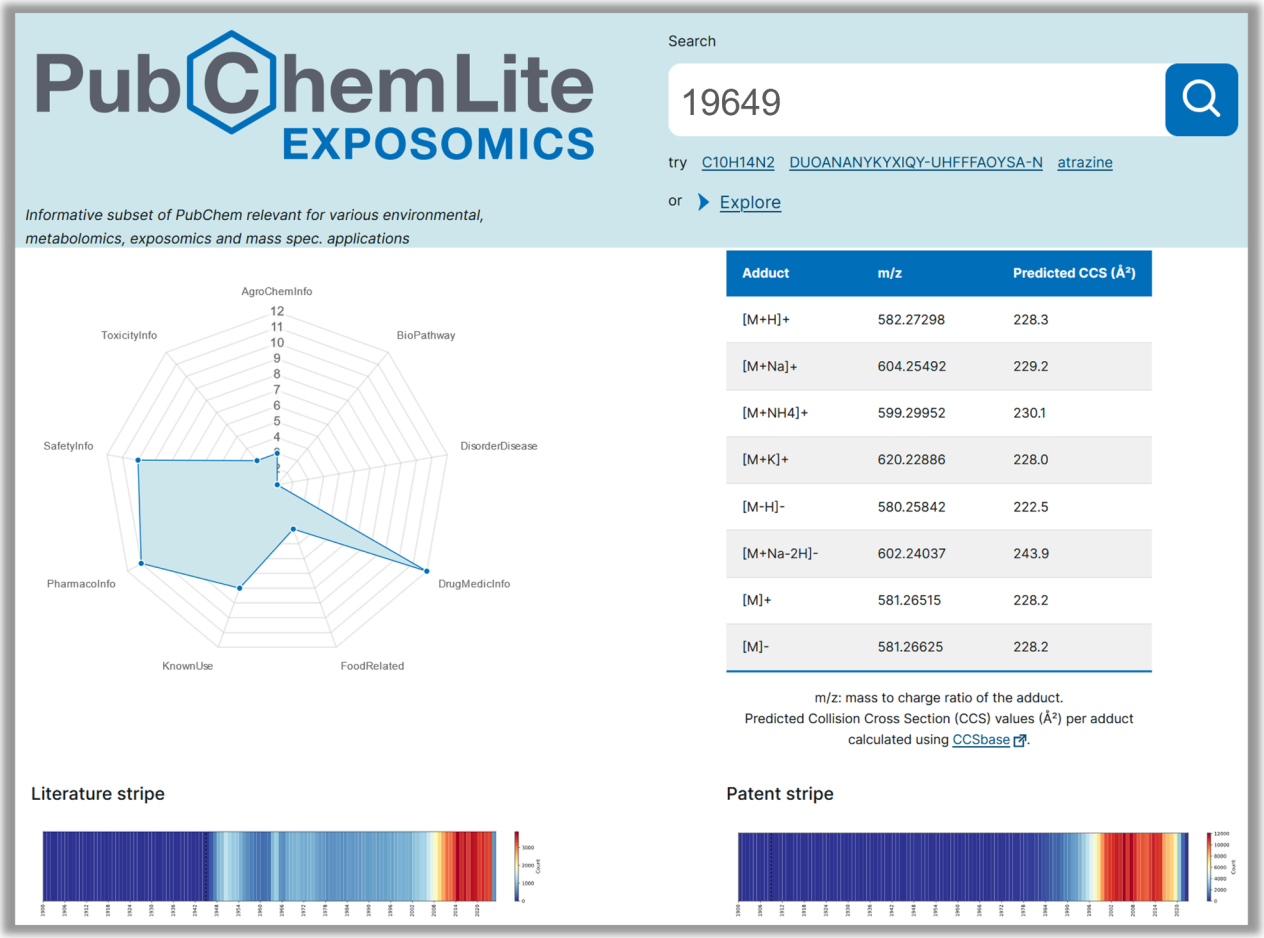


Figure 4. PubChemLite web interface (composite image), view of additional data including annotations, CCS values, and patent and literature stripes for Streptomycin (https://pubchemlite.lcsb.uni.lu/e/compound/19649). For high resolution live images, please click the embedded hyperlink. Logo image from GitLab.[28]

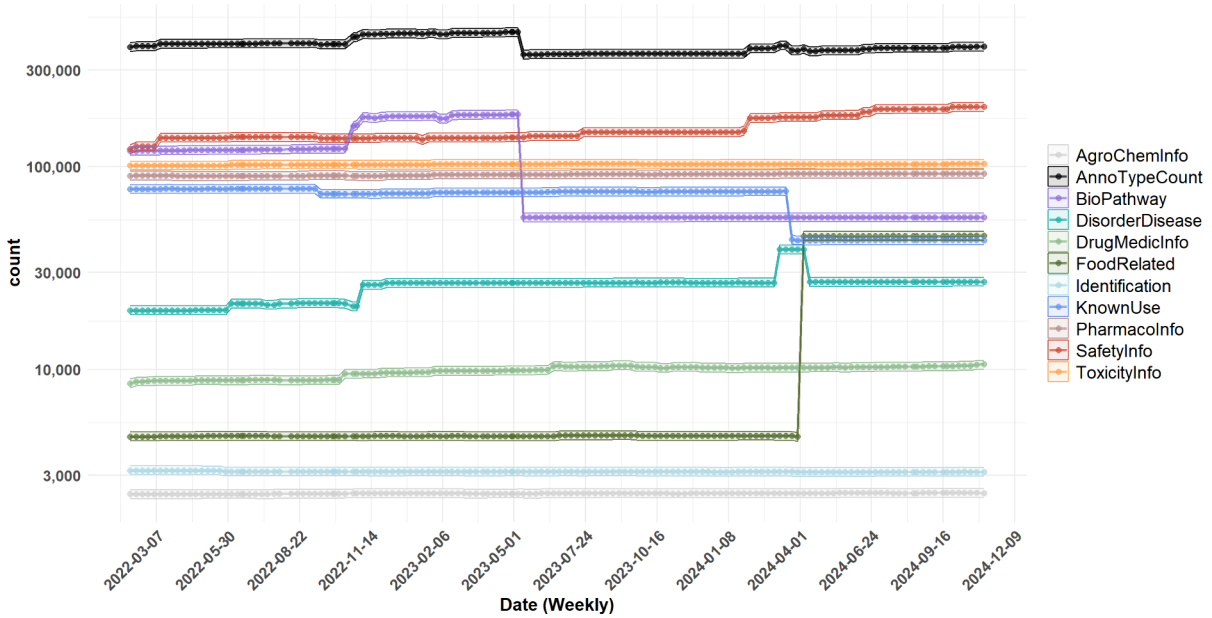


Figure 5. PubChemLite annotation content (total and by category) between 4 Feb. 2022, and 3 Nov. 2024.

collection of the experimental CCS data described in the next paragraph to improve the availability of relevant environmental CCS values for training ML approaches such as the CCSbase.

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), somewhat

https://doi.org/10.1021/acs.estlett.4c01003
Environ. Sci. Technol. Lett. XXXX, XXX, XXX−XXX

smaller than the recently released METLIN-CCS collections.[57,58] The contributions include 1554 CCS values for 1136 CIDs from the Baker Lab;[31,32] 17,187 CCS values for 6242 CIDs from CCSbase;[9] and 3451 CCS values corresponding to 869, 574, 148, and 205 CIDs from the NORMAN-SLE[19] collections S50 CCSCOMPEND,[33,34] S61 UJICCSLIB,[35,36] S79 UACCSCEC,[3,37] and S116 REFCCS,[38] 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).

Future Perspectives. PubChemLite continues to develop and improve as a resource for the environmental and exposomics communities based on user feedback. Collaborative research activities have helped trim less relevant content but also identified poor coverage for compounds in sediments, which requires the integration of additional annotation content into PubChem to address fully. Features to be added to PubChemLite in the short term include mass and CCS search options for the web interface and the addition of the "chemical classes" category to include more emerging contaminants like flame retardants. The CCS predictions will be updated once new versions of c3sdb (trained on more data) are available. Separate community efforts are underway to automate identification in NTS using ion mobility data, and the efforts presented here form an important basis for this.

PubChemLite (https://pubchemlite.lcsb.uni.lu) provides efficient candidate sets (tens to hundreds rather than thousands) and information-rich content for interpreting environmental NTS data, with 80% of the evaluation set ranked Top 1 and 98% Top 5, empowering identification in NTS studies. As CCS predictions improve with new technology, in particular, through incorporation of experimental reference values from higher resolving power IMS separations, this performance is likely to improve further. User feedback is welcome (see https://pubchemlite.lcsb.uni.lu/contact).

■ ASSOCIATED CONTENT

Data Availability Statement

The PubChemLite web interface (https://pubchemlite.lcsb.uni.lu) is openly available. PubChemLite is compiled weekly from openly available files downloaded from PubChem[50] and is archived monthly on Zenodo (DOI: https://doi.org/10.5281/zenodo.5995885). CCS values are added using open cs3db (https://github.com/dylanhross/c3sdb/) code[29] and the PubChemLite-CCS files are archived on Zenodo at DOI: https://doi.org/10.5281/zenodo.4081056. The Zenodo links redirect to the latest version. The code for the PubChemLite build system (https://gitlab.com/uniluxembourg/lcsb/eci/pclbuild),[25] inputs (https://gitlab.com/uniluxembourg/lcsb/eci/pubchemlite-input),[24] chemical stripes (https://gitlab.com/uniluxembourg/lcsb/eci/chemicalstripes)[53] and interface (https://gitlab.com/uniluxembourg/lcsb/eci/pubchemlite-web)[49] are openly available on the Environmental Cheminformatics (ECI) GitLab (https://gitlab.com/uniluxembourg/lcsb/eci/).[26] All resources are available under open licenses, see individual pages for details. This article was submitted as a preprint: Anjana Elapavalore, Dylan Ross, Valentin Groues, Dagny Aurich, Allison Krinsky, Sunghwan Kim, Paul Thiessen, Jian Zhang, James Dodds, Erin Baker, Evan Bolton, Libin Xu, Emma Schymanski. 2024. ChemRxiv. DOI: https://doi.org/10.26434/chemrxiv-2024-2xcsq.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.estlett.4c01003.

A document including additional details about the CCSbase training data sets (S1), using PubChemLite in MetFrag (S2), and additional rank and CCS results (S3) (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Evan E. Bolton − *National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), Bethesda, Maryland 20894, United States;* orcid.org/0000-0002-5959-6190; Phone: +1 301 451 1811; Email: bolton@ncbi.nlm.nih.gov; Fax: +1 301 480 9241

Libin Xu − *Department of Medicinal Chemistry, University of Washington, Seattle, Washington 98195, United States;* orcid.org/0000-0003-1021-5200; Phone: +1 206 543-1080; Email: libinxu@uw.edu; Fax: +1 206 685 3252

Emma L. Schymanski − *Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 4367 Belvaux, Luxembourg;* orcid.org/0000-0001-6868-8145; Phone: +352 46 66 44 5616; Email: emma.schymanski@uni.lu

Authors

Anjana Elapavalore − *Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 4367 Belvaux, Luxembourg*

Dylan H. Ross − *Department of Medicinal Chemistry, University of Washington, Seattle, Washington 98195, United States; Current Address: Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99352, United States;* orcid.org/0009-0005-2943-2282

Valentin Grouès − *Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 4367 Belvaux, Luxembourg*

Dagny Aurich − *Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 4367 Belvaux, Luxembourg;* orcid.org/0000-0001-8823-0596

Allison M. Krinsky − *Department of Medicinal Chemistry, University of Washington, Seattle, Washington 98195, United States*

Sunghwan Kim − *National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), Bethesda, Maryland 20894, United States;* orcid.org/0000-0001-9828-2074

Paul A. Thiessen − *National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), Bethesda, Maryland 20894, United States;* orcid.org/0000-0002-1992-2086

Jian Zhang − *National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), Bethesda, Maryland 20894, United States*

James N. Dodds − *Department of Chemistry, University of North Carolina, Chapel Hill, North Carolina 27599, United States;* orcid.org/0000-0002-9702-2294

https://doi.org/10.1021/acs.estlett.4c01003
Environ. Sci. Technol. Lett. XXXX, XXX, XXX−XXX

Erin S. Baker − *Department of Chemistry, University of North Carolina, Chapel Hill, North Carolina 27599, United States;* orcid.org/0000-0001-5246-2213

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.estlett.4c01003

Author Contributions

A.E.: Data curation, Methodology, Software (PubChemLite build, evaluation), Validation, Writing original draft preparation (joint), Writing review and editing. D.H.R.: Methodology, Software (CCSbase, cs3db), Validation, Writing review and editing. V.G.: Methodology, Software (PubChemLite-web), Visualization, Writing review and editing. D.A.: Methodology, Software (chemical stripes), Visualization, Writing review and editing. A.M.K.: Methodology, Software (CCSbase). S.K.: Methodology, Software (PubChem-CCS interface), Validation, Writing review and editing. P.A.T.: Methodology, Software (PubChemLite, PubChem-CCS interface, experimental CCS), Validation, Writing review and editing. J.Z.: Data curation, Methodology, Software (PubChemLite, PubChem-CCS interface, experimental CCS), Validation, Writing review and editing. J.N.D.: Data curation, Supervision, Writing review and editing. E.S.B.: Data curation, Project administration, Resources, Supervision, Writing review and editing. E.E.B.: Conceptualization, Data curation, Methodology, Project administration, Resources, Software (PubChemLite), Supervision, Validation, Writing review and editing. L.X.: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Software (CCSbase), Supervision, Writing review and editing. E.L.S.: 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.

Funding

A.E., D.A., and E.L.S. acknowledge funding support from the Luxembourg National Research Fund (FNR) for project A18/BM/12341006 (A.E., D.A., E.L.S.), the University of Luxembourg Institute for Advanced Studies (IAS) for the Audacity project "LuxTIME" (D.A., E.L.S.) and the European Union Research and Innovation program Horizon Europe for PARC, Grant No. 101057014 (A.E.). The work of S.K., P.A.T., J.Z., and E.E.B. was supported by the National Center for Biotechnology Information of the National Library of Medicine (NLM), National Institutes of Health. J.N.D. and E.S.B. 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). L.X. acknowledges financial support from the National Institute of Environmental Health Sciences, National Institutes of Health (R01 ES031927).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge the earlier efforts of Todor Kondic (now at LDNS) to parts of this work, Steffen Neumann (IPB Halle) for his merging of countless monthly pull requests into MetFrag, Rick Helmus (University of Amsterdam) for the continuing patRoon integration, 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, and finally the reviewers and editor for their comments and suggestions.

■ 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.; et al. NORMAN Guidance on Suspect and Non-Target Screening in Environmental Monitoring. *Environmental Sciences Europe* **2023**, *35* (1), 75.

(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.; et al. High-Resolution Mass Spectrometry for Human Exposomics: Expanding Chemical Space Coverage. *Environ. Sci. Technol.* **2024**, *58* (29), 12784−12822.

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

(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. *J. Chem. Inf. Model.* **2022**, *62* (22), 5425−5434.

(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. *Environ. Sci. Technol.* **2022**, *56* (13), 9463−9473.

(6) Ieritano, C.; Hopkins, W. S. Assessing Collision Cross Section Calculations Using MobCal-MPI with a Variety of Commonly Used Computational Methods. *Mater. Today Commun.* **2021**, *27*, No. 102226.

(7) 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. *Anal. Chem.* **2019**, *91* (7), 4346−4356.

(8) 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. *Nat. Commun.* **2020**, *11* (1), 4334.

(9) Ross, D. H.; Cho, J. H.; Xu, L. Breaking Down Structural Diversity for Comprehensive Prediction of Ion-Neutral Collision Cross Sections. *Anal. Chem.* **2020**, *92* (6), 4548−4557.

(10) 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.

(11) 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. *Anal. Chem.* **2019**, *91* (8), 5191−5199.

(12) 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. *Anal. Chem.* **2022**, *94* (50), 17456−17466.

(13) 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.; et al. HMDB 5.0: The Human Metabolome Database for 2022. *Nucleic Acids Res.* **2022**, *50* (D1), D622−D631.
(14) 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.
(15) 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 Res.* **2023**, *51*, D1373−D1380.
(16) Pence, H. E.; Williams, A. ChemSpider: An Online Chemical Information Resource. *J. Chem. Educ.* **2010**, *87* (11), 1123−1124.
(17) American Chemical Society. *CAS REGISTRY - The CAS Substance Collection*, 2024. https://www.cas.org/cas-data/cas-registry (accessed 2024-08-03).
(18) 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. *Environ. Sci. Technol.* **2020**, *54* (5), 2575−2584.
(19) 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, L.; D'Agostino, L. A.; Djoumbou Feunang, Y.; Dulio, V.; Fischer, S.; et al. 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.
(20) 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.
(21) Helmus, R.; ter Laak, T. L.; van Wezel, A. P.; de Voogt, P.; Schymanski, E. L. patRoon: Open Source Software Platform for Environmental Mass Spectrometry Based Non-Target Screening. *Journal of Cheminformatics* **2021**, *13* (1), 1.
(22) 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.
(23) NCBI/NLM/NIH. *PubChem Table of Contents Classification Browser*, 2024. https://pubchem.ncbi.nlm.nih.gov/classification/#hid=72 (accessed 2024-07-08).
(24) LCSB-ECI. Uniluxembourg/LCSB/Environmental Cheminformatics/Pubchemlite-Input. *GitLab*, 2024. https://gitlab.com/uniluxembourg/lcsb/eci/pubchemlite-input (accessed 2024-12-16).
(25) LCSB-ECI. Uniluxembourg/LCSB/Environmental Cheminformatics/Pubchemlite-Build-System. *GitLab*, 2024. https://gitlab.com/uniluxembourg/lcsb/eci/pclbuild (accessed 2024-12-16).
(26) LCSB-ECI. Uniluxembourg/LCSB/Environmental Cheminformatics. *GitLab*, 2024. https://gitlab.com/uniluxembourg/lcsb/eci/ (accessed 2024-12-16).
(27) Bolton, E.; Schymanski, E.; Kondic, T.; Thiessen, P.; Zhang, J. PubChemLite for Exposomics. *Zenodo* **2024**, DOI: 10.5281/zenodo.5995885.
(28) LCSB-ECI. Pubchemlite/Logos. *GitLab*, 2021. https://gitlab.com/uniluxembourg/lcsb/eci/pubchem/-/tree/master/pubchemlite/logos (accessed 2024-12-17).
(29) Ross, D. C3SDB (Combined Collision Cross Section DataBase) - Dylanhross/C3sdb. *GitHub*, 2024. https://github.com/dylanhross/c3sdb (accessed 2024-12-16).
(30) IPB Halle. *MetFrag Web*, 2024. https://msbi.ipb-halle.de/MetFrag/ (accessed 2024-12-03).
(31) 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. *J. Proteome Res.* **2022**, *21* (1), 232−242.
(32) 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. *Environ. Sci. Technol.* **2022**, *56* (12), 9133−9143.
(33) 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.
(34) Picache, J.; McLean, J. S50 CCSCOMPEND The Unified Collision Cross Section (CCS) Compendium. *Zenodo* **2019**, DOI: 10.5281/zenodo.2658162.
(35) 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. *Environ. Sci. Technol.* **2020**, *54* (23), 15120−15131.
(36) Celma, A.; Fabregat-Safont, D.; Ibàñez, M.; Bijlsma, L.; Hernandez, F.; Sancho, J. V. S61 UJICCSLIB Collision Cross Section (CCS) Library from UJI. *Zenodo* **2019**, DOI: 10.5281/zenodo.3549476.
(37) Belova, L.; Caballero-Casero, N.; Nuijs, A. L. N. van; Covaci, A. S79 UACCSCEC Collision Cross Section (CCS) Library from UAntwerp. *Zenodo* **2021**, DOI: 10.5281/zenodo.4704648.
(38) Muller, H.; Palm, E.; Schymanski, E. S116 REFCCS Collision Cross Section (CCS) Values from Literature. *Zenodo* **2024**, DOI: 10.5281/zenodo.10932895.
(39) PubChem. PubChem Classification Browser: CCSbase Classification, 2024. https://pubchem.ncbi.nlm.nih.gov/classification/#hid=104 (accessed 2024-12-17).
(40) PubChem. *PubChem Classification Browser: CCS Classification - Baker Lab*, 2024. https://pubchem.ncbi.nlm.nih.gov/classification/#hid=124 (accessed 2024-12-17).
(41) PubChem. *PubChem Classification Browser: NORMAN-SLE Classification*, 2024. https://pubchem.ncbi.nlm.nih.gov/classification/#hid=106 (accessed 2024-12-17).
(42) PubChem. *PubChem Classification Browser: Aggregated CCS Classification*, 2024. https://pubchem.ncbi.nlm.nih.gov/classification/#hid=106 (accessed 2024-12-17).
(43) Schymanski, E. Annotations/CCS/CCS_retrieval · Master · Uniluxembourg/LCSB/Environmental Cheminformatics/Pubchem. *GitLab*, 2024. https://gitlab.com/uniluxembourg/lcsb/eci/pubchem/-/tree/master/annotations/CCS/CCS_retrieval (accessed 2024-12-16).
(44) Schymanski, E.; Zhang, J.; Thiessen, P.; Bolton, E. Experimental CCS Values in PubChem. *Zenodo* **2024**, DOI: 10.5281/zenodo.6800138.
(45) Grouès, V.; Rocca-Serra, P.; Ded, V. Elixir-Luxembourg/Data-Catalog. *GitHub*, 2023. https://github.com/elixir-luxembourg/data-catalog (accessed 2024-08-04).
(46) 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.
(47) Landrum, G.. *RDKit: Open-Source Cheminformatics Software*, 2024. https://www.rdkit.org/ (accessed 2024-08-04).
(48) Landrum, G.; Tosco, P.; Kelley, B.; Rodriguez, R.; Cosgrove, D.; Vianello, R.; sriniker; gedeck; Jones, G.; NadineSchneider; Kawashima, E.; Nealschneider, D.; Dalke, A.; Swain, M.; Cole, B.; Turk, S.; Savelev, A.; Vaucher, A.; Wójcikowski, M.; Take, I. Rdkit/Rdkit: 2024_03_5 (Q1 2024) Release. *Zenodo* **2024**, DOI: 10.5281/zenodo.591637.
(49) Grouès, V. Uniluxembourg/LCSB/Environmental Cheminformatics/PubChemLite-Web. *GitLab*, 2024. https://gitlab.com/uniluxembourg/lcsb/eci/pubchemlite-web (accessed 2024-08-04).
(50) NCBI/NLM/NIH. *PubChem Download Pages*, 2024. https://ftp.ncbi.nlm.nih.gov/pubchem/ (accessed 2024-12-16).

https://doi.org/10.1021/acs.estlett.4c01003
Environ. Sci. Technol. Lett. XXXX, XXX, XXX−XXX

(51) 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. *Environ. Sci. Technol. Lett.* **2024**, *11* (10), 1046−1052.
(52) 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. *Environ. Sci. Technol.* **2023**, *57* (16), 6355−6359.
(53) Aurich, D. Uniluxembourg/LCSB/Environmental Cheminformatics/Chemicalstripes. *GitLab*, 2024. https://gitlab.com/uniluxembourg/lcsb/eci/chemicalstripes (accessed 2024-08-04).
(54) 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? *Environ. Sci. Technol.* **2024**, *58*, 4181−4192.
(55) WishartLab. *FooDB*, 2024. https://foodb.ca/ (accessed 2024-11-06).
(56) 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. *Environ. Int.* **2022**, *170*, No. 107585.
(57) Baker, E. S.; Hoang, C.; Uritboonthai, W.; Heyman, H. M.; Pratt, B.; MacCoss, M.; MacLean, B.; Plumb, R.; Aisporna, A.; Siuzdak, G. METLIN-CCS: An Ion Mobility Spectrometry Collision Cross Section Database. *Nat. Methods* **2023**, *20* (12), 1836−1837.
(58) Baker, E. S.; Uritboonthai, W.; Aisporna, A.; Hoang, C.; Heyman, H. M.; Connell, L.; Olivier-Jimenez, D.; Giera, M.; Siuzdak, G. METLIN-CCS Lipid Database: An Authentic Standards Resource for Lipid Classification and Identification. *Nature Metabolism* **2024**, *6* (6), 981−982.

https://doi.org/10.1021/acs.estlett.4c01003
Environ. Sci. Technol. Lett. XXXX, XXX, XXX−XXX