Title

Exploring chemical space in non-targeted analysis: a proposed ChemSpace tool

Journal

Analytical and Bioanalytical Chemistry, 415(1)

ISSN

1618-2642

Authors

Black, Gabrielle

Lowe, Charles

Anumol, Tarun

et al.

Publication Date

2023

DOI

10.1007/s00216-022-04434-4

Peer reviewed

Published in final edited form as:
Anal Bioanal Chem. 2023 January ; 415(1): 35–44. doi:10.1007/s00216-022-04434-4.

Exploring chemical space in non-targeted analysis: a proposed ChemSpace tool

Gabrielle Black[#1], **Charles Lowe**[#2], **Tarun Anumol**[3], **Jessica Bade**[4], **Kristin Favela**[5], **Yong-Lai Feng**[6], **Ann Knolhoff**[7], **Andrew Mceachran**[3], **Jamie Nuñez**[4], **Christine Fisher**[7], **Kathy Peter**[8], **Natalia Soares Quinete**[9], **Jon Sobus**[2], **Eric Sussman**[10], **William Watson**[5], **Samanthi Wickramasekara**[11], **Antony Williams**[2], **Tom Young**[1]

[1]Department of Civil & Environmental Engineering, University of California Davis, Davis, CA, USA

[2]U.S. EPA, Office of Research and Development, Center for Computational Toxicology and Exposure, Research Triangle Park, NC, USA

[3]Agilent Technologies, Inc., Santa Clara, CA, USA

[4]Pacific Northwest National Laboratory, Richland, WA, USA

[5]Southwest Research Institute, San Antonio, TX, USA

[6]Exposure and Biomonitoring Division, Environmental Health Science and Research Bureau, Health Canada, Ottawa, ON, Canada

[7]U.S. Food and Drug Administration, Center for Food Safety and Applied Nutrition, College Park, MD, USA

[8]Center for Urban Waters, University of Washington Tacoma, Tacoma, WA 98421, USA

[9]Department of Chemistry and Biochemistry, Institute of Environment, Florida International University, North Miami, FL, USA

[10]MCRA LLC, Washington, DC, USA

[11]U.S. Food and Drug Administration, Center for Devices and Radiological Health, Silver Spring, MD, USA

[#] These authors contributed equally to this work.

Abstract

Non-targeted analysis (NTA) using high-resolution mass spectrometry allows scientists to detect and identify a broad range of compounds in diverse matrices for monitoring exposure and toxicological evaluation without a priori chemical knowledge. NTA methods present an opportunity to describe the constituents of a sample across a multidimensional swath of chemical properties, referred to as “chemical space.” Understanding and communicating which region of chemical space is extractable and detectable by an NTA workflow, however, remains challenging and non-standardized. For example, many sample processing and data analysis steps influence

Gabrielle Black gnpecora@ucdavis.edu.

Conflicts of interest The authors declare no conflicts of interest in the publication of this manuscript.

the types of chemicals that can be detected and identified. Accordingly, it is challenging to assess whether analyte non-detection in an NTA study indicates true absence in a sample (above a detection limit) or is a false negative driven by workflow limitations. Here, we describe the need for accessible approaches that enable chemical space mapping in NTA studies, propose a tool to address this need, and highlight the different ways in which it could be implemented in NTA workflows. We identify a suite of existing predictive and analytical tools that can be used in combination to generate scores that describe the likelihood a compound will be detected and identified by a given NTA workflow based on the predicted chemical space of that workflow. Higher scores correspond to a higher likelihood of compound detection and identification in a given workflow (based on sample extraction, data acquisition, and data analysis parameters). Lower scores indicate a lower probability of detection, even if the compound is truly present in the samples of interest. Understanding the constraints of NTA workflows can be useful for stakeholders when results from NTA studies are used in real-world applications and for NTA researchers working to improve their workflow performance. The hypothetical ChemSpaceTool suggested herein could be used in both a prospective and retrospective sense. Prospectively, the tool can be used to further curate screening libraries and set identification thresholds. Retrospectively, false detections can be filtered by the plausibility of the compound identification by the selected NTA method, increasing the confidence of unknown identifications. Lastly, this work highlights the chemometric needs to make such a tool robust and usable across a wide range of NTA disciplines and invites others who are working on various models to participate in the development of the ChemSpaceTool. Ultimately, the development of a chemical space mapping tool strives to enable further standardization of NTA by improving method transparency and communication around false detection rates, thus allowing for more direct method comparisons between studies and improved reproducibility. This, in turn, is expected to promote further widespread applications of NTA beyond research-oriented settings.

Keywords

Non-targeted analysis; Mass spectrometry; Cheminformatics; Chemical space; Quality assurance/control

Introduction

Non-targeted analysis (NTA), using high-resolution mass spectrometry (HRMS), is a comprehensive approach for screening and identifying broad suites of organic compounds without a priori knowledge of their presence in samples. Although NTA, in principle, can detect and identify compounds with nearly any characteristics, no single method can detect or identify chemicals over the totality of the chemical universe. To access different domains of chemical space, numerous methods and software tools have been developed, but it can be challenging to determine the capabilities and limitations of a given workflow. In light of this challenge, the Benchmarking and Publications for Non-Targeted Analysis (BP4NTA) [1] group is working to standardize and promote guidelines for consistent and transparent NTA study reporting; the recently published Study Reporting Tool (SRT) [2] was an initial step toward this end. While we expect overall improvement in study reporting quality from use of

the SRT, considerable uncertainty remains in reporting the chemical space, or applicability domain, of an NTA method.

Current approaches to NTA generally involve individual laboratories developing their own analysis procedure(s) based on instrument types, acquisition and analysis methods, personnel expertise, research goals, and numerous other factors. The combination of such factors directly influences the compounds detectable and, in turn, makes comparing results from separate experiments (particularly across laboratories) difficult to nearly impossible. Therefore, a need exists for approaches that can delineate the region of chemical space accessible by an NTA method. Such approaches are expected to enhance annotation confidence using NTA methods by minimizing false-positive candidates (i.e., undetectable compounds reported as present) and confirming the plausibility of putative positive identifications that fall within the defined method applicability domain. The development of chemical space mapping tools could also give researchers the ability to reduce the vast known chemical universe into lists of plausibly detectable and identifiable compounds. Such amenable compound lists (ACLs) could then be used as identification libraries for annotation efforts as part of data processing workflows. Ultimately, understanding method boundaries would allow researchers to better assess method needs on a project-by-project basis and better communicate and compare methods and results.

Accordingly, we propose development of a tool that would allow any laboratory or research group to apply a systematic workflow to define and easily communicate the regions of chemical space accessible using a given NTA method. Defining detectable spaces would enable the direct comparison of inter-laboratory results. In cases where laboratories utilize different methods and workflows resulting in different chemical space coverage, comparison would still be possible within the overlapping portion(s) of the chemical spaces. The proposed tool could provide a consistent basis for achieving and evaluating reproducibility, reliability, and accuracy of results across laboratories, which is critical for working toward acceptance of NTA studies and results for routine monitoring and regulatory use. It can also aid in experimental design when there are compound classes of interest (e.g., for validation purposes or minimum reporting goals). At the time of this publication, there is currently no tool available to define chemical space coverage, making the proposed tool the first of its kind.

This work describes the proposed development and intended uses of a Chemical Space Tool (ChemSpaceTool), which would support numerous prospective and retrospective research activities (Fig. 1). While the proposed tool is in the early stages of development and not yet available, many existing tools and models can be leveraged to begin building the ChemSpaceTool. However, additional efforts are required for full conception of a comprehensive after conventional NTA workflows are conducted. ACLs can be used as a filter to prioritize plausible structure annotations, rather than as an annotation library itself, and can be used to improve the accuracy of performance metrics through a better understanding of true versus false identifications based on method-amenable compounds and robust product. We focus here on the framework for this tool and its anticipated benefit to NTA researchers. Specifically, this manuscript is aimed at:

1. *Introducing the proposed ChemSpaceTool* involving a series of query-based filtering steps to pare down a suspect library to a list of compounds plausibly detectable given input constraints
2. *Describing anticipated applications* of the tool to support prospective method development and retrospective method evaluation. Illustrative examples are provided to demonstrate its potential applicability and impact on NTA studies
3. *Providing a call to action* for NTA, cheminformatics, and other interested communities to design and advance chemical prediction models with the ability to contribute to this tool, ultimately enhancing chemical identification capabilities

The proposed ChemSpace tool

The ChemSpaceTool is envisioned to incorporate multiple filtering steps based on method and instrumentation parameters to reduce input libraries to ACLs that contain compounds that are plausibly detectable and identifiable in analyzed samples. These filtering steps would be based on an ensemble of chemometric tools to determine which compounds (within a definable degree of uncertainty) are likely within the boundaries of the chemical space of a given method. The tool would partition chemical space into three parts: (1) the detectable space, (2) the identifiable space, and (3) the region that includes compounds not detectable or identifiable using the selected methods (Fig. 2).

The detectable space

Detectable space is defined by the compounds amenable to detection using the methods applied for sampling, sample preparation, and data acquisition. This space comprises compounds that are likely to be *present* in a sample matrix type (e.g., water-soluble compounds in water samples, nonpolar compounds in fatty tissues), *extractable* by a sample preparation method, and *detectable* on the instrument platform used (e.g., polar and semi-polar compounds via LC–MS versus volatile and semi-volatile lower-polarity compounds via GC–MS). Thus, researchers make numerous decisions in the initial steps of NTA method selection and development that influence the types of compounds covered by that method. Mapping this detectable space requires application of various chemometric tools and filters to pare down the potential chemical universe ($> 10^{60}$) of organic compounds [3] to ACLs defined by the intersection of compounds that are plausibly present in a sample type, extractable, *and* detectable.

Eight separate filtering parameters that are exceptionally influential in defining chemical space have been identified as a starting point for detectable space mapping: (1) sample matrix type, (2) extraction solvent, (3) extract pH, (4) extraction/cleanup media, (5) elution buffers, (6) instrument platform, (7) ionization type, and (8) ionization mode. Each step would produce an ACL that occupies a defined region of chemical space (e.g., compounds that are water soluble, extractable via a hydrophilic-lipophilic solid-phase extraction media, LC–MS amenable, etc.). Ultimately, the eight ACLs resulting from these filtering steps would be compared with the overlapping compounds representing the detectable space (Fig. 3). These compounds would comprise the detectable space ACL which would

include accompanying quantitative molecular descriptors (e.g., water solubility, logK_{ow}, etc.) for each compound. Descriptors used to filter compounds into ACLs will be selected based on chemometric modeling of experimental data. For instance, many sorbents and chromatographic materials are proprietary, but data on individual compound retainability is prolific enough to inform predictive models allowing subsequent deduction of specific molecular descriptors driving retention. These descriptors would support filtering ACLs for categories such as chromatographic or extraction/cleanup media. These molecular descriptors can be used to assign a plausibility score and an applicability domain index (ADI), indicating the likelihood that the compound is within the defined chemical space and how appropriately the compound fits into each model used for prediction.

This initial vision for the ChemSpaceTool addresses attributes known to fundamentally affect the types of chemicals that are detected in a sample. However, given the diversity of options available for sample preparation, the need for additional tool functionality is expected. For instance, including a step to encompass inclusion or loss due to concentration (solvent reduction) techniques (e.g., nitrogen drying and solid-phase microextraction (SPME)) or solvent-exchange steps may help to further refine the detectable space. Additionally, allowing researchers to filter ACLs based on use of derivatization steps, such as hydrazine-derivatized carbonyls for LC–MS analysis or to improve volatility for GC–MS analysis may also be useful. Including multidimensional separation options such as ion mobility, two-dimensional chromatography (GC × GC and LC × LC), and HILIC-to-reverse-phase chromatography would be expected to further improve the resolution of the chemical space boundaries. Including physical and chemical characteristics of a given sample (e.g., dissolved organic carbon, total suspended solids, % lipid, % moisture, pH, etc.) could be useful as additional refinement options for sample type filters. Similarly, advancements in analytical hardware and software solutions will inevitably influence chemical space and an analysis of chemical space should be repeated with any workflow improvements/modifications.

Additional conditions, such as instrument cleanliness and salt contamination, may also impact the detectable chemical space. However, these conditions are significantly more difficult to predict and model. As with all NTA tools, quality control (QC) mixtures are critical for validating the ChemSpaceTool.

ChemSpace QC mixtures

Based on the minimum, median, and maximum values of the quantitative molecular descriptors used to create the detectable space ACL, a set of validation steps with quality control (QC) compounds would be suggested. Based on the instrument platform used, a standardized QC-NTA mix (similar to that proposed by Knolhoff et al. [4]) would serve as the foundation of the ChemSpaceTool QC mix (QC mix), with additional compounds suggested based on ChemSpaceTool filtering parameters; the suggested requirements for such a mixture are outlined in the referenced paper. The suggested ChemSpaceTool QC mix would be used for both validation of the detectable space boundary and for quality control during the general NTA workflow. The QC mix would be spiked (1) into the matrix and processed alongside samples (when possible), (2) in the extract of a sample (or pooled

sample), and (3) in solvent directly before data acquisition, where each compound would be evaluated for extractability and detectability using targeted data processing methodologies (e.g., searching the data directly for these standard compounds). The aim of a matrix spike is to evaluate whether the compound is extractable and detectable, when present in matrix. If a compound is spiked and is detected despite being in a dirty matrix, it is within the detectable space. If it is not detected in the matrix spike, but is detectable in spiked solvent, it is understood that it is not detectable when in the presence of matrix components. In some instances, matrix interferences may suppress signal below detection levels, and therefore, high and low spike levels should be considered. Successfully detected compounds would be confirmed in the ChemSpaceTool and the boundary adjusted accordingly. Furthermore, this standard mixture can be used routinely to ensure that the method performance and data quality is consistent to ensure that the measurable chemical space has not been negatively impacted by experimental factors (e.g., changes in sensitivity, dynamic range, etc.). The standard mixture could also be analyzed at different concentrations to determine the dynamic range for different chemical classes.

In addition to making initial suggestions for compounds to include in a QC mix, alternative suggestions may also be provided to allow researchers to utilize compounds already on hand or appropriate to specific research goals, thus reducing the number of additional standards that need to be purchased for each project. For example, if the ChemSpaceTool suggests levorphanol (a schedule 2 opioid) as a compound appropriate for a project's QC mix, but the researcher does not have DEA approval for a scheduled drug, an alternative list of suggestions might indicate dextrorphan (a metabolite of dextromethorphan, an over-the-counter cough suppressant), an optical isomer [5] that is easier to obtain and is equally informative of chemical space coverage.

"Decoy" compounds, or compounds not predicted to be present or extractable, would be included in the QC mix. Decoy compounds would improve confidence in the location of the outer boundaries of the predicted detectable space (by confirming/rejecting their extractability and detectability retroactively in the ChemSpaceTool) allowing researchers to understand baseline thresholds for the plausibility score. For example, if a researcher is investigating contaminants in sewage sludge using an extraction pH of 2 and LC–MS analysis is conducted in ESI negative mode, nonylphenol ethoxylates (NPnEOs) would not be within the predicted detectable space (because they are not extracted under acidic conditions) [6]. The researcher would then spike these (as part of their QC mix) into the matrix to confirm that they were not extracted and/or detected and thus be able to (1) evaluate the performance of the ChemSpaceTool in delimiting the occupied chemical space and (2) report that these compounds would not have been detected in the sample, even if they had been present, because their detection is outside of the method scope. Furthermore, if any of the decoy compounds are identified during annotation steps in the NTA analysis, their match scores (e.g., for molecular formula generation, MS/MS matches, etc.) can help define thresholds for true-versus false-positive candidates, since they are known to be outside of the scope of the project. Decoy compounds just outside of the defined chemical space will be particularly useful because they enable researchers to better understand scoring thresholds in downstream identification efforts. It is also possible that some unexpected compounds may be detected (or not detected) despite being predicted to

be outside (or inside) the chemical space of the method. Additional investigations of these "fringe" compounds can be used to help determine which molecular descriptors are most important and/or are not being modeled well, which can then inform improvements to the component models and the overall robustness of the ChemSpaceTool.

The identifiable space

The identifiable space of an NTA method refers to the types of chemicals feasibly identified during data processing. There are a variety of unique data processing approaches (e.g., libraries, databases, filters, thresholds) that are vital for data prioritization and compound identification in NTA; however, each of these tools would affect the identifiable space boundary. For example, using exclusive lists such as US EPA's CompTox PFAS library would limit the identifiable space of the method to only those compounds present in the list. Similarly, tools such as HaloSeeker [7] and FluoroMatch [8] would limit the identifiable space to halogenated species and fluorinated species, respectively. Filters, including mass defect filters, can also be used to prioritize (or deprioritize) homologous series of compounds of interest [9], where only compounds in (or outside) the homologous series would be included in the identifiable space. In addition, various thresholds (e.g., retention time, intensity, mass range, elemental composition) may be used to filter data based on data quality and/or method performance. For example, it is common to match masses/molecular formulae or MS/MS spectra to libraries or databases only if they surpass a defined signal intensity threshold to ensure sufficient ion statistics for isotopic fit or MS/MS spectral matching. Likewise, the optimal performance of a specific tool may indicate an advisable threshold, such as limiting the retention time range of considered analytes based on the variable accuracy of a retention time prediction tool across a chromatogram. Ideally, through a series of inquiries/filtering steps, the ChemSpaceTool described herein would take user-defined inputs in the form of chemical lists that are considered for each of the tools or filtering parameters used (e.g., chemicals in the libraries/databases used, possible elemental compositions, etc.), and provide a summary of the overall identifiable chemical space and molecular descriptors used to position each compound in multidimensional space. An additional feature of the proposed ChemSpaceTool would allow the user to upload a list of putative identifications determined by their workflow (non-targeted identified compounds) for comparison to the summarized identifiable chemical space.

Anticipated applications

A primary advantage of NTA is the ability to cast a wide net for chemical detection and identification. Fully understanding the chemical space (scope) of any NTA method/study remains challenging; researchers may optimize various parameters to improve the extractability and detectability of a handful of surrogate compounds only to inadvertently reduce that of other compounds. The ChemSpaceTool would strive to establish the chemical space boundary based on these varied and highly influential method steps. By understanding this boundary, researchers would be able to make prospective adjustments to method procedures to capture more of, or a different region of, chemical space. Retrospectively, being aware of the applicability domain of a project provides important context for results and aids in accurately evaluating NTA performance.

Knowledge of the detectable space can be leveraged in two primary ways. First, researchers can identify their detectable ACL prior to performing any annotation and identification efforts, then use that list as their primary annotation and/or identification library. Using a highly curated, project-specific library can significantly reduce false detection rates (e.g., compounds that are not ionizable on the analytical platform should not be matched to prioritized features). Furthermore, decoy compounds can help users define thresholds for inclusion in their libraries by setting the plausibility or match scoring thresholds to just above that of the decoys. Decoy compounds, or other "fringe" compounds identified, can be retrospectively reported in the tool to further aid in improving the tool in subsequent releases. Secondly, the detectable ACL can be used retrospectively to cross reference non-targeted annotations using conventional annotation/identification libraries and techniques, where there is greater confidence that matched compounds are true positives because the method could detect such compounds. Both methods of implementation come with advantages and disadvantages, but both increase the confidence of non-targeted identifications.

Prospective use

Breaking down NTA methods into discrete steps, as outlined and promoted by the SRT, is also effective for filtering down the chemical universe into plausible chemical space [2]. In keeping with the goals of the SRT, defining the chemical space boundary further supports method transparency and more direct method comparability among studies. For example, if permethrin was reported in a study on household dust, and a researcher is interested in identifying a suite of pyrethroid pesticides in dust, understanding the chemical space covered by the existing method allows the researcher to evaluate whether those methods are appropriate for their analysis. Communication of method intricacies and chemical space coverage would allow existing methods to be readily reused, circumventing or significantly shortening lengthy method development steps.

In some cases, NTA researchers want to detect and/or identify as many chemicals of interest as possible, but no single method is capable of extracting, detecting, and identifying *all* chemicals. Some researchers may strive to capture as many types of chemical classes as possible; others may focus on compounds associated with a specific toxicological endpoint (e.g., estrogenicity), or try to ensure maximal coverage of compounds within a single class (e.g., PFAS). The ChemSpaceTool may aid researchers in understanding the influence of their study design, data acquisition, and data processing and analysis methods on the chemical space covered. Deconstructing each of these steps can play a key role in method development and allow researchers to adjust their methods in ways that expand upon or refine their chemical space prior to analyzing samples.

For example, if a researcher is investigating chemical classes in drinking water that may activate selected bioassays, it is important to develop sample processing methods that capture the broad array of bioassay-active compound types. These may include polar compounds like hormones, pesticides, antiseptics, per- and polyfluorinated alkyl substances (PFAS), and bisphenols, in addition to semi- and nonpolar compounds like polychlorinated biphenyls (PCBs), dioxins, and polycyclic aromatic hydrocarbons (PAHs). With these

compound classes as suspects, the researcher can build a method with an applicability domain that encompasses them and expands along influential molecular descriptor ranges to increase the coverage of project-appropriate chemical space. Sample preparation methods for the project may also be impacted; for example, common solid-phase extraction media like hydrophilic-lipophilic balance (HLB) cartridges, with a single elution and acquisition platform, may not provide sufficient extraction capabilities for the compound classes of interest. To expand the chemical space captured by this method, other SPE media, elution buffers, and analytical platforms can be explored. Upon reviewing the chemical space covered by each of the three workflows outlined in Fig. 4 (A), a mixed-mode cartridge with HLB *and* anion exchange resins, multi-step elutions with polar and nonpolar solvents with acidic buffers, and dual-platform data acquisition (LC and GC) may be required to encompass the desired chemical space coverage. Compounds within the suspect chemical classes can be used to supplement the ChemSpace QC mixtures to further validate the chemical space coverage suggested by the method-based filtering models and better describe the limitations of the selected method.

Once the chemical space boundary has been generally defined for a method, the ACLs provided by the ChemSpaceTool can be used as a molecular formula and annotation library. Having a highly curated and project-appropriate screening list can notably reduce false-positive rates by eliminating the compounds unamenable to the method that may be matched otherwise. Although screening with the ACL during molecular formula matching and structure elucidation steps has advantages, it should also be noted that with complex modeling such as this, there is a margin of error that could lead to false negatives in some instances.

Retrospective use

The ChemSpaceTool may prove to be equally as beneficial in a retrospective sense as it pertains to annotation and performance evaluation. Often in NTA, the number of plausible annotations far exceeds the number of features requiring identification. Frequently, researchers use filtering tools to eliminate unlikely or implausible structures based on retention time, platform amenability, etc. before attempting to annotate a feature. The ACL offers the opportunity to further eliminate implausible structures in tandem with other filtering steps, thus increasing the confidence in annotation (Fig. 5 (A)). Leveraging ACLs in a retrospective sense at the point of annotation may prove to be a valuable tool in prioritizing plausible structures that fit within the defined chemical space relative to those that are defined as outside of a methods' applicability domain.

Implementing the ChemSpaceTool in a post hoc example can allow for third-party evaluation of vastly different methods by adjusting the evaluation based on chemical space coverage of the individual methods. For example, normalizing sensitivity based on the different chemical spaces covered by each lab's methods results in more comparable sensitivity (Fig. 5 (B)). Importantly, the ChemSpace ACL can be used to define the size of the applicable chemical space versus much larger databases that are conventionally used. Typically, compounds in databases that are *not* identified are considered true negatives (TN); however, upon implementation of the ChemSpaceTool, the distinction can be made whether

a compound was not detected because it was not present in the sample, or because it was likely not identifiable with the method.

Performance assessment of NTA has historically been challenging. When the chemical space of a method is not defined, both compound annotation and communication of results remain hindered. The ChemSpaceTool offers a distinct opportunity to improve feature annotation and increase confidence and communication in performance assessment of NTA.

Outlook

The expected benefits of the ChemSpaceTool include streamlining method development by allowing researchers to predict chemical space coverage, improving annotation prioritization and overall accuracy, enhancing method transferability, and providing context for methods and results. In addition to transparent and detailed reporting of all workflow steps, chemical space delineation would allow researchers to compare results on an inter-laboratory or inter-project basis, and also allow for more confident adaptation of existing methods to new projects. Perhaps most importantly, understanding chemical space provides important context for results, thus allowing researchers and readers to discern whether negative detections correspond to compounds that are likely not present in a sample or that are not amenable to the method.

While the framework for *what* this tool would encompass has been developed, there are many details regarding *how* this tool would be built and perform that are still evolving and/or would be required before it is fully realized. Leveraging chemometric tools like Lowe et al.'s LC amenability tool [10], Liigand et al.'s ionization prediction model [11], and Nuñez et al.'s multidimensional chemical mapping [12] puts development of this tool in motion, but additional models are required for the various filtering steps outlined here. Namely, models to categorize the likelihood of compound presence in various matrix types and their extractability under different pHs, extraction solvents, extraction and cleanup media, and elution buffers are missing. Much is understood in terms of liquid versus gas chromatography amenability, and tools currently exist for ionization prediction in electrospray modes, but tools to predict ionization using different technologies (i.e., atmospheric pressure photoionization (APPI), or atmospheric pressure chemical ionization (APCI)) are needed. As this tool continues to take shape and enters the beginning stages of building and testing, we ask that chemometrics experts and chemists alike step forward to help fill these model gaps. Modeling and datasets that are needed are available at www.nontargetedanalysis.com/chemspacetool in addition to existing chemometric tools that can be used to discuss chemical space in NTA reporting. The authors encourage researchers working in these topic areas to contact us through the website if their work can contribute to any of these specific areas or in the general advancement of this tool.

Acknowledgements

This work was partially supported by the PNNL Laboratory Directed Research and Development program, the m/q Initiative. PNNL is operated by Battelle for the DOE under contract DE-AC05-76RL01830. The participation of Yong-Lai Feng is on behalf of the Government of Canada. The findings and conclusions in this paper have not been formally disseminated by the Food and Drug Administration and should not be construed to represent any agency determination or policy. The mention of commercial products, their sources, or their use in connection with material reported herein is not to be construed as either an actual or implied endorsement of such products by the

Department of Health and Human Services. The views expressed in this article are those of the authors and do not necessarily represent the views or policies of the US Environmental Protection Agency.

References

1. Place BJ, Ulrich EM, Challis JK, Chao A, Du B, Favela K, et al. An introduction to the benchmarking and publications for non-targeted analysis working group. Anal Chem. 2021;93(49):16289–96. [PubMed: 34842413]
2. Peter KT, Phillips AL, Knolhoff AM, Gardinali PR, Manzano CA, Miller KE, et al. Nontargeted analysis study reporting tool: a framework to improve research transparency and reproducibility. Anal Chem [Internet]. 2021;93:13870–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/34618419. [PubMed: 34618419]
3. Reymond JL, Ruddigkeit L, Blum L, van Deursen R. The enumeration of chemical space. Wiley Interdiscip Rev Comput Mol Sci. 2012;2(5):717–33.
4. Knolhoff AM, Premo JH, Fisher CM. A proposed quality control standard mixture and its uses for evaluating nontargeted and suspect screening LC/HR-MS method performance. Anal Chem. 2021;93(3):1596–603. [PubMed: 33274925]
5. Pechnick RN. Comparison of the effects of dextromethorphan, dextrorphan, and levorphanol on the hypothalamo-pituitary-adrenal axis. J Pharmacol Exp Ther. 2004;309(2):515–22. [PubMed: 14742749]
6. Black G, He G, Denison M, Young T. Using estrogenic activity and nontargeted chemical analysis to identify contaminants in sewage sludge. Environ Sci Technol. 55(10):6729–39. [PubMed: 33909413]
7. Léon A, Cariou R, Hutinet S, Hurel J, Guitton Y, Tixier C, et al. HaloSeeker 1.0: a user-friendly software to highlight halogenated chemicals in nontargeted high-resolution mass spectrometry data sets. Anal Chem. 2019;91(5):3500–7. [PubMed: 30758179]
8. Koelmel JP, Stelben P, McDonough CA, Dukes DA, Aristizabal-Henao JJ, Nason SL, et al. FluoroMatch 2.0—making automated and comprehensive non-targeted PFAS annotation a reality. Anal Bioanal Chem. 2022;414(3):1201–15. [PubMed: 34014358]
9. Loos M, Singer H. Nontargeted homologue series extraction from hyphenated high resolution mass spectrometry data. J Cheminform. 2017;9(1):1–11. [PubMed: 28316652]
10. Lowe CN, Isaacs KK, McEachran A, Grulke CM, Sobus JR, Ulrich EM, et al. Predicting compound amenability with liquid chromatography-mass spectrometry to improve non-targeted analysis. Anal Bioanal Chem [Internet]. 2021;413(30):7495–508. Available from: 10.1007/s00216-021-03713-w. [PubMed: 34648052]
11. Liigand J, Wang T, Kellogg J, Smedsgaard J, Cech N, Kruve A. Quantification for non-targeted LC/MS screening without standard substances. Sci Rep. 2020;10(1):1–10. [PubMed: 31913322]
12. Nuñez JR, McGrady M, Yesiltepe Y, Renslow RS, Metz TO. Chespa: streamlining expansive chemical space evaluation of molecular sets. J Chem Inf Model. 2020;60(12):6251–7. [PubMed: 33283505]

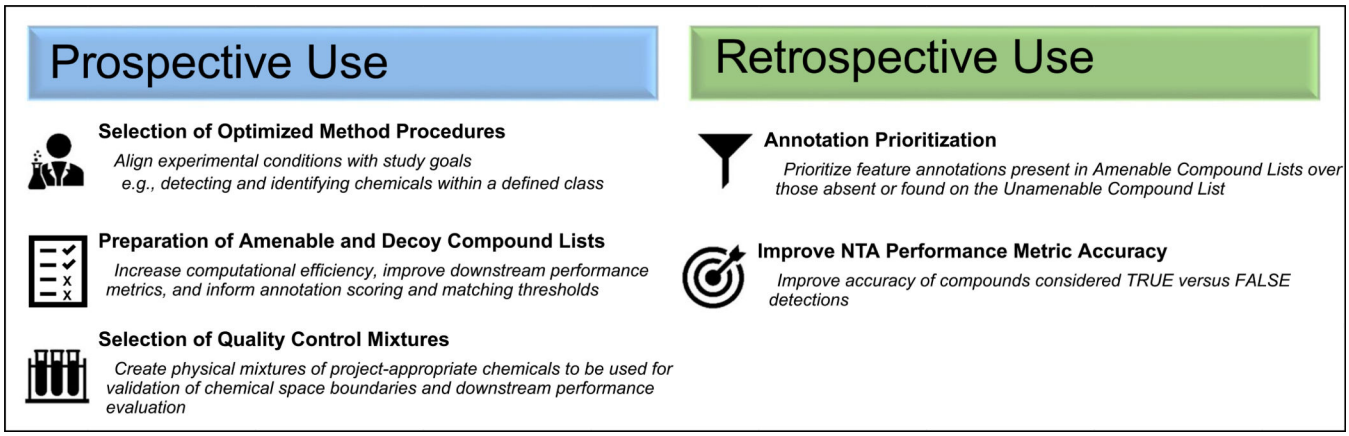

Fig. 1.

The ChemSpaceTool can be used prospectively and/or retrospectively in NTA. Prospective use refers to implementation before or during non-targeted analyses by using predicted chemical space analysis to inform sample preparation and data acquisition methods used and/or using amenable compound lists (ACL) as annotation libraries. Retrospective use refers to implementation of facets of the tool

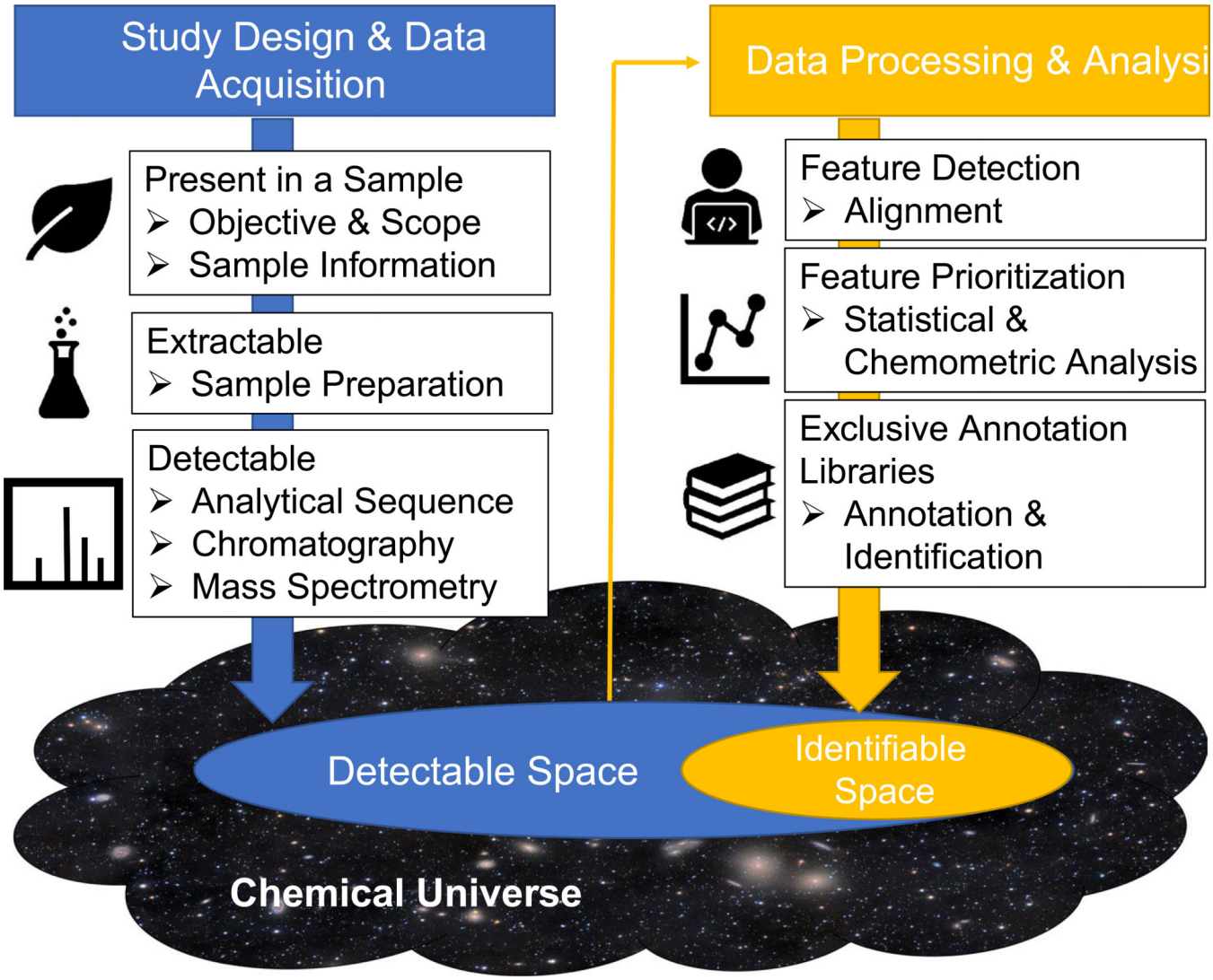

Fig. 2.

Division (not to scale) of the detectable space (blue shaded) and the identifiable space (yellow shaded). Many of these steps closely align with the framework built by the Study Reporting Tool [2]. The detectable space is informed by the ability of a compound to be present in a sample, extractable by analytical parameters. The identifiable space is informed by data processing workflows starting with feature detection, alignment, and binning parameters; various statistical or chemometric tests; exclusive annotation libraries (those that require matches for additional identification efforts); and expert knowledge

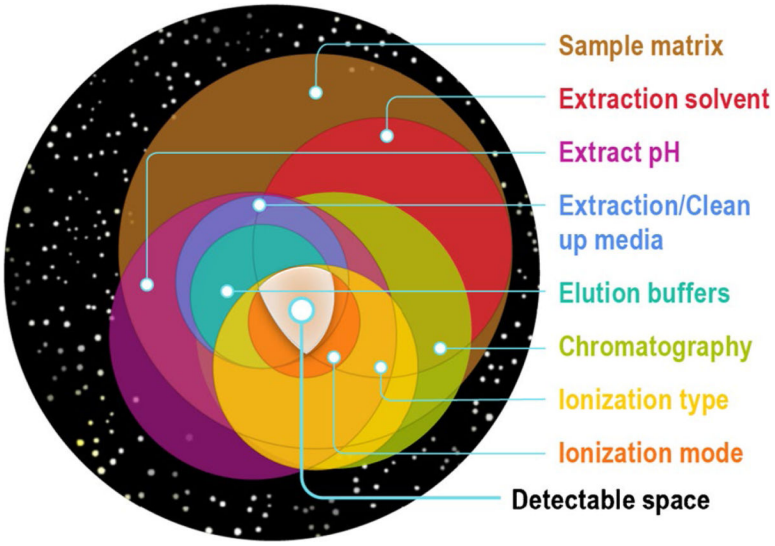

Fig. 3.

Illustration of the intersection of individual filtering steps to define the detectable space of a particular analysis (white area) out of the chemical universe (largest, black circle)

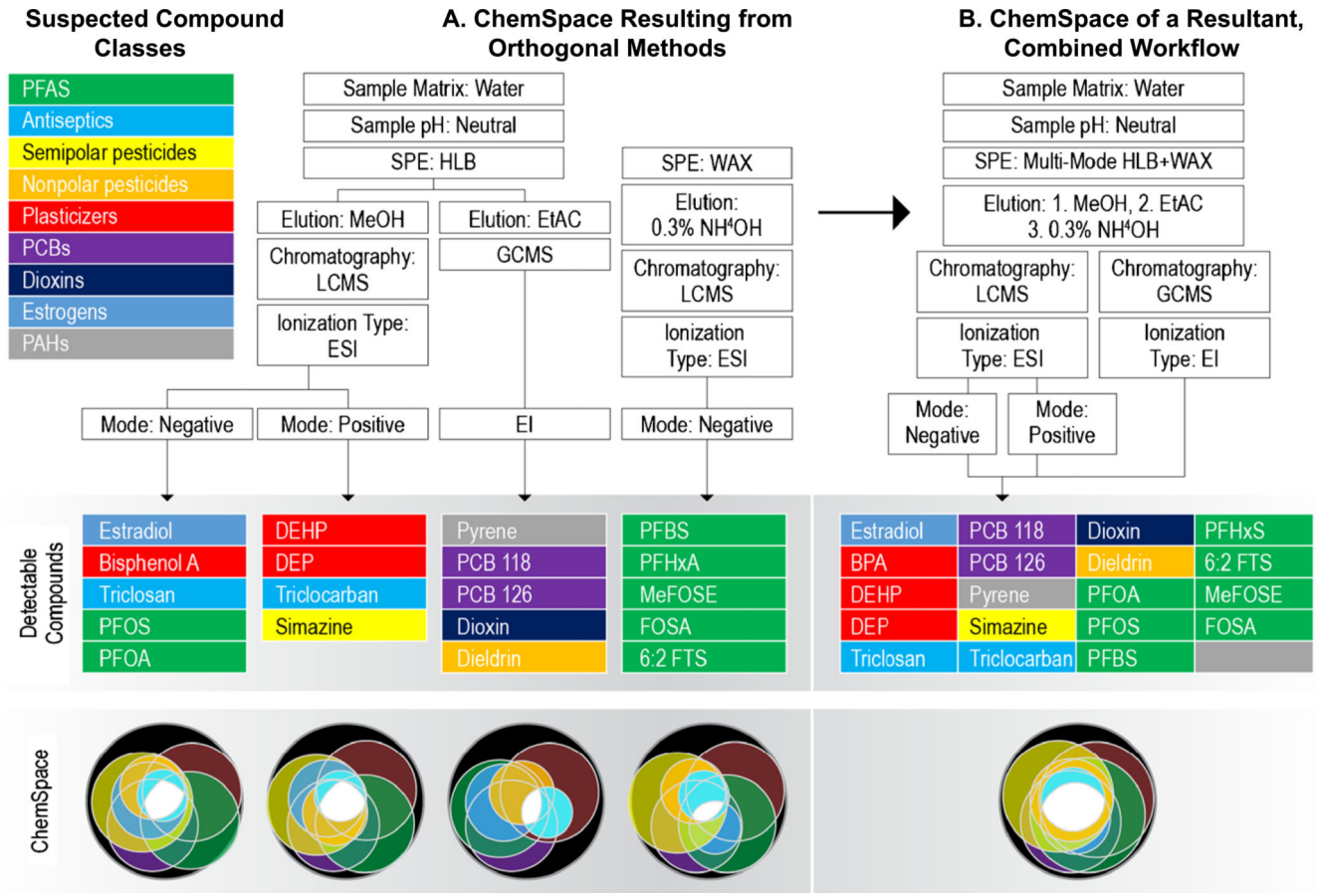

Fig. 4.

Prior to choosing a sample preparation workflow, researchers can use predictive models in the proposed ChemSpaceTool to evaluate the chemical space coverage of different methods under consideration. (A) An example of orthogonal sample preparation and data acquisition steps that provide coverage of all compound classes of interest in four separate workflows; (B) a combined workflow of the three methods in (A) to capture the same types of chemical classes in a single, more comprehensive workflow

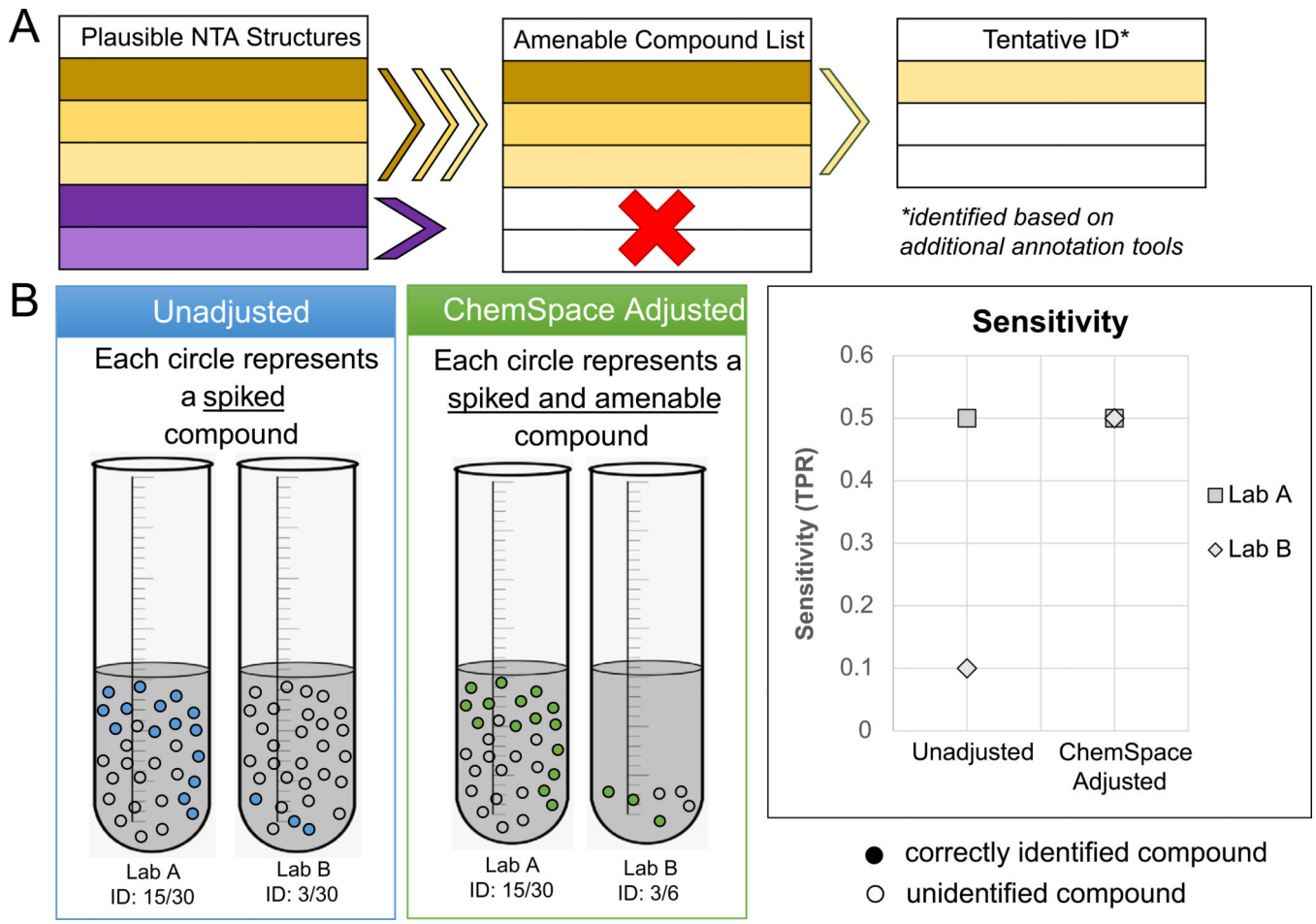

Fig. 5.

Examples of retrospective uses. (A) Amenable compound lists (ACLs) can be used to prioritize plausible structures found via conventional NTA workflows to those most likely to be amenable by a method. (B) Inter-laboratory comparison of sensitivity (and other performance metrics) can be "normalized" by evaluating reported results in the context of chemical space. In this example, lab B's chemical space covers only 3 of 30 spiked compounds. The true-positive rate of their detections is 10% when left unadjusted for chemical space, but increases to 50% when chemical space coverage is considered. When considering amenable chemical space, the sensitivities of the two labs are comparable