Data-Driven Compound Identification in Atmospheric Mass Spectrometry

*Hilda Sandström, Matti Rissanen, Juho Rousu, and Patrick Rinke**

Aerosol particles found in the atmosphere affect the climate and worsen air quality. To mitigate these adverse impacts, aerosol particle formation and aerosol chemistry in the atmosphere need to be better mapped out and understood. Currently, mass spectrometry is the single most important analytical technique in atmospheric chemistry and is used to track and identify compounds and processes. Large amounts of data are collected in each measurement of current time-of-flight and orbitrap mass spectrometers using modern rapid data acquisition practices. However, compound identification remains a major bottleneck during data analysis due to lacking reference libraries and analysis tools. Data-driven compound identification approaches could alleviate the problem, yet remain rare to non-existent in atmospheric science. In this perspective, the authors review the current state of data-driven compound identification with mass spectrometry in atmospheric science and discuss current challenges and possible future steps toward a digital era for atmospheric mass spectrometry.

1. Introduction

In this perspective article, we review the current state of data-driven mass spectrometry in atmospheric science. We focus on automated compound identification, which refers to the large-scale identification of molecules facilitated by digital tools, open knowledge, and data sharing practices. The past 50 years have seen the emergence of large mass spectral databases, which are filled with mass spectra for a variety of compounds.[1,2] Mass spectral databases are used during compound identification and the development of data-driven identification tools. As a result, many research fields, which rely on high-throughput mass spectrometry, have been able to improve, accelerate, and automate data analysis of mass spectrometry experiments. However, in atmospheric science, we believe that there is room for a broader application and more specific development of such tools. Here, we outline the potential and current barriers for data-driven compound identification in atmospheric mass spectrometry.

Atmospheric science includes the study of all chemical and physical processes that occur in the atmosphere. These processes drive a complex, interlinked system with global impact. The chemical composition of the atmosphere mostly consists of nitrogen and oxygen gas (around 99%), followed by noble gases (about 1%), water vapor (≈ 0.01–4%), and carbon dioxide (0.04%). In addition, the atmospheric gas mixture contains a vast number of trace gases, including methane and carbon monoxide (around 2 ppm and 100 ppb, respectively); inorganic vapors, such as nitrogen and sulfur compounds (e.g., NO, NO_2 and HNO_3, and SO_2, COS, and CS_2); and a substantial number of organic compounds from either biogenic or anthropogenic emissions (e.g., terpenes and polyaromatics). These trace gases all transform in the atmosphere through reactions initiated by sunlight.[3–5]

Trace gases can alter the atmospheric composition at any given time. Certain trace gases are very reactive and have short lifetimes, while others are practically nonreactive and persist for far longer periods, allowing them to transport over long distances. Trace gas emissions of organic compounds enter the atmosphere mainly in reduced and poorly water-soluble forms. Through oxidation, the organic compounds increase their affinity for the condensed phase (see **Figure 1**). This means they can be scavenged by liquid droplets and airborne particles. One example of this complex multi-phase chemistry is secondary organic aerosol particle generation. Secondary organic aerosol particles form via rapid gas-phase oxidation of emitted volatile organic compounds (VOCs) into low-volatile reaction products that can grow atmospheric aerosol particles,[6–8] or form them directly.[9–11] An autoxidation process drives this gas-to-particle conversion by

H. Sandström, P. Rinke
Department of Applied Physics
Aalto University
P.O. Box 11000, FI-00076 Aalto, Espoo, Finland
E-mail: patrick.rinke@aalto.fi

M. Rissanen
Aerosol Physics Laboratory
Tampere University
FI-33720 Tampere, Finland

M. Rissanen
Department of Chemistry
University of Helsinki
P.O. Box 55, A.I. Virtasen aukio 1, FI-00560 Helsinki, Finland

J. Rousu
Department of Computer Science
Aalto University
P.O. Box 11000, FI-00076 Aalto, Espoo, Finland

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/advs.202306235

DOI: 10.1002/advs.202306235

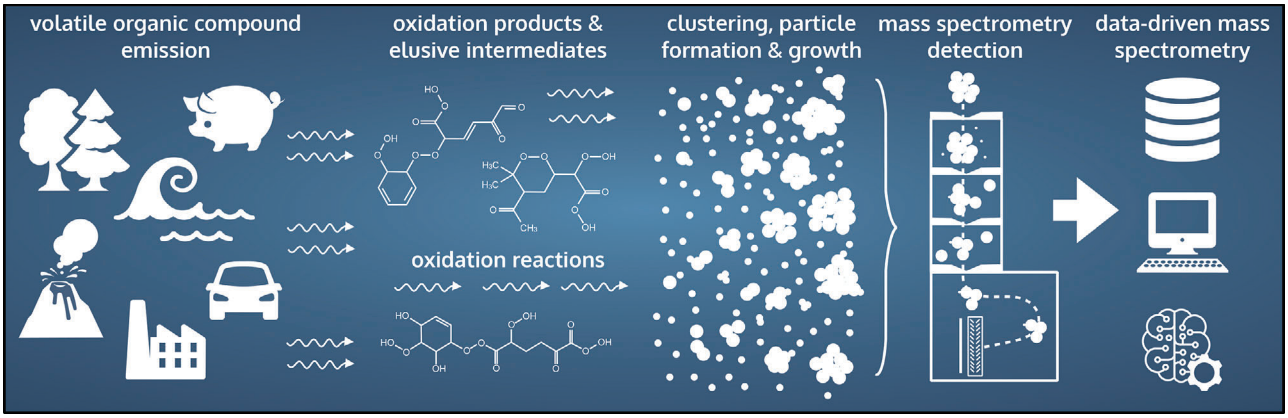
volatile organic compound emission
oxidation products & elusive intermediates
oxidation reactions
clustering, particle formation & growth
mass spectrometry detection
data-driven mass spectrometry

Figure 1. Particles in the atmosphere form through complex processes spanning multiple spatial-scales. First, emissions of volatile compounds enter the atmosphere and oxidize into lower volatility compounds. These low-volatility compounds eventually form clusters which, in turn, can grow into atmospheric nanoparticles. Mass spectrometry has become the measurement method of choice to study atmospheric molecular processes like these. Introducing data-driven methods such as machine learning to the mass spectrometry workflow can help unlock the full analytical potential of mass spectrometry and provide unprecedented insight into atmospheric processes.

generating a sequence of progressively more oxygenated and often isomeric, reaction products from the same parent hydrocarbon.[12,13] With each oxygenation step, the reactant molecules become better at condensing into smaller nanoparticles.[14,15]

The volatility of a compound and its tendency to form atmospheric secondary organic aerosol particles can be described conceptually by the volatility basis set.[14,16,17] The basis set contains information on the vapor concentration and oxygen content (the oxygen to carbon ratio, O:C, or the average carbon oxidation state, OSc) and correlates the volatility evolution with structural changes. The most oxygenated, and generally also the most polar, compounds contribute most to aerosol particle formation and typically have the highest O:C ratios and lowest saturation vapor concentrations. The most extreme cases are the so-called ultra-low volatile organic compounds (ULVOCs) with saturation vapor concentrations lower than 3×10^{-9} µg m^{-3}.[10,14,16,17] At the opposite end of the volatility basis set scale, we find the most volatile, and the least polar, organic compound gases.

The shear number of emitted volatile organic compounds, combined with the many aforementioned reaction schemes, lead to a combinatorial explosion of possible reaction products. The number of different, emitted volatile organic molecules is estimated to lie in the thousands or even millions.[18,19] Through atmospheric reactions, each emitted volatile organic compound multiplies into thousands of reaction products. For example, a decane molecule (10-carbon alkane) with around 100 isomers could already yield over one million distinct compounds.[18]

Understanding the complex atmospheric chemistry behind aerosol particle formation is an important and challenging task. Efforts to map atmospheric compounds and processes contribute to a better basic knowledge of the chemistry in one of Earth's largest and most complex systems. The atmospheric chemistry leading to particle formation also contributes to air pollution and climate change. Aerosol particle pollution has adverse effects on air quality and human health,[20] contributing to 7–9 million premature deaths annually.[21,22] Additionally, aerosol particles impact the climate by reflecting and absorbing solar radiation, an effect addressed in climate models used by the Intergovernmental Panel on Climate Change (IPCC) to inform and guide legislation and action plans for climate change mitigation.[23] In this context, compound identification could, for example, help to develop a better understanding of particle growth, an important factor in determining aerosol–cloud interactions.[24] Small changes in our understanding of aerosol particle growth could alter the number of cloud condensation nuclei by 50% and, thus, affect the outcome of climate models.[14] In this perspective, we propose merging experimental mass spectrometry techniques with data-driven approaches, such as machine learning, to accelerate identification of new atmospheric compounds (see Figure 1).

Atmospheric scientists utilize a combination of laboratory and field-campaign spectrometry experiments to map out the intricacies of atmospheric chemistry leading to particle formation (**Figure 2**). Field-campaigns generate numerous experimental spectra of compound mixtures. Such mixtures often contain unknown compounds and have a composition that varies between measurement sites. Meanwhile, laboratory experiments can, for example, be used to create reference spectra to aid the identification and tracking of atmospheric compounds.[25–28] In a data-driven approach, existing experimental infrastructures would be coupled to data science frameworks. Reference compounds shared in data infrastructures can function as training data for automated compound identification tools. Such digitization of atmospheric mass spectrometry could then expedite compound identification in laboratories and field measurements and help us to gain basic knowledge of the chemistry guiding particle formation (Figure 2).

2. Mass Spectrometry as a Window into Molecular-Level Atmospheric Processes

Much of what is currently known about atmospheric molecular-level processes was obtained with mass spectrometry. While mass spectrometers primarily provide data on the molecular mass and formula, the molecular formula alone often cannot uniquely identify a compound.[29] To gain additional insight into

© 2023 The Authors. Advanced Science published by Wiley-VCH GmbH

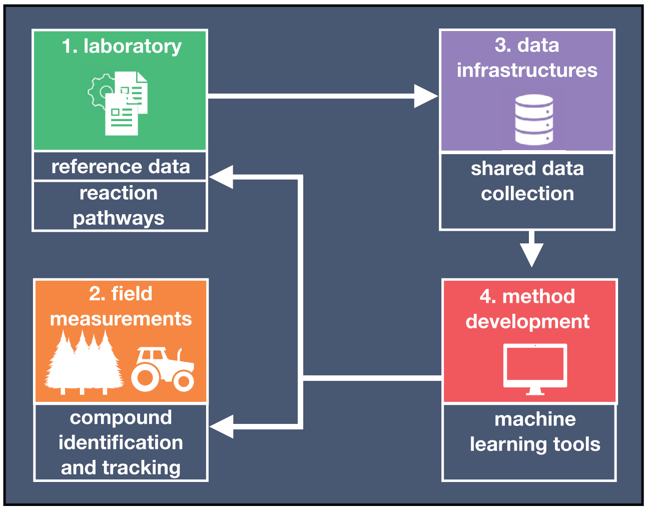

Figure 2. Data-driven compound identification in atmospheric mass spectrometry requires an integration of experiments and data science frameworks. Laboratory experiments can be used to create reference spectra for atmospheric compounds (1). Field measurements produce large amounts of mass spectrometry data of unknown compounds (2). Reference spectra and field measurements can be collected in shared data repositories (3). Data-driven (e.g., machine learning-based) compound identification tools can be trained with reference spectra and be used to identify new compounds measured in field campaigns or laboratories thereby increasing our basic knowledge of atmospheric processes (4).

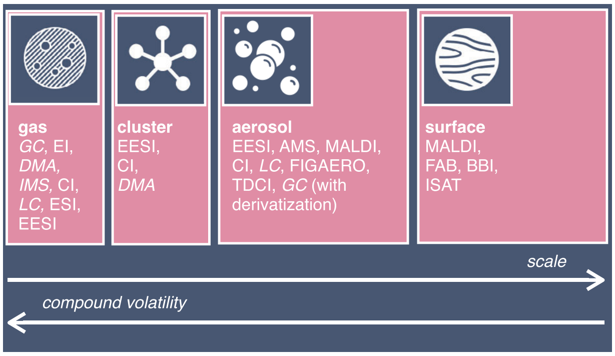
gas
GC, EI, *DMA*, *IMS*, CI, *LC*, ESI, EESI
cluster
EESI, CI, *DMA*
aerosol
EESI, AMS, MALDI, CI, *LC*, FIGAERO, TDCI, *GC* (with derivatization)
surface
MALDI, FAB, BBI, ISAT
scale
compound volatility

Figure 3. Example overview of mass spectrometric techniques and complementary separation techniques (in italicized font), used to study atmospheric compounds ranging from molecules in the gas-phase, clusters to aerosols, and aerosol surfaces. The arrows at the bottom of the figure indicate the inverse relation between measurable scale and detectable volatility. EI, electron ionization; DMA, differential mobility analysis; IMS, ion mobility spectrometry; CI, chemical ionization; ESI, electrospray ionization; EESI, extractive electrospray ionization; AMS, aerosol mass spectrometry; MALDI, matrix-assisted laser desorption ionization; FIGAERO, filter inlet for gas and aerosols; TDCI, thermal desorption chemical ionization; FAB, fast atom bombardment; BBI, bursting bubble ionization; ISAT, interfacial sampling with an acoustic transducer.

molecular structures, mass spectrometry can be combined with techniques such as chromatographic separation,[30] induced fragmentation (MS/MS[31,32] and electron ionization [EI] mass spectrometry[33]) ion mobility spectrometry,[34,35] ionization characteristics,[36–38] and spectroscopy methods.[19] Such combined approaches have the potential to identify compounds and address a wide range of research questions, including those requiring high-throughput analysis. However, the use of mass spectrometry in atmospheric science faces many challenges, which we outline below.

Figure 3 shows examples of mass spectrometric techniques used to study different compounds in atmospheric chemistry.[39] In the introduction, we alluded to the fact that atmospheric chemistry (gas, molecular clusters, and particles) involves compounds with widely different volatility. Since mass spectrometry is inherently a gas-phase detection method, any specimen must first be volatilized. For this purpose, specialized techniques have been developed to study low-volatile molecules with mass spectrometry.

The experimentally resolvable fraction of compounds, in terms of their volatility, has expanded steadily, as techniques have improved.[31,40] For example, large biomolecules have been detected using several spray ionization sources (e.g., electrospray ionization [ESI][41,42] and atmospheric pressure photoionization [APPI]),[43–45] and surface-bound species by desorption techniques such as matrix-assisted laser desorption ionization (MALDI).[31,46] Particulate bound targets, the constituents of nanoparticles, can be detected through direct aerosol sampling by, for example, using an aerodynamic lens with subsequent flash vaporization and EI ionization in aerosol mass spectrometry (AMS),[47] or by collecting the particles onto a filter (or wire) with subsequent rapid thermal desorption vaporization of the condensed-phase constituents. The latter is, for example, applied in chemical ionization mass spectrometry (CIMS)[48,49] detection (with, e.g., filter inlets for gas and aerosols [FIGAERO][50] or thermal desorption multi-scheme chemical ionization inlet [TD-MION][51]).

Of the atmospheric compounds, the volatile gas-phase organic molecules are commonly investigated with either gas-chromatography mass spectrometry (GC-MS)[52] or proton transfer reaction mass spectrometry (PTRMS).[53] The least volatile fraction (corresponding to the lowest gas-phase concentrations) can generally only be measured by atmospheric pressure interface (Api) CIMS methods employing anion attachment.[7,10,54] Finding techniques that are applicable to the whole range of molecular species present in the atmosphere is a major challenge in atmospheric mass spectrometry, and multiple techniques are currently required to cover the whole volatility range (Figure 3).

Besides a broad compound coverage, the ideal mass spectrometric technique in atmospheric science should be able to analyze ambient gas-phase samples directly without the need for sample pre-treatment.[55] However, such techniques are rare and are often limited by, for example, sampling requirements (e.g., limited time resolution resulting from the necessary temporal spacing of compounds as they pass through a chromatographic column), sensitivity, and interference from background compounds (e.g., spectral overlaps in spectroscopic techniques).[56,57] Api-CIMS is popular because it can sample ambient air, usually through a differentially pumped interface (see, e.g., ref. [58]). Samples do not need to be pre-treated, which enables direct, online analysis. While various methods exist for analyzing aerosols in real-time, such as resonance multiphoton ionization[59,60] and secondary electrospray ionization,[61] we will focus here on Api-CIMS due to its user-friendliness, reliability, and robustness. Api-CIMS can operate continuously for months, even in field conditions. Without sample pre-treatment, Api-CIMS can be coupled with other research methodologies, which provide complimentary information, such as ion mobility.[34,35] Api-CIMS is most commonly applied in ambient field measurements and environmental chamber campaigns where it is combined with several other measurement techniques.[7,62–66]

The atmospheric composition at a research site can be monitored for days, weeks, or sometimes even years. These time-consuming field campaigns are characteristic of atmospheric mass spectrometry and set atmospheric science apart from other research fields that use mass spectrometry (e.g., metabolomics or pharmaceutics).[67] Field instruments usually produce relatively long time series for a selected group of target ion signals.[36,37] At the opposite end of the time spectrum, specimen can also be collected on a filter or a filament and then analyzed within a few minutes in an Api-CIMS[38,50,51] enabling high-throughput studies of aerosol particles. While early quadrupole-based Api-CIMS instruments were by necessity only monitoring selected target ions, modern mass spectrometric methods measure the whole mass spectrum continuously.[31] The field measurements are often performed up to a mass resolution of 200 000 (the higher the mass resolution, the smaller the resolvable changes in the target mass), which generates large amounts of data that make data analysis challenging. Currently, only a fraction of compounds in atmospheric mass spectrometry measurements are definitively identified due to the various challenges we will review in the next section.[19] Two possible mass spectrometry approaches exist that are suitable for compound identification following or during field campaigns. For example, compounds collected on-site can be analyzed later in the laboratory with chromatography and fragmentation mass spectrometry.[68–70] Alternatively, current developments for improved compound identification by other mass spectrometry techniques used during field-campaigns are ongoing and outlined below.

Field campaigns often employ soft ionization approaches such as Api-CIMS, which minimize ion fragmentation. In Api-CIMS, reagent ions attach to target molecules (adduction mode), revealing molecular formula information. Details on the molecular structure can be obtained by coupling Api-CIMS with molecular fragmentation techniques (MS/MS).[71] Varying the reagent ion increases sensitivity and selectivity, with detectable target ion concentrations ranging down to 10^{-4} cm^{-3}.[15,54,72,73] New methods, for example, selected ion flow tube mass spectrometry (SIFT-MS) and specialized CIMS,[74] have been developed to improve compound identification by varying the ion–molecule interaction. Noteworthy is the 2019 development of the MION inlet platform,[55] facilitating rapid transitions between ionization modes (e.g., nitrate in anion mode[75] and aminium- or proton-transfer in the cation mode[76]). MION has already increased the number of detectable atmospheric molecules[55,77] and further methodological synergy promises even better compound identification in atmospheric mass spectrometry.[72,78]

Summarizing this section, atmospheric science is in a state of dichotomy. Field campaigns have produced large amounts of data, but these data are not labeled and have not been uploaded to mass spectral databases (see following sections). Moreover, the development of data-driven compound identification tools and the accuracy of the tools after deployment relies on the production and analysis of coherent high-quality reference data.[68–70] The vast atmospheric compound space, the heterogeneity of studies (field vs laboratory), and the multiple mass spectrometric techniques have produced a data landscape that is difficult to navigate. Standardization procedures for data collection, processing, and analysis are still lacking. Combined, these challenges have aggravated compound identification in atmospheric science.

3. Compound Identification with Mass Spectrometry

The identification of unknown compounds and processes is the holy grail of atmospheric mass spectrometry. To identify unknown processes and compounds is challenging, requiring suitable identification techniques and a high-accuracy identification method. Since only a few hundred atmospheric compounds out of potentially millions have been identified in aerosol samples,[68–70] the chemical space of atmospheric compounds remains largely uncharted. We also note that, while compound identification is important for gaining basic knowledge of atmospheric chemistry and for use in particle formation modeling,[79] atmospheric mass spectrometry studies are diverse in type and aim. Some studies do not require compound identification, such as: I) inventorying compounds based on their properties, II) real-time monitoring, or III) monitoring known sources or processes (for a review, see ref. [19]). In these example cases, it can be sufficient to track a molecular or elemental composition, or specific compounds and sources, which are easier objectives than compound identification.

In this perspective, we focus on compound identification. We have identified three factors that most affect the accuracy of compound identification in mass spectrometry that we will present in more detail in the following: the chosen experimental technique, the compound identification method (or tool), and the existence of reference standards.

Mass spectrometry methods are able to identify compounds to a varying degree. In 2015, Nozière et al. introduced the I-factor to quantify the identification accuracy of a mass spectrometry technique in terms of the ability to narrow down the number of plausible candidate structures.[19] In the best case, only one plausible structure is identified and the I-factor is equal to one. If the identification method is not able to discern between isomers of the molecular formula, the I-factor goes up to the number of isomers (two or higher). Uncertainties in the determination of the molecular formula can further increase the I-factor.

Nozière et al. used the I-factor to compare atmospheric mass spectrometric techniques in terms of their compound identification ability.[19] The best I-factors were achieved when two or more techniques, such as chromatography and mass spectrometry, were combined. Fragmentation mass spectrometry methods such as tandem mass spectrometry and EI mass spectrometry, coupled to chromatography methods, reached I-factors of 1–3. The I-factor of soft ionization techniques like CIMS were estimated around 4–40 at the time of publication. The newly developed MION-CIMS method, which uses multiple ion chemistries (see Section 2), has the potential to achieve similarly low I-factors as the combination of two or more techniques given above.[55,56] The data produced by mass spectrometry techniques are used to isolate candidate structures with the help of a compound identification method.

The identification accuracy of compound identification methods and tools varies and is determined by their ability to match a recorded spectrum to a molecular structure. In Section 5, we summarize these tools and their principles. The performance of a compound identification tool is measured by the *Top-k* accuracy. Unlike the I-factor, which quantifies the ability of a mass spectrometry technique to resolve the identity of a compound, the *Top-k* accuracy gives the percentage of instances in which the correct compound is found among the *k* best matching compounds during a compound search. For example, a benchmark study in ref. [80] reported a Top-1 accuracy of 39.4 (and a Top-10 accuracy of 74.8) for their highest-ranking identification tool. This means that the tool identified the correct molecular structure in two out of five cases (Top-1 accuracy of 39.4) and found it among the ten best matches in three fourths of all cases (Top-10 accuracy of 74.8). Here, it should be noted that the absolute numbers are highly dependent on both the data size used in training and the molecular database used to retrieve candidate molecular structures. Moreover, the recorded mass spectrum's quality and type can limit the compound identification method's ability to provide reasonable candidate structure suggestions.

The accuracy of a compound identification tool often depends on the existence of appropriate reference standards, that is, measured mass spectra of compounds, which are either identical or similar to the unknown compound. In the compound identification process, most approaches search for the measured spectrum, or a very similar one, in a database. Even if the identification method does not employ a spectral database search, it has still likely been developed, parameterized, or trained with data from one or more such databases. In atmospheric science, the lack of reference standards is a large barrier for effective compound identification,[15,19,56] which we will return to later in this perspective.

In the digitization of compound identification in atmospheric mass spectrometry, machine learning will naturally play a large role. As we will detail in the next section, machine learning tools are already utilized to automate and improve analysis and processing of mass spectrometry data in other fields (see a recent review in ref. [81]). **Figure 4** illustrates a typical mass spectrometry data acquisition process. In atmospheric mass spectrometry, machine learning is already applied to some, but not all, of the steps outlined in Figure 4. Machine learning models have been trained on different atmospheric mass spectrometry data (like AMS, PTRMS, ESI-mass spectrometry, single particle mass spectrometry, and inductively coupled plasma mass spectrometry) for aerosol classification and source apportionment and[82–90] prediction of composition[91–94] and properties.[95,96] Moreover, a recent review highlighted the role of machine learning in data pre-processing during measurements of volatile organic compounds.[97] Thus, machine learning is being integrated into the data analysis of atmospheric mass spectrometry, but little attention is currently devoted to compound identification. GC-MS machine learning models for molecular formula annotation of atmospheric, halogenated compounds,[98] or for molecular property and quantification factor prediction,[69] are two notable exceptions.

We will next address the reasons for the gap between the perceived demand and utility of smart, high-throughput compound identification tools for atmospheric mass spectrometry and the lack of corresponding availability of such tools. We will also identify the major barriers for introducing compound identification techniques in atmospheric mass spectrometry. A key to both these points are currently available mass spectral databases and their link to the success story of machine learning for compound identification in the field of metabolomics.

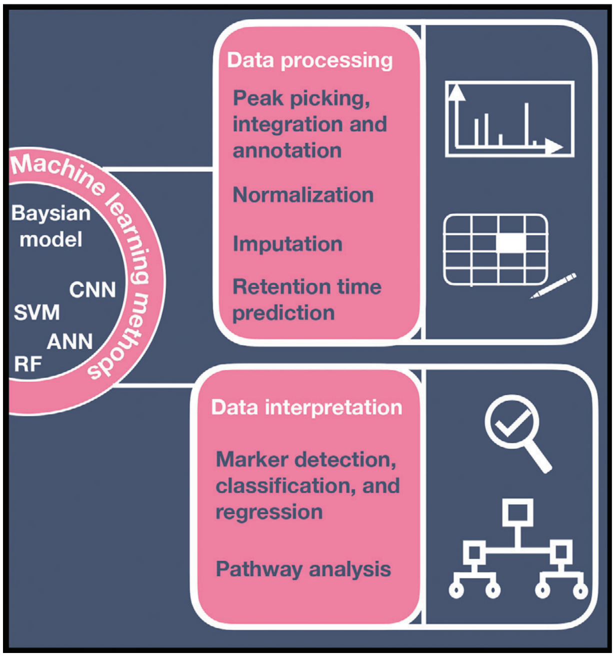

Figure 4. Data processing and analysis steps in a mass spectrometry experiment which have been performed using machine learning methods. Spectral information is extracted through data processing and analysis. Data processing serves to mitigate statistical effects such as batch-to-batch variations, or missing data. Other processing steps include peak processing, alignment, integration, and annotation. Conversely, data analysis aids in the classification or detection of molecules and the identification of chemical pathways to the observed molecules. ANN, artificial neural network; CNN, convolutional neural network; RF, random forest model; SVM, support vector machine.

4. Mass Spectral Databases

Digital mass spectrometry libraries with reference mass spectra, so called mass spectral databases, have been used for compound identification since the 1960s.[1,2] Over time, mass spectral databases have grown in size and usage, partly as a result of increased data processing and storage capabilities as well as adoption of open science practices. **Table 1** summarizes a selection of mass spectral databases that are hosted by research institutions, or distributed by companies and mass spectrometry vendors. The mass spectral data are either collected through research community contributions (e.g., refs. [99–105]), or curation of scientific publications, measurements, and computations (e.g., refs. [106–113]).

By design, mass spectral databases either cover a specific compound space or aim for some level of generality. However, in reality, the data in large mass spectral databases tend to reflect the interest of the primary users and contributors. This is evident in Table 1, which includes specific mass spectral databases created

Table 1. List of select mass spectrometry databases. The list is divided into open access (top) and commercial (bottom). Data volumes reflect the state in August 2023 (the data were taken from an associated webpage or publication). GC, gas chromatography; MS, mass spectrometry; FAB, fast atom bombardment; MS/MS, tandem MS; LC, liquid chromatography; MS^n, tandem mass spectrometry done with *n* fragmentation stages.

Name	Website	Description	Reference
Global Natural Product Social Molecular Networking (GNPS)	gnps.ucsd.edu	The database contains 26 485 unique structures (when full structure is available). The GNPS database contains data contributions from the public and other mass spectral libraries.	[99]
Golm Metabolome Database	gmd.mpimp-golm.mpg.de	Public database maintained by the Max Planck Institute of Molecular Plant Physiology containing 26 590 mass spectra. Has GC-MS spectra for 2222 metabolites and 3651 reference substances.	[100]
Human Metabolome Database (HMDB), v5	hmdb.ca	Freely available database containing experimental and predicted mass spectra. The database has predicted and experimental GC-MS spectra for 74 944 and 3000 compounds, respectively, as well as predicted and experimental LC-MS/MS spectra for 206 809 and 4064 compounds, respectively. HMDB also contains predicted retention times and collision cross sections.	[106]
LipidBank	lipidbank.jp	Curated database containing > 6000 lipids and their spectral information (EI-MS, FAB-MS),	[107]
LipidBlast	fiehnlab.ucdavis.edu	an in silico tandem mass spectral library for lipid identification containing predicted spectra for 119 200 compounds. Provides a tool for users to predict new spectra for their molecules, available in MS-Dial software.	[108]
Lipid Maps Structure Database (LMSD)	lipidmaps.org	LMSD is a database of >48 169 lipid structures, 26 122 of which were determined experimentally and 22 047 of which were generated computationally. LMSD has links to in-house (500 lipid standards) and external (54 877 MS and MS/MS spectra for 7210 lipids from MassBank of North America) mass spectrometry resources.	[101]
MaConDa, v1	maconda.bham.ac.uk	Freely available, manually annotated database of 200 known small molecule contaminants and their LC-MS and GC-MS peaks. Contains un-annotated data. Downloadable and searchable in batch format.	[109]
MassBank (EU), v2023.09	massbank.eu	Public repository of >96 449 mass spectra of $\geqslant$ 15 500 molecules in metabolomics, exposomics, and environmental samples.	[102]
MassBank of North America (MoNA)	mona.fiehnlab.ucdavis.edu	Auto-curated public database with experimental and computational mass spectra of > 650 292 compounds. Includes quality estimation of the mass spectra.	[103]
Advanced Mass Spectral Database (mzCloud)	beta.mzcloud.org	Commerical database maintained by HighChem LLC, Slovakia with manually curated high-resolution LC-MS/MS spectra for 26 417 compounds.	[105]
RIKEN tandem mass spectral database (ReSpect) for phytochemicals	spectra.psc.riken.jp	A curated database with 8649 tandem mass spectra of 3595 plant metabolite compounds collected from scientific literature in 2011 and authentic standards. Has grown since and now contains 9017 (+368) spectra.	[104]
Maurer/Wissenbach/Weber LC-MS^n Library of Drugs, Poisons, and their Metabolites, (2nd edition)	sciencesolutions.wiley.com	LC-MS^n library of over 2270 compounds and over 3600 of their metabolites curated for forensic use.	[112,113]
Metlin Gen2 (Mass consortium)	massconsortium.com	METLIN is a highly curated commercial database with experimental spectra on over 930 000 molecular standards (2023) (LC-MS/MS). All molecular standards were analyzed in positive and negative ionization modes and at four different collision energies (0, 10, 20, and 40 eV).	[114–116]
NIST Tandem and Electron Ionization Mass spectral library, 2023 release	chemdata.nist.gov	Curated spectra of 51501 compounds (tandem) and 347 100 (EI), mainly metabolites, drugs, pesticides, peptides, and lipids. Also contains a retention index database, including predicted values.	[110]
LipidSearch (Thermofisher)	thermofisher.com	Computational database containing in-silico LC-MS and LC-MS/MS spectra for > 1.7 million lipid compounds.	[111]
Wiley Registry of Mass Spectral Data 2023	sciencesolutions.wiley.com	A curated GC-MS library with 873 000 spectra of 741 000 unique compounds with relevance to applications in environmental, forensics/toxicology, metabolomics, pharmaceutical, biotech, food/cosmetics, defense/homeland security, and more.	[117]
Wiley Registry of Tandem Mass Spectral Data - MS for ID	sciencesolutions.wiley.com	A curated LC-MS/MS library with spectra for 1163 compounds including illicit drugs, pharmaceutical compounds, pesticides, and other small bioorganic molecules.	[118]

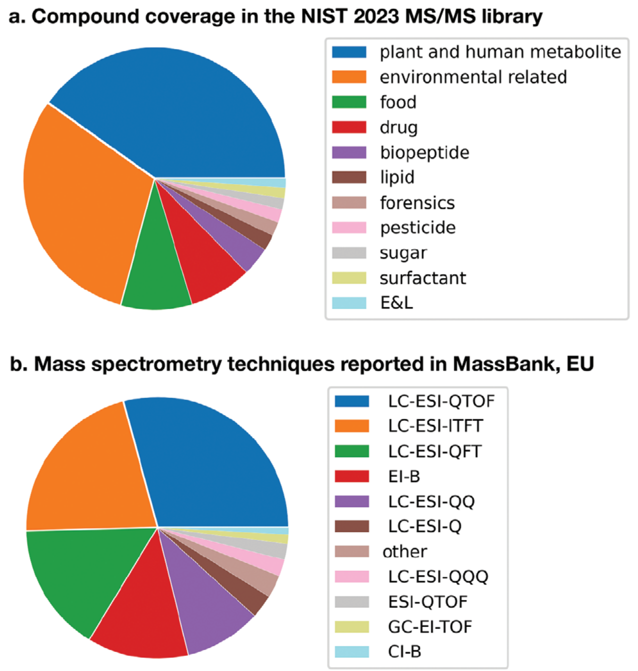

Figure 5. Example of listed contents in mass spectral databases. a) The reported compound coverage of the NIST 23 tandem mass spectral library. b) The different reported mass spectrometric techniques in the European MassBank. These two databases represent general mass spectral databases. E&L, extractables and leachables; CI, chemical ionization; B, bombardment; GC, gas chromatography; EI, electron ionization; TOF, time-of-flight; ESI, electrospray ionization; Q, QQ, QQQ, single, double, triple quadrupole instrument; LC, liquid chromatography; EI-B, electron bombardment ionization; QFT, quadrupole Fourier transform; ITFT, inductively coupled plasma Fourier transform.

for and by the metabolomics community. These databases contain predominantly small molecules called metabolites, found in organisms, cells, or tissues. As in atmospheric science, mass spectrometry is used in metabolomics to identify and quantify molecules of interest. The plethora of mass spectral databases in metabolomics can be attributed to open science initiatives in the research field and the ensuing rapid growth over the past 25 years. As a result, large, general mass spectral databases contain mostly metabolites (see also **Figure** 5a),[102,103,110] despite no stated limitation or constraints on the compound coverage. For this reason, we have decided to highlight metabolomics in this perspective and to use it as a comparative example for developments in atmospheric science. Besides metabolites, other common compound classes in general databases include molecules found in drug or environmental samples (see an overview of NIST 2023 tandem mass spectral library in Figure 5a).

Mass spectral databases provide data collected with a variety of mass spectrometric techniques. As can be seen in Table 1, some databases focus on only one technique, such as LC-MS/MS,[101,104,105,108,111–113,116,118] or GC-MS,[100,107,117] while others provide data from two or more techniques.[99,102,103,106,109,110] The most common technique is LC-MS/MS mass spectrometry followed by GC-MS. For example, the MassBank of North America contains approximately 30 times fewer MS1 spectra (22 500) than tandem mass spectra (including all MS^n) (May, 2023). As expected, these most common mass spectrometric techniques found in mass spectral databases are those that facilitate compound identification (see Section 3).

The number of compounds in the mass spectral databases of Table 1 varies considerably, although a direct comparison of the database size is complicated by the non-standardized way in which the size is reported (e.g., number of ions, number of unique compounds, or number of spectra). The reported data volume of mass spectral libraries either increases continuously or with new versions. The data volumes listed in Table 1 reflect the state in August 2023. LipidSearch by Thermofisher is the largest mass spectral database with spectra for over 1.7 million lipid ions. Massbank of North America is the largest open access database with spectra for over 650 000 compounds. The smallest database reports spectra for only 200 compounds.[109] The median size of all databases reported in Table 1 is 26 485 (average > 290 000). However, the databases overlap in terms of the compounds they cover.[119] The total amount of compounds offered by all databases together is therefore likely less than the sum of their individual compound counts.

Synthetic (i.e., computational) mass spectra have been important for creating large mass spectral databases. Table 1 also lists mass spectral libraries with computationally predicted (so called in silico) tandem mass spectra or GC-MS spectra.[101,103,106,108,111,116] For example, LipidBlast is a purely computational database, which also provides a tool for users to build their own tandem mass spectrometry database.[108] The motivation for generating computational databases, and sometimes combining them with experimental ones, is the need to accelerate data collection. The large number of predicted mass spectra can greatly increase the average mass spectral database size. For example, HMDB contains experimental LC-MS/MS spectra for approximately 4000 compounds, but computational spectra for more than 200 000 compounds. The quality and information content of in silico spectra is, however, still a subject of debate.

The retention time provides useful additional information and is often enough for correct compound annotation in LC- and GC-mass spectrometry. However, for certain isomeric compounds, even the simple chromatographic separation does not provide a positive compound identification and further separation can be necessary.[120] Retention times in GC-MS are collected in MassBanks,[102,103,121] GMD,[100] and NIST23,[110] among others. In addition, computationally predicted retention times are supplied in, for example, HMDB.[106] However, retention times tend to vary significantly between laboratories, which hampers their utility for compound identification. Machine learning techniques can help in alleviating this problem (see Section 5).

Vinaixa and colleagues have reviewed features of mass spectral databases in 2016.[119] They identified beneficial features such as open access, downloadable, large size, curation, data from different platforms, functionality to merge spectra, inclusion of chemical standards, and addition of unknown compounds. On the adverse side, they list commercial licenses, lack of curation and spectrum information, limited sample sources, only negative polarity mode, or only computational data. The review also surmises that there might be a trade-off between too many and too few instrument types as well as collision energies. Following Vinaixa et al., we summarize some features of the mass spectral databases in Tables 1 and 2.

Table 2. Features of the mass spectrometry databases. Open access, partial or full free access to mass spectral data; Data upload, users can contribute with data; Comp. data, contains computationally (in silico) generated mass spectra; Exp. data, experimental mass spectrometry data; collects unknowns, collects and adds unknown spectral queries; machine learning tools, has associated machine learning tools.

	Open access	Data upload	Computational data	Experimental data	Collects unknowns	Machine learning tools
Global Natural Products Social Molecular Networking (GNPS)	✓	✓		✓	✓	✓
Golm Metabolome Database	✓[a)]		[b)]	✓	✓	✓
Human Metabolome Database (HMDB), v5	✓		✓	✓	✓	✓
LipidBank	✓			✓		
LipidBlast	✓	✓[c)]	✓			
Lipid Maps Structure Database (LMSD)	✓	✓	✓	✓		
MaConDa, v1	✓		✓	✓		
MassBank (EU), v2023.09	✓	✓		✓	✓[d)]	
MassBank of North America (MoNA)	✓	✓	✓	✓		
Advanced Mass Spectral Database (mzCloud)	✓			✓		
RIKEN tandem mass spectral database (ReSpect) for phytochemicals	✓			✓		
Maurer/Wissenbach/Weber LC-MS^n Library of Drugs, Poisons, and their Metabolites, (2nd edition)				✓		
Metlin Gen2 (Mass consortium)				✓		
NIST Tandem and Electron Ionization Mass spectral library, 2023 release				✓		
LipidSearch (Thermofisher)		✓	✓			
Wiley Registry of Mass Spectral Data 2023				✓		
Wiley Registry of Tandem Mass Spectral Data—MS for ID				✓		

[a)] For academic and non-commercial use; [b)] Download page contains non-redundant mass spectra that were calculated from available multiple replicate spectra; [c)] Provides a tool to make your own database with computational data; [d)] Stores spectra of compounds tentatively identified.

Mass spectrometry data pipelines and infrastructures are important to further grow mass spectral databases and to facilitate data management, curation, and reproducibility.[122] For example, Pedrioli and colleagues developed the open, vendor-independent data representation `mzXML` in 2004, which enables cross-platform data analysis and management.[123] In addition, a plethora of freely available software has been developed to facilitate mass spectrometry data processing and upload, such as OpenMS,[124] TidyMass,[125] XCMS,[126,127] metaboscape,[128] progenesis,[129] mztab-m,[130] mzMine,[131] and MS-DIAL.[132] Furthermore, the GNPS database offers a feature-based molecular networking tool, which connects feature processing to molecular network modeling.[133]

Another important data management feature mitigates provenance variability. In LC-MS/MS mass spectrometry (as in other soft ionization techniques), data collected at different experimental conditions can vary in appearance. To mitigate such spectral variability, certain database providers have developed the concept of spectral trees[114] and merged spectra[121] that combine spectra collected under different conditions for the same analyte.

5. Compound Identification: Approaches and Software

Compound identification is the primary purpose of mass spectral databases. Traditionally, compounds were identified by searching libraries or databases for matches. With the emergence of digital mass spectral databases, more sophisticated approaches were developed, such as in silico fragmentation,[134–138] fragmentation trees,[125,139–141] and machine learning approaches.[139,142–145]

In the traditional library search, the measured mass spectrum is compared to all spectra in a mass spectral database. The compound is identified (be it correctly or not) as the one with the most similar mass spectrum, out of those in the database. A mass spectral library search is inherently limited by the size of the database, which typically is some orders of magnitude smaller than the target compound space.[146]

State-of-the art compound identification methods also use database information but go significantly beyond library searches. Classical rule-based in silico fragmentation methods rely on a pre-defined set of chemical bond fragmentation rules to predict mass spectra,[146] while combinatorial in silico fragmentation methods search all possible fragmentation paths.[134–137] During compound identification, spectra predictions are made for all entries in a compound database and compared to the measured spectrum to find the best match. In contrast to traditional mass spectral library searches, in silico fragmentation methods search through compound databases (e.g., PubChem) and not through mass spectral libraries. Compound databases cover a larger portion of chemical space than mass spectral databases and are thus less limited in content and size. Rule-based in silico fragmentation methods are limited by the available fragmentation models that rely on heuristic bond energies (measured or estimated), while combinatorial methods generally need to limit

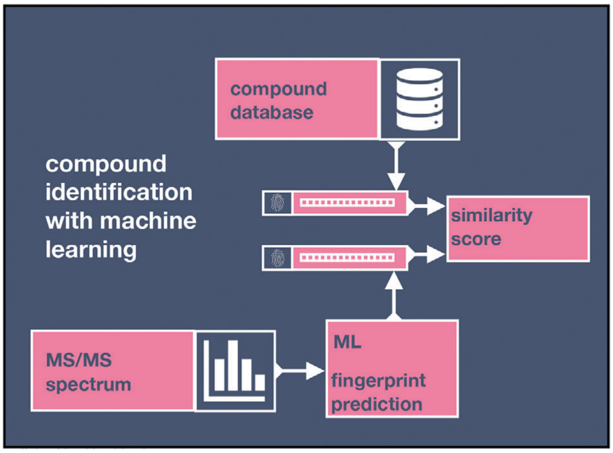

Figure 6. Schematic of the operating principle of most machine learning based compound identification tools. A machine learning model learns to map a mass spectrum to a feature space, here represented by a molecular fingerprint vector. In a second step, the similarity is scored between the predicted fingerprint and the molecular fingerprints of a compound database. ML, machine learning; MS/MS, tandem mass spectrometry.

the amount of fragmentation allowed by the model. In a similar vein, fragmentation tree methods find the optimal fragmentation tree that matches a recorded spectrum. Fragmentation trees are used for de novo molecular formula annotation through Gibbs sampling and Bayesian statistics.[141,147] In in silico fragmentation and fragmentation tree methods, machine learning is not necessarily a component but can be included (e.g., competitive fragmentation modeling [CFM] method).[136–138]

The third category of compound identification algorithms is referred to as machine learning approaches, which are emerging as powerful property and structure inference tools in spectrometry.[148] **Figure 6** illustrates the working principle of most compound identification machine learning algorithms.[139,142–145] In the first step, a mass spectrum is mapped to a feature space represented by a so-called fingerprint. A fingerprint is a vector that encodes the presence or absence of certain molecular features or their counts. Molecular fingerprints can be calculated in different ways from a molecular representation, like a 2D molecular geometry (e.g., refs. [149, 150]). The mapping from spectra to molecular fingerprints requires a reference dataset of spectrum–molecule pairs. Supervised machine learning algorithms are then trained to assign fingerprints to spectra. Examples include kernel methods, such as support vector machines,[142] vector valued kernel ridge regression,[143,151,152] and multiple kernel learning support vector machines,[80,125,139,144] or a combination of deep learning and multiple kernel learning.[145] In the second step, the fingerprint vector is compared to the molecular fingerprints of compounds in compound databases. Moreover, compounds not present in any database can be annotated through hybrid searches.[110,153–155] Additional information channels such as LC retention times,[154,156–159] pairwise retention orders[160] or retention indices[154,156–159] (both relating to the retention order of compounds from LC), or collision cross sections[161] can further improve the identification success. For retention time data, the heterogeneity of data across different laboratories is a hindrance because the retention times depend on the configuration of the chromatograph. Machine learning techniques have been developed to standardize retention times across different laboratories[162] and learn from the relative retention times of molecules,[163,164] which are known to be more invariant across laboratories than absolute retention times.[160]

Open access mass spectral databases containing high-quality reference mass spectra have been essential for the development of machine learning-based compound identification. For example, FingerID,[142] IOKR,[143] Adaptive,[145] CSI:FingerID 1.0,[139] and CSI:FingerID 1.1[80] were all trained using different sets of compounds from different libraries (MassBank, GNPS, MassHunter Forensics/Toxicology PCDL library [Agilent Technologies, Inc.], and NIST17), with sizes ranging from approximately 1200 to 16 083 compounds. The increase in compound identification accuracy during the past decade can largely be attributed to the growth of the spectral databases. In these examples, Agilent Technologies, Inc. and the NIST mass spectral library are the only commercial datasets.

In summary, a variety of approaches and software are now available for compound identification. Open access mass spectral databases have been integral to the development of machine learning approaches and have facilitated the emergence of data-driven mass spectrometry in metabolomics. We will review in the next section how this insight, concepts, tools, and infrastructures can be transferred to atmospheric science.

6. Toward Data-Driven Compound Identification in Atmospheric Mass Spectrometry

In principle, all compound identification approaches we reviewed in this perspective could be directly used in atmospheric science. Suitable training or reference data, however, might be a limiting factor. The identification success rate would strongly depend on the number of atmospheric compounds in available mass spectral databases, or at least on the similarity between these compounds and those in the databases. Furthermore, the preferred mass spectrometric techniques in atmospheric science may differ from those prevalent in current databases. While compound identification algorithms may be able to extrapolate to the chemical space of atmospheric compounds, such generalization would be algorithm dependent and likely incur large uncertainties. We will address these points and propose an action plan to improve data-driven compound identification in atmospheric science. We start off by highlighting general challenges faced in the adoption of mass spectral databases for data-driven compound identification.

6.1. Data Heterogeneity in Mass Spectrometry Databases

The content coverage of current mass spectrometry databases is heterogeneous in terms of compounds, instruments, and experimental procedures. Tool and method developers, therefore, face the challenge of balancing the available data volume, more of which is beneficial for, for example, machine learning methods, against the increased effort of handling the heterogeneity appropriately. Another challenge is the aforementioned coverage overlap, which could introduce biases in data-driven tools derived from more than one database. The current extent of this overlap is, unfortunately, not known, since the last investigation by Vinaixa et al. dates back to 2016.[119] The heterogeneity of available mass spectrometry techniques (see Figure 5b) presents a further challenge but also an opportunity. The characteristics of spectra produced by different mass spectrometry techniques differ, which necessitates dedicated tool and method development. In the long run, however, this technique diversity could be advantageous since different spectrometries could complement each other synergistically. With transfer learning, multivariate machine learning models could be trained to convert between techniques or operate directly on heterogeneous datasets.

In summary, in atmospheric science, much work is still required to assess the utility of existing databases, determine which training data to include in new models, and to establish initial identification tools for atmospherically relevant compounds. Below, we provide a first assessment of the relevance of current mass spectral libraries for data-driven atmospheric mass spectrometry. Investments in improved compound identification for atmospheric science can be justified by the progress achieved in other application domains, such as metabolomics, which have been able to collect experimental data for tens of thousands of compounds (see Section 4).

6.2. Compound Coverage of Atmospheric Molecules

As alluded to in Section 4, atmospheric compounds are currently under-represented in mass spectral databases. Compound identification approaches that were developed for specific database compounds will almost certainly perform worse for atmospheric compounds than for compound classes in the databases. This is true for traditional library searches, which can only identify structures stored in a mass spectral database, as well as for algorithms built with database compounds and spectra.

How well compound identification algorithms perform for atmospheric compounds depends on the overlap of atmospheric compound space with available mass spectral databases. **Figure 7** shows a first visualization of this overlap. The figure presents a t-stochastic neighborhood embedding (t-SNE) analysis for three atmospheric molecular datasets (here referred to as Gecko,[165,166] Wang,[167] and Quinones[168,169]) and two datasets of drug and metabolite compounds, representative of those in mass spectral databases (nablaDFT[170,171] and Massbank of North America[103]). t-SNE clustered the compounds according to the similarity of their (molecular) topological fingerprints.[149,150] Figure 7 shows that the atmospheric compounds cluster closer together and are therefore more similar. Their clusters do, however, not overlap strongly, which indicates that these three datasets cover different parts of atmospheric compound space. The drug and metabolite compounds form their own clusters, most notable is the dense ring of MassBank molecules surrounding the clusters of the other datasets. The two drug and metabolite datasets share some similarity in the inside of the ring, but only the MassBank has some small overlap with the three atmospheric datasets. The implications of Figure 7 are: i) most atmospheric compound classes are absent from mass spectral databases; ii) most atmospheric compounds therefore belong to a chemical space unknown by current compound identification algorithms; iii) the performance of compound identification algorithms in atmospheric science is unpredictable. Three traditional library searches report identification rates of only 2–35% for atmospheric molecules,[68–70] providing further evidence for our three suppositions.

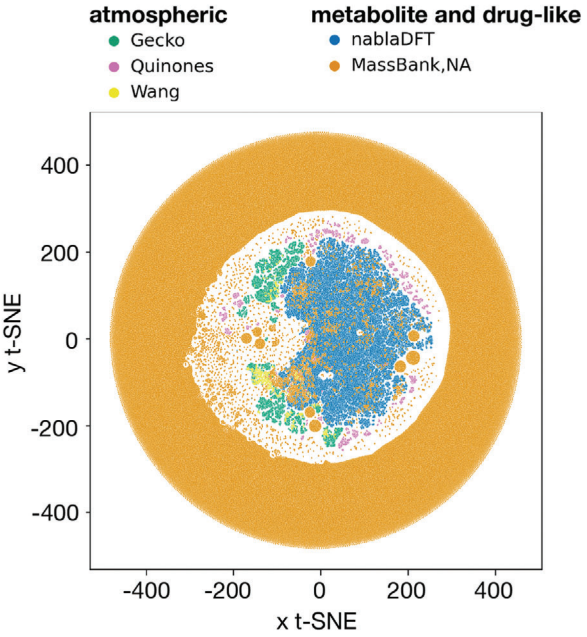

Figure 7. Similarity between molecular datasets containing drug molecules (nablaDFT), metabolites (Massbank of North America), and atmospheric molecules (Gecko, Wang, and quinones) shown through t-SNE clustering. The molecules were compared based on their topological fingerprint.

The fact that atmospheric compounds differ from those in available mass spectral databases implies that compound identification algorithms would have to be able to extrapolate to be applicable in atmospheric science in the short term. Yet, classical rule-based in silico fragmentation algorithms generalize poorly due to built-in rule-sets for chemical bond fragmentation,[146] while in silico fragmentation methods based on combinatorial search (e.g., MetFrag, CFM-ID) are expected to do slightly better. On the other hand, generalization is a common challenge for machine learning models in chemistry.[172] For example, a machine learning model is forced to generalize when it evaluates a new elemental composition,[173] molecular size,[174] or functional group[175] that was not in the training data. Methods for quantifying uncertainty or confidence in a model's prediction have been developed through ensemble methods,[174,176] Bayesian neural networks,[177] Gaussian process regression,[178] support vector machines,[179] and Monte Carlo dropout.[180] In metabolomics, it has been shown that machine learning methods predicting molecular fingerprints from spectra out-perform in silico fragmentation approaches.[80,164] However, it is not known if this also holds true in atmospheric science, where the coverage of the reference spectra of the relevant chemical space is significantly smaller.

Until atmospheric data are available in large enough quantities in mass spectral databases, it would seem prudent to not develop new compound identification methods or workflows immediately for atmospheric science. Machine learning-based approaches, for example, could instead evolve from existing methods developed in other application domains by means of transfer learning. For mass spectrometric techniques commonly found in mass spectral databases, such as tandem mass spectrometry or EI-MS, transfer learning would be particularly well-suited, as already developed models would likely only have to be retrained on atmospheric data. However, for under-represented techniques such as Api-CIMS, transfer learning would not be applicable and new approaches would have to be developed. Api-CIMS applications are currently flourishing in atmospheric science (see Section 2)[34,35,50,51,55,74,76,77,181] but are practically absent from current databases (e.g., less than 0.1% of the European MassBank[102] data, see Figure 5b). If atmospheric science is moving toward data-driven compound identification, this severe lack of data needs to be addressed. In the following, we outline an action plan to fill this data vacuum.

6.3. Action Plan

In this perspective, we reviewed the current challenges of implementing data-driven methods for mass spectrometry in atmospheric science. We next present practical strategies to overcome the identified barriers. Our recommendations are summarized in **Figure 8** and expanded on in the following.

6.3.1. A1—Relevant Data

A paradigm shift toward data-driven mass spectrometry in atmospheric science could begin with access to relevant data (Section 6). For atmospheric mass spectrometry, reference spectra would have to be collected for the compounds taking part in atmospheric chemistry, including the atmospheric gas-phase, small clusters, and nanoparticles (see Section 2). The collection could begin with representative compounds and expand from there. Finding such relevant molecules is no simple feat because the chemical space of atmospheric compounds is large and largely uncharted. We suggest to use data-driven approaches, possibly based on the volatility basis set description of atmospheric compound space (see Section 1), to ensure data collection of compounds with varying properties of interest, such as, for example, volatility and O:C ratio. Data collection should furthermore include the multiple mass spectrometry techniques used in atmospheric science for compatibility with existing databases and compound identification tools, as well as for a holistic description of atmospheric chemistry. It is particularly important to include presently under-represented techniques (e.g., Api-CIMS, as addressed in Section 6.2) to improve their data coverage in the databases. The methodology portfolio could be augmented with synthetic data generated with computational tools as discussed further in A4 below. For example, computational studies in atmospheric chemistry have shown that the binding energy between molecules and reagent ions can be used to predict the experimentally measured CIMS sensitivity (e.g., refs. [72, 78, 182]).

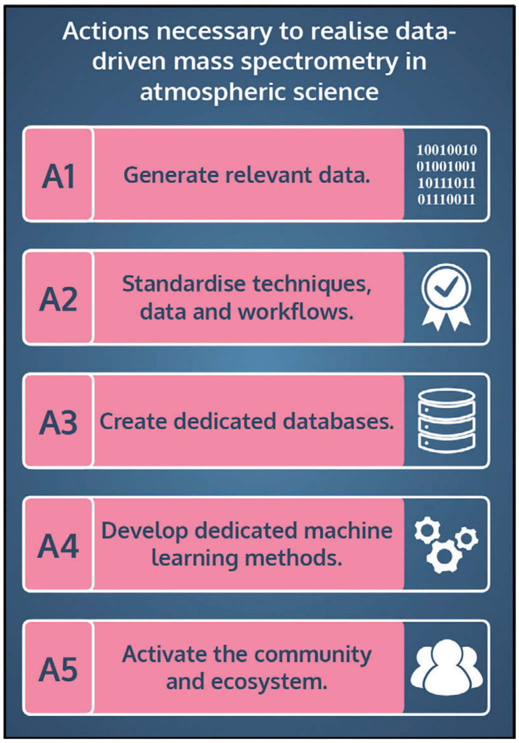
Actions necessary to realise data-driven mass spectrometry in atmospheric science
A1
Generate relevant data.
10010010
01001001
10111011
01110011
A2
Standardise techniques, data and workflows.
A3
Create dedicated databases.
A4
Develop dedicated machine learning methods.
A5
Activate the community and ecosystem.

Figure 8. Our proposed action plan is designed to overcome the challenges hindering a successful implementation of data-driven mass spectrometry in atmospheric science. The plan contains five steps A1–A5.

6.3.2. A2—Standardization

To utilize the collected data in atmospheric science to its full extent, standards and standardized practices for data collection, curation, management, and sharing need to be agreed on and implemented. For certain mass spectrometric techniques (e.g., EI-MS and MS/MS), such practices have already been developed in other fields (e.g., metabolomics, see Section 4) to ensure data standardization and reproducibility (e.g., platform-independent data formats, data analysis pipelines, and spectral trees or merged spectra). They could be directly applied to atmospheric mass spectral data and should be embraced by atmospheric scientists. Conversely, for techniques currently under-represented in mass spectral databases (e.g., Api-CIMS), appropriate standardization practices still need to be developed. Such practices also need to consider the specific use cases in atmospheric science (e.g., the lack of sample pre-treatment and separation by chromatography). For example, Api-CIMS data should be easy to standardize, because the number of different Api-CIMS instruments used in the field has stayed relatively small, with a dominant fraction of the data being acquired by similar methods, such as chemical ionization atmospheric interface time-of-flight (CI-Api-ToF) instrumentation, or the recently introduced orbitrap CIMS systems.[54,58,181,183,184] For Api-CIMS, the standardization of ion production and gas-phase sample introduction is crucial for ensuring fully reproducible measurements. The signal depends on specific ion–molecule reactions and interaction time. Gas-phase chemical ionization is typically linear and scalable, allowing for a wide range of ion concentrations for increased sensitivity. Normalizing measured signals with the number of charge carriers (i.e., reagent ions) is essential in Api-CIMS analysis to account for differences in the initial ion pool. Digital CI-Api-ToF twins can aid in the standardization.[185]

6.3.3. A3—Infrastructure

Data collection and sharing require dedicated infrastructures. En route toward data-driven science, atmospheric science could proceed in two different ways: i) establish dedicated mass spectral databases for atmospheric science data that are operated by the atmospheric science community, or ii) contribute atmospheric science data to existing mass spectral databases. A dedicated database in option (i) offers better control over the data (for example, data curation, labeling, and quality control) but requires concerted actions of key stakeholders and sustained funding.[186] Adopting existing mass spectral databases as in option (ii) is therefore easier in the short term. Contributing to an existing, interdisciplinary mass spectral base promotes data sharing with the broader mass spectrometry community, which expands the user base. We recommend a third option, which is an amalgamation of the two approaches above: curating dedicated databases that can be local to research groups or consortia, but are regularly uploaded and synchronized with large open access databases (such as the MassBanks or GNPS). Dedicated databases could, for example, be linked to collections of reference spectra of atmospheric compounds (e.g., refs. [25–28]). Such collections need to grow to provide access to curated high-quality training data for the data-driven method development. Meanwhile, data from field campaign repositories containing data of unknown compounds can be shared for compound identification. In addition, community datasets, such as refs. [68, 166, 187, 188], could complement data infrastructures. They offer distinct advantages such as having been purposefully curated with design criteria like similarity and balance in mind.

6.3.4. A4—Dedicated Machine Learning Methods

In Sections 5 and 6.2, we reviewed the potential and challenges of available machine learning-based compound identification tools in atmospheric science and observed that the identification performance depends strongly on the availability of relevant data (see A1). For tandem and EI-mass spectrometry, data are available for other compounds, and we propose to begin applying existing machine learning techniques to atmospheric data and to then refine the models accordingly. Over time, such models could be improved through transfer-learning, possibly coupled to active learning schemes, as new atmospheric data become available (Section 6.2). For mass spectrometric techniques, which lack existing machine learning models, but are used for compound identification in atmospheric science (e.g., MION-CIMS), new, dedicated models need to be developed. **Figure** 9 outlines our proposal for a machine learning-based compound identification scheme for MION-CIMS. The CIMS sensitivity for different reagent ions acts as the molecule-specific MION-CIMS fingerprint. The machine learning model learns how to map the MION-CIMS fingerprint to a molecular representation. The development of such a new machine learning-based model could make use of computational mass spectral databases until experimental counterparts become available (see A1). To that end, machine learning could also assist in building computational databases by expediting calculations of the binding energies used to predict CIMS sensitivity.

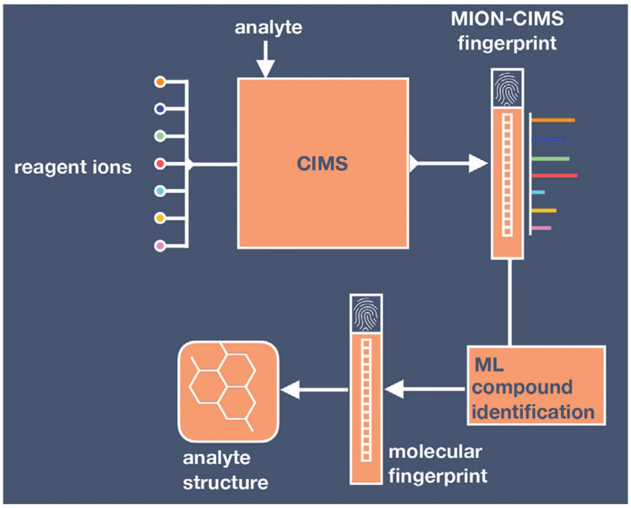

Figure 9. A proposed workflow for machine learning based compound identification with MION-CIMS. The model learns how to map the molecule specific MION-CIMS fingerprint (set of CIMS sensitivity values for different reagent ions) to a molecular representation.

6.3.5. A5—Community Endorsement

Wide-spread adoption of standardized data practices requires a community wide effort. Together, the atmospheric science community needs to commit to open data sharing and publishing. The data should preferably be shared through open access databases, or with FAIR sharing rights,[189] if published with commercial parties. Adoption of community-wide data practices can be encouraged through education in data literacy and machine learning, for example, in summer schools, webinars, or workshops. Further dissemination at atmospheric science conferences and through research networks would create awareness and rally the community to endorse the new paradigm.

7. Take-Home Message

In this perspective, we reviewed the current state and potential for data-driven compound identification in atmospheric mass spectrometry. Although developments of experimental techniques now enable monitoring and tracking of atmospheric chemical processes, an accurate method for high-throughput compound identification is still missing. Community-wide efforts to improve data standardization and collection can support the transition toward reliable identification of atmospheric compounds with mass spectrometry. Integration of data-driven approaches, such as machine learning, into mass spectrometric data analysis will facilitate knowledge gain. Concomitantly, a true paradigm change requires a community endorsement and a combined effort to collect, curate, and share data in a standardized manner.

Although the development of data-driven approaches requires an initial time and resource investment, data-driven approaches promise to be more efficient than the manual processing currently employed. Successful examples in parallel fields can be used to guide and inform this shift toward a digital era in atmospheric mass spectrometry.

Acknowledgements

This project has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme under Grant No. 101002728. The support from the Academy of Finland (334790, 339421, 345802, 346373, 346377, 353836, 358066) is greatly appreciated.

Conflict of Interest

The authors declare no conflict of interest.

Keywords

aerosol, database, machine learning, mass spectrometry, open science

Received: August 31, 2023
Revised: November 4, 2023
Published online:

[1] S. Stein, *Anal. Chem.* **2012**, *84*, 7274.
[2] *Atlas of Mass Spectral Data* (Eds: E. Stenhagen, S. Abrahamsson, F. W. McLafferty), John Wiley & Sons, New York **1969**.
[3] R. P. Wayne, *Chemistry of Atmospheres: An Introduction to the Chemistry of the Atmospheres of Earth, the Planets, and Their Satellites*, 3rd ed., Oxford University Press, Oxford **2000**.
[4] J. H. Seinfeld, S. N. Pandis, *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*, Wiley-Interscience, New York **2006**.
[5] B. J. Finlayson-Pitts, J. N. Pitts, *Chemistry of the Upper and Lower Atmosphere*, Academic Press, San Diego, CA **2000**.
[6] J. H. Kroll, J. H. Seinfeld, *Atmos. Environ.* **2008**, *42*, 3593.
[7] M. Ehn, J. A. Thornton, E. Kleist, M. Sipilä, H. Junninen, I. Pullinen, M. Springer, F. Rubach, R. Tillmann, B. Lee, F. Lopez-Hilfiker, S. Andres, I. H. Acir, M. Rissanen, T. Jokinen, S. Schobesberger, J. Kangasluoma, J. Kontkanen, T. Nieminen, T. Kurtén, L. B. Nielsen, S. Jørgensen, H. G. Kjaergaard, M. Canagaratna, M. D. Maso, T. Berndt, T. Petäjä, A. Wahner, V. M. Kerminen, M. Kulmala, et al., *Nature* **2014**, *506*, 476.
[8] N. M. Donahue, J. H. Kroll, S. N. Pandis, A. L. Robinson, *Atmos. Chem. Phys.* **2012**, *12*, 615.
[9] J. Kirkby, J. Duplissy, K. Sengupta, C. Frege, H. Gordon, C. Williamson, M. Heinritzi, M. Simon, C. Yan, J. Almeida, J. Trostl, T. Nieminen, I. K. Ortega, R. Wagner, A. Adamov, A. Amorim, A. K. Bernhammer, F. Bianchi, M. Breitenlechner, S. Brilke, X. Chen, J. Craven, A. Dias, S. Ehrhart, R. C. Flagan, A. Franchin, C. Fuchs, R. Guida, J. Hakala, C. R. Hoyle, et al., *Nature* **2016**, *533*, 521.
[10] M. Simon, L. Dada, M. Heinritzi, W. Scholz, D. Stolzenburg, L. Fischer, A. C. Wagner, A. Kürten, B. Rörup, X. C. He, J. Almeida, R. Baalbaki, A. Baccarini, P. S. Bauer, L. Beck, A. Bergen, F. Bianchi, S. Bräkling, S. Brilke, L. Caudillo, D. Chen, B. Chu, A. Dias, D. C. Draper, J. Duplissy, I. El-Haddad, H. Finkenzeller, C. Frege, L. Gonzalez-Carracedo, H. Gordon, et al., *Atmos. Chem. Phys.* **2020**, *20*, 9183.
[11] C. Rose, Q. Zha, L. Dada, C. Yan, K. Lehtipalo, H. Junninen, S. B. Mazon, T. Jokinen, N. Sarnela, M. Sipilä, T. Petäjä, V. M. Kerminen, F. Bianchi, M. Kulmala, *Sci. Adv.* **2018**, *4*, eaar5218.
[12] J. D. Crounse, L. B. Nielsen, S. Jørgensen, H. G. Kjaergaard, P. O. Wennberg, *J. Phys. Chem. Lett.* **2013**, *4*, 3513.
[13] M. P. Rissanen, T. Kurtén, M. Sipilä, J. A. Thornton, J. Kangasluoma, N. Sarnela, H. Junninen, S. Jørgensen, S. Schallhart, M. K. Kajos, R. Taipale, M. Springer, T. F. Mentel, T. Ruuskanen, T. Petäjä, D. R. Worsnop, H. G. Kjaergaard, M. Ehn, *J. Am. Chem. Soc.* **2014**, *136*, 15596.
[14] J. Tröstl, W. K. Chuang, H. Gordon, M. Heinritzi, C. Yan, U. Molteni, L. Ahlm, C. Frege, F. Bianchi, R. Wagner, M. Simon, K. Lehtipalo, C. Williamson, J. S. Craven, J. Duplissy, A. Adamov, J. Almeida, A. K. Bernhammer, M. Breitenlechner, S. Brilke, A. Dias, S. Ehrhart, R. C. Flagan, A. Franchin, C. Fuchs, R. Guida, M. Gysel, A. Hansel, C. R. Hoyle, T. Jokinen, et al., *Nature* **2016**, *533*, 527.
[15] F. Bianchi, T. Kurtén, M. Riva, C. Mohr, M. P. Rissanen, P. Roldin, T. Berndt, J. D. Crounse, P. O. Wennberg, T. F. Mentel, J. Wildt, H. Junninen, T. Jokinen, M. Kulmala, D. R. Worsnop, J. A. Thornton, N. Donahue, H. G. Kjaergaard, M. Ehn, *Chem. Rev.* **2019**, *119*, 3472.
[16] N. M. Donahue, S. A. Epstein, S. N. Pandis, A. L. Robinson, *Atmos. Chem. Phys.* **2011**, *11*, 3303.
[17] M. Schervish, N. M. Donahue, *Atmos. Chem. Phys.* **2020**, *20*, 1183.
[18] A. H. Goldstein, I. E. Galbally, *Environ. Sci. Technol.* **2007**, *41*, 1514.
[19] B. Nozière, M. Kalberer, M. Claeys, J. Allan, B. D'Anna, S. Decesari, E. Finessi, M. Glasius, I. Grgić, J. F. Hamilton, T. Hoffmann, Y. Iinuma, M. Jaoui, A. Kahnt, C. J. Kampf, I. Kourtchev, W. Maenhaut, N. Marsden, S. Saarikoski, J. Schnelle-Kreis, J. D. Surratt, S. Szidat, R. Szmigielski, A. Wisthaler, *Chem. Rev.* **2015**, *115*, 3919.
[20] A. Pozzer, S. C. Anenberg, S. Dey, A. Haines, J. Lelieveld, S. Chowdhury, *GeoHealth* **2023**, *7*, e2022GH000711.
[21] S. Khomenko, M. Cirach, E. Pereira-Barboza, N. Mueller, J. Barrera-Gómez, D. Rojas-Rueda, K. de Hoogh, G. Hoek, M. Nieuwenhuijsen, *The Lancet Planetary Health* **2021**, *5*, e121.
[22] J. Lelieveld, A. Pozzer, U. Pöschl, M. Fnais, A. Haines, T. Münzel, *Cardiovasc. Res.* **2020**, *116*, 1910.
[23] *Climate Change 2021: The Physical Science Basis, Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (Eds: V. Masson-Delmotte, A. Z. P. Pirani, S. C. S. L. Péan, C. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, B. Zhou), Cambridge University Press, Cambridge **2021**.
[24] P. A. Arias, N. Bellouin, E. Coppola, R. G. Jones, G. Krinner, J. Marotzke, V. Naik, M. Palmer, G.-K. Plattner, J. Rogelj, M. Rojas, J. Sillmann, T. Storelvmo, P. Thorne, B. Trewin, K. Achuta Rao, B. Adhikary, R. Allan, K. Armour, G. Bala, R. Barimalala, S. Berger, J. Canadell, C. Cassou, A. Cherchi, W. Collins, W. D. Collins, S. L. Connors, S. Corti, F. Cruz, et al., *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (Eds: V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, B. Zhou), Cambridge University Press, Cambridge **2021**, pp. 33–144.
[25] M. Claeys, R. Szmigielski, I. Kourtchev, P. V. D. Veken, R. Vermeylen, W. Maenhaut, M. Jaoui, T. E. Kleindienst, M. Lewandowski, J. H. Offenberg, E. O. Edney, *Environ. Sci. Technol.* **2007**, *41*, 1628.
[26] J. Parshintsev, J. Nurmi, I. Kilpeläinen, K. Hartonen, M. Kulmala, M.-L. Riekkola, *Anal. Bioanal. Chem.* **2008**, *390*, 913.
[27] Y. H. Lin, Z. Zhang, K. S. Docherty, H. Zhang, S. H. Budisulistiorini, C. L. Rubitschun, S. L. Shaw, E. M. Knipping, E. S. Edgerton, T. E.

Kleindienst, A. Gold, J. D. Surratt, *Environ. Sci. Technol.* **2012**, *46*, 250.

[28] A. van Eijck, T. Opatz, D. Taraborrelli, R. Sander, T. Hoffmann, *Atmos. Environ.* **2013**, *80*, 122.

[29] E. S. Baker, G. J. Patti, *J. Am. Soc. Mass Spectrom.* **2019**, *30*, 2031.

[30] S. K. Grebe, R. J. Singh, *Clin. Biochem. Rev.* **2011**, *32*, 5.

[31] G. L. Glish, R. W. Vachet, *Nat. Rev. Drug Discovery* **2003**, *2*, 140.

[32] R. G. Sadygov, D. Cociorva, J. R. Yates, *Nat. Methods* **2004**, *1*, 195.

[33] Q. Yang, H. Ji, Z. Xu, Y. Li, P. Wang, J. Sun, X. Fan, H. Zhang, H. Lu, Z. Zhang, *Nat. Commun.* **2023**, *14*, 3722.

[34] A. Skyttä, J. Gao, R. Cai, M. Ehn, L. R. Ahonen, T. Kurten, Z. Wang, M. P. Rissanen, J. Kangasluoma, *J. Phys. Chem. A* **2022**, *126*, 5040.

[35] J. E. Krechmer, M. Groessl, X. Zhang, H. Junninen, P. Massoli, A. T. Lambe, J. R. Kimmel, M. J. Cubison, S. Graf, Y. H. Lin, S. H. Budisulistiorini, H. Zhang, J. D. Surratt, R. Knochenmuss, J. T. Jayne, D. R. Worsnop, J. L. Jimenez, M. R. Canagaratna, *Atmospheric Measurement Techniques* **2016**, *9*, 3245.

[36] C. Rose, M. P. Rissanen, S. Iyer, J. Duplissy, C. Yan, J. B. Nowak, A. Colomb, R. Dupuy, X. C. He, J. Lampilahti, Y. J. Tham, D. Wimmer, J. M. Metzger, P. Tulet, J. Brioude, C. Planche, M. Kulmala, K. Sellegri, *Atmos. Chem. Phys.* **2021**, *21*, 4541.

[37] H. Berresheim, T. Elste, C. Plass-Dülmer, F. L. Eisele, D. J. Tanner, *Int. J. Mass Spectrom.* **2000**, *202*, 91.

[38] J. N. Smith, K. C. Barsantia, H. R. Friedlia, M. Ehnd, M. Kulmala, D. R. Collins, J. H. Scheckman, B. J. Williams, P. H. McMurry, *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 6634.

[39] J. Laskin, A. Laskin, S. A. Nizkorodov, *Anal. Chem.* **2018**, *90*, 166.

[40] S. Tamara, M. A. D. Boer, A. J. Heck, *Chem. Rev.* **2022**, *122*, 7269.

[41] M. Wilm, *Mol. Cell. Proteomics* **2011**, *10*, M111.009407.

[42] J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, C. M. Whitehouse, *Science* **1989**, *246*, 64.

[43] M. Gaudin, L. Imbert, D. Libong, P. Chaminade, A. Brunelle, D. Touboul, O. Laprévote, *J. Am. Soc. Mass Spectrom.* **2012**, *23*, 869.

[44] A. Bagag, A. Giuliani, M. Réfrégiers, F. L. Naour, *Int. J. Mass Spectrom.* **2012**, *328-329*, 23.

[45] S. Sedláčková, M. Hubálek, V. Vrkoslav, M. Blechová, J. Cvačka, *Separations* **2022**, *9*, 42.

[46] X. Fei, G. Wei, K. K. Murray, *Anal. Chem.* **1996**, *68*, 1143.

[47] J. T. Jayne, D. C. Leard, X. Zhang, P. Davidovits, K. A. Smith, C. E. Kolb, D. R. Worsnop, *Aerosol Sci. Technol.* **2000**, *33*, 49.

[48] M. S. Munson, F. H. Field, *J. Am. Chem. Soc.* **1966**, *88*, 2621.

[49] B. Munson, *Anal. Chem.* **1977**, *49*, 772A.

[50] F. D. Lopez-Hilfiker, C. Mohr, M. Ehn, F. Rubach, E. Kleist, J. Wildt, T. F. Mentel, A. Lutz, M. Hallquist, D. Worsnop, J. A. Thornton, *Atmospheric Measurement Techniques* **2014**, *7*, 983.

[51] F. Partovi, J. Mikkilä, S. Iyer, J. Mikkilä, J. Kontro, S. Ojanperä, P. Juuti, J. Kangasluoma, A. Shcherbinin, M. Rissanen, *ACS Omega* **2023**, *8*, 25749.

[52] K. D. Bartle, P. Myers, *TrAC Trends in Analytical Chemistry* **2002**, *21*, 547.

[53] W. Lindinger, A. Hansel, A. Jordan, *Chem. Soc. Rev.* **1998**, *27*, 347.

[54] T. Jokinen, M. Sipilä, H. Junninen, M. Ehn, G. Lönn, J. Hakala, T. Petäjä, R. L. Mauldin, M. Kulmala, D. R. Worsnop, *Atmos. Chem. Phys.* **2012**, *12*, 4117.

[55] M. P. Rissanen, J. Mikkilä, S. Iyer, J. Hakala, *Atmospheric Measurement Techniques* **2019**, *12*, 6635.

[56] M. Rissanen, *J. Phys. Chem. A* **2021**, *125*, 9027.

[57] A. Tiusanen, J. Ruiz-Jimenez, K. Hartonen, S. K. Wiedmer, *Environ. Sci.: Processes Impacts* **2023**, *25*, 1263.

[58] H. Junninen, M. Ehn, Petäjä, L. Luosujärvi, T. Kotiaho, R. Kostiainen, U. Rohner, M. Gonin, K. Fuhrer, M. Kulmala, D. R. Worsnop, *Atmospheric Measurement Techniques* **2010**, *3*, 1039.

[59] T. W. Adam, R. Chirico, M. Clairotte, M. Elsasser, U. Manfredi, G. Martini, M. Sklorz, T. Streibel, M. F. Heringa, P. F. Decarlo, U. Baltensperger, G. D. Santi, A. Krasenbrink, R. Zimmermann, A. S. Prevot, C. Astorga, *Anal. Chem.* **2011**, *83*, 67.

[60] A. Laskin, J. Laskin, S. A. Nizkorodov, *Environ. Chem.* **2012**, *9*, 163.

[61] T. Z. Semren, S. Majeed, M. Fatarova, C. Laszlo, C. Pak, S. Steiner, G. Vidal-De-Miguel, A. Kuczaj, A. Mazurov, M. C. Peitsch, N. V. Ivanov, J. Hoeng, P. A. Guy, *J. Am. Soc. Mass Spectrom.* **2022**, *33*, 2147.

[62] J. Almeida, S. Schobesberger, A. Kürten, I. K. Ortega, O. Kupiainen-Määttä, A. P. Praplan, A. Adamov, A. Amorim, F. Bianchi, M. Breitenlechner, A. David, J. Dommen, N. M. Donahue, A. Downard, E. Dunne, J. Duplissy, S. Ehrhart, R. C. Flagan, A. Franchin, R. Guida, J. Hakala, A. Hansel, M. Heinritzi, H. Henschel, T. Jokinen, H. Junninen, M. Kajos, J. Kangasluoma, H. Keskinen, A. Kupc, et al., *Nature* **2013**, *502*, 359.

[63] M. Sipilä, N. Sarnela, T. Jokinen, H. Henschel, H. Junninen, J. Kontkanen, S. Richters, J. Kangasluoma, A. Franchin, O. Peräkylä, M. P. Rissanen, M. Ehn, H. Vehkamäki, T. Kurten, T. Berndt, T. Petäjä, D. Worsnop, D. Ceburnis, V. M. Kerminen, M. Kulmala, C. O'Dowd, *Nature* **2016**, *537*, 532.

[64] R. L. Mauldin, T. Berndt, M. Sipilä, P. Paasonen, T. Petäjä, S. Kim, T. Kurtén, F. Stratmann, V. M. Kerminen, M. Kulmala, *Nature* **2012**, *488*, 193.

[65] M. Wang, W. Kong, R. Marten, X. C. He, D. Chen, J. Pfeifer, A. Heitto, J. Kontkanen, L. Dada, A. Kürten, T. Yli-Juuti, H. E. Manninen, S. Amanatidis, A. Amorim, R. Baalbaki, A. Baccarini, D. M. Bell, B. Bertozzi, S. Bräkling, S. Brilke, L. C. Murillo, R. Chiu, B. Chu, L. P. D. Menezes, J. Duplissy, H. Finkenzeller, L. G. Carracedo, M. Granzin, R. Guida, A. Hansel, et al., *Nature* **2020**, *581*, 184.

[66] F. L. Eisele, D. J. Tanner, *Journal of Geophysical Research: Atmospheres* **1993**, *98*, 9001.

[67] L. M. Mayr, D. Bojanic, *Curr. Opin. Pharmacol.* **2009**, *9*, 580.

[68] D. R. Worton, M. Decker, G. Isaacman-VanWertz, A. W. Chan, K. R. Wilson, A. H. Goldstein, *Analyst* **2017**, *142*, 2395.

[69] E. B. Franklin, L. D. Yee, B. Aumont, R. J. Weber, P. Grigas, A. H. Goldstein, *Atmospheric Measurement Techniques* **2022**, *15*, 3779.

[70] J. F. Hamilton, P. J. Webb, A. C. Lewis, J. R. Hopkins, S. Smith, P. Davy, *Atmos. Chem. Phys.* **2004**, *4*, 1279.

[71] S. Tomaz, D. Wang, N. Zabalegui, D. Li, H. Lamkaddam, F. Bachmeier, A. Vogel, M. E. Monge, S. Perrier, U. Baltensperger, C. George, M. Rissanen, M. Ehn, I. E. Haddad, M. Riva, *Nat. Commun.* **2021**, *12*, 300.

[72] N. Hyttinen, R. V. Otkjær, S. Iyer, H. G. Kjaergaard, M. P. Rissanen, P. O. Wennberg, T. Kurtén, *J. Phys. Chem. A* **2018**, *122*, 269.

[73] S. Iyer, X. He, N. Hyttinen, T. Kurtén, M. P. Rissanen, *J. Phys. Chem. A* **2017**, *121*, 6778.

[74] P. Brophy, D. K. Farmer, *Atmospheric Measurement Techniques* **2015**, *8*, 2945.

[75] N. Hyttinen, O. Kupiainen-Määttä, M. P. Rissanen, M. Muuronen, M. Ehn, T. Kurtén, *J. Phys. Chem. A* **2015**, *119*, 6339.

[76] T. Berndt, B. Mentler, W. Scholz, L. Fischer, H. Herrmann, M. Kulmala, A. Hansel, *Environ. Sci. Technol.* **2018**, *52*, 11069.

[77] X.-C. He, J. Shen, S. Iyer, P. Juuti, J. Zhang, M. Koirala, M. M. Kytökari, D. R. Worsnop, M. Rissanen, M. Kulmala, N. M. Maier, J. Mikkilä, M. Sipilä, J. Kangasluoma, *Atmos. Meas. Tech.* **2023**, *16*, 4461.

[78] S. Iyer, F. Lopez-Hilfiker, B. H. Lee, J. A. Thornton, T. Kurtén, *J. Phys. Chem. A* **2016**, *120*, 576.

[79] J. Elm, J. Kubečka, V. Besel, M. J. Jääskeläinen, R. Halonen, T. Kurtén, H. Vehkamäki, *J. Aerosol Sci.* **2020**, *149*, 105621.

[80] K. Dührkop, M. Fleischauer, M. Ludwig, A. A. Aksenov, A. V. Melnik, M. Meusel, P. C. Dorrestein, J. Rousu, S. Böcker, *Nat. Methods* **2019**, *16*, 299.

[81] U. W. Liebal, A. N. Phan, M. Sudhakar, K. Raman, L. M. Blank, *Metabolites* **2020**, *10*, 243.

[82] D. J. Phares, K. P. Rhoads, A. S. Wexler, D. B. Kane, M. V. Johnston, *Anal. Chem.* **2001**, *73*, 2338.
[83] D. M. Murphy, A. M. Middlebrook, M. Warshawsky, *Aerosol Sci. Technol.* **2003**, *37*, 382.
[84] M. Äijälä, K. R. Daellenbach, F. Canonaco, L. Heikkinen, H. Junninen, T. Petäjä, M. Kulmala, A. S. H. Prévôt, M. Ehn, *Atmos. Chem. Phys.* **2019**, *19*, 3645.
[85] M. A. Zawadowicz, K. D. Froyd, D. M. Murphy, D. J. Cziczo, *Atmos. Chem. Phys.* **2017**, *17*, 7193.
[86] C. D. Christopoulos, S. Garimella, M. A. Zawadowicz, O. Möhler, D. J. Cziczo, *Atmospheric Measurement Techniques* **2018**, *11*, 5687.
[87] H. L. Lu, Z. M. Su, L. Li, X. Li, *Anal. Chem.* **2022**, *94*, 17861.
[88] F. Wang, H. Yu, Z. Wang, W. Liang, G. Shi, J. Gao, M. Li, Y. Feng, *Sci. Total Environ.* **2021**, *762*, 144095.
[89] G. D. Bland, M. Battifarano, Q. Liu, X. Yang, D. Lu, G. Jiang, G. V. Lowry, *Environ. Sci. Technol. Letters* **2022**, *56*, 2990.
[90] X. Gong, H. Wex, T. Müller, S. Henning, J. Voigtländer, A. Wiedensohler, F. Stratmann, *Atmos. Chem. Phys* **2022**, *22*, 5175.
[91] C. Giri, H. J. Cleaves, M. Meringer, K. Chandru, *Sustainability* **2021**, *13*, 7614.
[92] P. Pande, M. Shrivastava, J. E. Shilling, A. Zelenyuk, Q. Zhang, Q. Chen, N. L. Ng, Y. Zhang, M. Takeuchi, T. Nah, Q. Z. Rasool, Y. Zhang, B. Zhao, Y. Liu, *ACS Earth Space Chem.* **2022**, *6*, 932.
[93] J. J. Y. Zhang, L. Sun, D. Rainham, T. J. B. Dummer, A. J. Wheeler, A. Anastasopolos, M. Gibson, M. Johnson, *Sci. Total Environ.* **2022**, *806*, 150149.
[94] T. Feng, T. Chen, M. Li, J. Chi, H. Tang, T. Zhang, H. Li, *Chemom. Intell. Lab. Syst.* **2022**, *231*, 104691.
[95] J. Ruiz-Jimenez, M. Okuljar, O.-M. Sietiö, G. Demaria, T. Liangsupree, E. Zagatti, J. Aalto, K. Hartonen, J. Heinonsalo, J. Bäck, T. Petäjä, M.-L. Riekkola, *Atmos. Chem. Phys.* **2021**, *21*, 8775.
[96] H. Jiang, J. Li, R. Sun, C. Tian, J. Tang, B. Jiang, Y. Liao, C.-E. Chen, G. Zhang, *Environ. Sci. Technol.* **2021**, *55*, 10268.
[97] Y. Sun, Y.-B. Chen, M.-J. Chu, X.-H. Jiang, Y. Wang, B.-Q. Guo, *Journal of Chinese Mass Spectrometry Society* **2018**, *39*, 513.
[98] M. Guillevic, A. Guillevic, M. K. Vollmer, P. Schlauri, M. Hill, L. Emmenegger, S. Reimann, *Journal of Cheminformatics* **2021**, *13*, 78.
[99] M. Wang, J. J. Carver, V. V. Phelan, L. M. Sanchez, N. Garg, Y. Peng, D. D. Nguyen, J. Watrous, C. A. Kapono, T. Luzzatto-Knaan, C. Porto, A. Bouslimani, A. V. Melnik, M. J. Meehan, W. T. Liu, M. Crüsemann, P. D. Boudreau, E. Esquenazi, M. Sandoval-Calderón, R. D. Kersten, L. A. Pace, R. A. Quinn, K. R. Duncan, C. C. Hsu, D. J. Floros, R. G. Gavilan, K. Kleigrewe, T. Northen, R. J. Dutton, D. Parrot, et al., *Nat. Biotechnol.* **2016**, *34*, 828.
[100] J. Hummel, N. Strehmel, C. Bölling, S. Schmidt, D. Walther, J. Kopka, *Mass Spectral Search and Analysis Using the Golm Metabolome Database*, John Wiley & Sons, New York **2013**.
[101] M. Sud, E. Fahy, D. Cotter, A. Brown, E. A. Dennis, C. K. Glass, A. H. Merrill, R. C. Murphy, C. R. Raetz, D. W. Russell, S. Subramaniam, *Nucleic Acids Res.* **2007**, *35*, D527.
[102] MassBank consortium, MassBank/MassBank-data: Release version 2022.12, https://zenodo.org/record/7436494 (accessed: August 2023).
[103] MassBank of North America, https://mona.fiehnlab.ucdavis.edu/ (accessed: August 2023).
[104] Y. Sawada, R. Nakabayashi, Y. Yamada, M. Suzuki, M. Sato, A. Sakata, K. Akiyama, T. Sakurai, F. Matsuda, T. Aoki, M. Y. Hirai, K. Saito, *Phytochemistry* **2012**, *82*, 38.
[105] Advanced mass spectral database (mzcloud), www.mzcloud.org (accessed: August 2023).
[106] D. S. Wishart, A. Guo, E. Oler, F. Wang, A. Anjum, H. Peters, R. Dizon, Z. Sayeeda, S. Tian, B. L. Lee, M. Berjanskii, R. Mah, M. Yamamoto, J. Jovel, C. Torres-Calzada, M. Hiebert-Giesbrecht, V. W. Lui, D. Varshavi, D. Varshavi, D. Allen, D. Arndt, N. Khetarpal, A. Sivakumaran, K. Harford, S. Sanford, K. Yee, X. Cao, Z. Budinski, J. Liigand, L. Zhang, et al., *Nucleic Acids Res.* **2022**, *50*, D622.
[107] K. Watanabe, E. Yasugi, M. Oshima, *Trends in Glycoscience and Glycotechnology* **2000**, *12*, 175.
[108] T. Kind, K. H. Liu, D. Y. Lee, B. Defelice, J. K. Meissen, O. Fiehn, *Nat. Methods* **2013**, *10*, 755.
[109] R. J. Weber, E. Li, J. Bruty, S. He, M. R. Viant, *Bioinformatics* **2012**, *28*, 2856.
[110] W. E. Wallace, A. S. Moorthy, *Journal of Forensic Sciences* **2023**, *68*, 1484.
[111] R. Taguchi, M. Ishikawa, *J. Chromatogr. A* **2010**, *1217*, 4229.
[112] D. K. Wissenbach, M. R. Meyer, D. Remane, A. A. Weber, H. H. Maurer, *Anal. Bioanal. Chem.* **2011**, *400*, 79.
[113] D. K. Wissenbach, M. R. Meyer, D. Remane, A. A. Philipp, A. A. Weber, H. H. Maurer, *Anal. Bioanal. Chem.* **2011**, *400*, 3481.
[114] C. A. Smith, G. O'Maille, E. J. Want, C. Qin, S. A. Trauger, T. R. Brandon, D. E. Custodio, R. Abagyan, G. Siuzdak, *Therapeutic Drug Monitoring* **2005**, *27*, 747.
[115] C. Guijas, J. R. Montenegro-Burke, X. Domingo-Almenara, A. Palermo, B. Warth, G. Hermann, G. Koellensperger, T. Huan, W. Uritboonthai, A. E. Aisporna, D. W. Wolan, M. E. Spilker, H. P. Benton, G. Siuzdak, *Anal. Chem.* **2018**, *90*, 11.
[116] J. R. Montenegro-Burke, C. Guijas, G. Siuzdak, *Methods in Molecular Biology* **2020**, *2104*, 149.
[117] F. W. McLafferty, *Wiley Registry of Mass Spectral Data*, 12th ed., Wiley, New York **2020**.
[118] H. Oberacher, *Wiley Registry of Tandem Mass Spectral Data: MS for ID*, Wiley, New York **2012**.
[119] M. Vinaixa, E. L. Schymanski, S. Neumann, M. Navarro, R. M. Salek, O. Yanes, *TrAC - Trends in Analytical Chemistry* **2016**, *78*, 23.
[120] M. N. Eckberg, L. E. Arroyo-Mora, D. R. Stoll, A. P. Decaprio, *J. Anal. Toxicol.* **2019**, *43*, 170.
[121] H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M. Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, et al., *J. Mass Spectrom.* **2010**, *45*, 703.
[122] K. M. Mendez, L. Pritchard, S. N. Reinke, D. I. Broadhurst, *Metabolomics* **2019**, *15*, 150.
[123] P. G. Pedrioli, J. K. Eng, R. Hubley, M. Vogelzang, E. W. Deutsch, B. Raught, B. Pratt, E. Nilsson, R. H. Angeletti, R. Apweiler, K. Cheung, C. E. Costello, H. Hermjakob, S. Huang, R. K. Julian, E. Kapp, M. E. McComb, S. G. Oliver, G. Omenn, N. W. Paton, R. Simpson, R. Smith, C. F. Taylor, W. Zhu, R. Aebersold, *Nat. Biotechnol.* **2004**, *22*, 1459.
[124] H. L. Röst, T. Sachsenberg, S. Aiche, C. Bielow, H. Weisser, F. Aicheler, S. Andreotti, H.-C. Ehrlich, P. Gutenbrunner, E. Kenar, X. Liang, S. Nahnsen, L. Nilse, J. Pfeuffer, G. Rosenberger, M. Rurik, U. Schmitt, J. Veit, M. Walzer, D. Wojnar, W. E. Wolski, O. Schilling, J. S. Choudhary, L. Malmström, R. Aebersold, K. Reinert, O. Kohlbacher, *Nat. Methods* **2016**, *13*, 741.
[125] H. Shen, K. Dührkop, S. Böcker, J. Rousu, *Bioinformatics* **2014**, *30*, i157.
[126] C. A. Smith, E. J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, *Anal. Chem.* **2006**, *78*, 779.
[127] R. Tautenhahn, C. Bottcher, S. Neumann, *BMC Bioinformatics* **2008**, *9*, 504.
[128] Metaboscape, https://www.bruker.com/en/products-and-solutions/mass-spectrometry/ms-software/metaboscape.html (accessed: June 2023).
[129] Progenesis QI, https://www.nonlinear.com/progenesis/qi/(accessed: June 2023).

[130] N. Hoffmann, J. Rein, T. Sachsenberg, J. Hartler, K. Haug, G. Mayer, O. Alka, S. Dayalan, J. T. Pearce, P. Rocca-Serra, D. Qi, M. Eisenacher, Y. Perez-Riverol, J. A. Vizcaíno, R. M. Salek, S. Neumann, A. R. Jones, *Anal. Chem.* **2019**, *91*, 3302.
[131] R. Schmid, S. Heuckeroth, A. Korf, A. Smirnov, O. Myers, T. S. Dyrlund, R. Bushuiev, K. J. Murray, N. Hoffmann, M. Lu, A. Sarvepalli, Z. Zhang, M. Fleischauer, K. Dührkop, M. Wesner, S. J. Hoogstra, E. Rudt, O. Mokshyna, C. Brungs, K. Ponomarov, L. Mutabdžija, T. Damiani, C. J. Pudney, M. Earll, P. O. Helmer, T. R. Fallon, T. Schulze, A. Rivas-Ubach, A. Bilbao, H. Richter, et al., *Nat. Biotechnol.* **2023**, *41*, 447.
[132] H. Tsugawa, T. Cajka, T. Kind, Y. Ma, B. Higgins, K. Ikeda, M. Kanazawa, J. Vandergheynst, O. Fiehn, M. Arita, *Nat. Methods* **2015**, *12*, 523.
[133] L. F. Nothias, D. Petras, R. Schmid, K. Dührkop, J. Rainer, A. Sarvepalli, I. Protsyuk, M. Ernst, H. Tsugawa, M. Fleischauer, F. Aicheler, A. A. Aksenov, O. Alka, P. M. Allard, A. Barsch, X. Cachet, A. M. Caraballo-Rodriguez, R. R. D. Silva, T. Dang, N. Garg, J. M. Gauglitz, A. Gurevich, G. Isaac, A. K. Jarmusch, Z. Kameník, K. B. Kang, N. Kessler, I. Koester, A. Korf, A. L. Gouellec, et al., *Nat. Methods* **2020**, *17*, 905.
[134] S. Wolf, S. Schmidt, M. Müller-Hannemann, S. Neumann, *BMC Bioinformatics* **2010**, *11*, 148.
[135] C. Ruttkies, E. L. Schymanski, S. Wolf, J. Hollender, S. Neumann, *Journal of Cheminformatics* **2016**, *8*, 3.
[136] F. Allen, A. Pon, M. Wilson, R. Greiner, D. Wishart, *Nucleic Acids Res.* **2014**, *42*, W94.
[137] F. Allen, R. Greiner, D. Wishart, *Metabolomics* **2015**, *11*, 98.
[138] Y. Djoumbou-Feunang, A. Pon, N. Karu, J. Zheng, C. Li, D. Arndt, M. Gautam, F. Allen, D. S. Wishart, *Metabolites* **2019**, *9*, 72.
[139] K. Dührkop, H. Shen, M. Meusel, J. Rousu, S. Böcker, *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 12580.
[140] S. Böcker, K. Dührkop, *J. Cheminform* **2016**, *8*, 5.
[141] M. Ludwig, M. Fleischauer, K. Dührkop, M. A. Hoffmann, S. Böcker, *Methods in Molecular Biology* **2020**, *2104*, 185.
[142] M. Heinonen, H. Shen, N. Zamboni, J. Rousu, *Bioinformatics* **2012**, *28*, 2333.
[143] C. Brouard, H. Shen, K. Dührkop, F. D'Alché-Buc, S. Böcker, J. Rousu, *Bioinformatics* **2016**, *32*, i28.
[144] D. H. Nguyen, C. H. Nguyen, H. Mamitsuka, *Bioinformatics* **2018**, *34*, i323.
[145] D. H. Nguyen, C. H. Nguyen, H. Mamitsuka, *Bioinformatics* **2019**, *35*, i164.
[146] D. H. Nguyen, C. H. Nguyen, H. Mamitsuka, *Briefings in Bioinformatics* **2019**, *20*, 2028.
[147] M. Ludwig, L. F. Nothias, K. Dührkop, I. Koester, M. Fleischauer, M. A. Hoffmann, D. Petras, F. Vargas, M. Morsy, L. Aluwihare, P. C. Dorrestein, S. Böcker, *Nature Machine Intelligence* **2020**, *2*, 629.
[148] H. J. Kulik, T. Hammerschmidt, J. Schmidt, S. Botti, M. A. L. Marques, M. Boley, M. Scheffler, M. Todorović, P. Rinke, C. Oses, A. Smolyanyuk, S. Curtarolo, A. Tkatchenko, A. P. Bartók, S. Manzhos, M. Ihara, T. Carrington, J. Behler, O. Isayev, M. Veit, A. Grisafi, J. Nigam, M. Ceriotti, K. T. Schütt, J. Westermayr, M. Gastegger, R. J. Maurer, B. Kalita, K. Burke, R. Nagai, et al., *Electro. Struct.* **2022**, *4*, 023004.
[149] C. James, D. Weininger, J. Delany, *Daylight Theory Manual*, Daylight Chemical Information Systems, Inc., Irvine, CA **1995**.
[150] G. Landrum, Open-source cheminformatics, **2006**, www.rdkit.org (accessed: December 2023).
[151] C. Brouard, E. Bach, S. Böcker, J. Rousu, *Proceedings of Machine Learning Research* **2017**, *77*, 407.
[152] C. Brouard, A. Bassé, F. D'alché-Buc, J. Rousu, *Metabolites* **2019**, *9*, 160.
[153] K. Dührkop, L. F. Nothias, M. Fleischauer, R. Reher, M. Ludwig, M. A. Hoffmann, D. Petras, W. H. Gerwick, J. Rousu, P. C. Dorrestein, S. Böcker, *Nat. Biotechnol.* **2020**, *39*, 462.
[154] M. A. Stravs, K. Dührkop, S. Böcker, N. Zamboni, *Nat. Methods* **2022**, *19*, 865.
[155] B. T. Cooper, X. Yan, Y. Simón-Manso, D. V. Tchekhovskoi, Y. A. Mirokhin, S. E. Stein, *Anal. Chem.* **2019**, *91*, 13924.
[156] F. Qiu, Z. Lei, L. W. Sumner, *Anal. Chim. Acta* **2018**, *1037*, 316.
[157] V. K. Gupta, H. Khani, B. Ahmadi-Roudi, S. Mirakhorli, E. Fereyduni, S. Agarwal, *Talanta* **2011**, *83*, 1014.
[158] T. H. Miller, A. Musenga, D. A. Cowan, L. P. Barron, *Anal. Chem.* **2013**, *85*, 10330.
[159] M. Jalali-Heravi, M. H. Fatemi, *J. Chromatogr. A* **2001**, *915*, 177.
[160] M. Witting, S. Böcker, *J. Sep. Sci.* **2020**, *43*, 1746.
[161] P.-L. Plante, E. Francovic-Fontaine, J. C. May, J. A. Mclean, E. S. Baker, F. Laviolette, M. Marchand, J. Corbeil, *Anal. Chem.* **2019**, *91*, 5191.
[162] J. Stanstrup, S. Neumann, U. Vrhovsek, *Anal. Chem.* **2015**, *87*, 9421.
[163] E. Bach, S. Szedmak, C. Brouard, S. Böcker, J. Rousu, *Bioinformatics* **2018**, *34*, i875.
[164] E. Bach, E. L. Schymanski, J. Rousu, *Nature Machine Intelligence* **2022**, *4*, 1224.
[165] G. Isaacman-Vanwertz, B. Aumont, *Atmos. Chem. Phys.* **2021**, *21*, 6541.
[166] V. Besel, M. Todorović, T. Kurtén, P. Rinke, H. Vehkamäki, *Sci. Data* **2023**, *10*, 450.
[167] C. Wang, T. Yuan, S. Wood, K. U. Goss, J. Li, Q. Ying, F. Wania, *Atmos. Chem. Phys.* **2017**, *17*, 7529.
[168] M. Krüger, J. Wilson, M. Wietzoreck, B. A. M. Bandowe, G. Lammel, B. Schmidt, U. Pöschl, T. Berkemeier, C. T. Berkemeier, *Nat. Sci.* **2022**, *2*, e20220016.
[169] D. P. Tabor, R. Gómez-Bombarelli, L. Tong, R. G. Gordon, M. J. Aziz, A. Aspuru-Guzik, *J. Mater. Chem. A* **2019**, *7*, 12833.
[170] K. Khrabrov, I. Shenbin, A. Ryabov, A. Tsypin, A. Telepov, A. Alekseev, A. Grishin, P. Strashnov, P. Zhilyaev, S. Nikolenko, A. Kadurin, *Phys. Chem. Chem. Phys.* **2022**, *24*, 25853.
[171] D. Polykovskiy, A. Zhebrak, B. Sanchez-Lengeling, S. Golovanov, O. Tatanov, S. Belyaev, R. Kurbanov, A. Artamonov, V. Aladinskiy, M. Veselov, A. Kadurin, S. Johansson, H. Chen, S. Nikolenko, A. Aspuru-Guzik, A. Zhavoronkov, *Frontiers in Pharmacology* **2020**, *11*, 1931.
[172] L. H. Mervin, S. Johansson, E. Semenova, K. A. Giblin, O. Engkvist, *Drug Discovery Today* **2021**, *26*, 474.
[173] Y. Hu, J. Musielewicz, Z. W. Ulissi, A. J. Medford, *Machine Learning: Science and Technology* **2022**, *3*, 045028.
[174] A. Ghose, M. Segal, F. Meng, Z. Liang, M. S. Hybertsen, X. Qu, E. Stavitski, S. Yoo, D. Lu, M. R. Carbone, *Phys. Rev. Res.* **2023**, *5*, 013180.
[175] G. Scalia, C. A. Grambow, B. Pernici, Y.-P. Li, W. H. Green, *J. Chem. Inf. Model* **2020**, *60*, 29.
[176] S. Wan, R. C. Sinclair, P. V. Coveney, *Philos. Trans. R. Soc., A* **2021**, *379*.
[177] Y. Zhou, B. Yang, *Journal of Energy Chemistry* **2023**, *81*, 118.
[178] L. Fang, E. Makkonen, M. Todorović, P. Rinke, X. Chen, *J. Chem. Theory Comput* **2021**, *17*.
[179] M. A. Hoffmann, L.-F. Nothias, M. Ludwig, M. Fleischauer, E. C. Gentry, M. Witting, P. C. Dorrestein, K. Dührkop, S. Böcker, *Nat. Biotechnol.* **2022**, *40*, 411.
[180] M. Wen, E. B. Tadmor, *npj Comput. Mater.* **2020**, *6*, 1.
[181] M. Riva, M. Ehn, D. Li, S. Tomaz, F. Bourgain, S. Perrier, C. George, *Anal. Chem.* **2019**, *91*, 9419.
[182] N. Hyttinen, M. P. Rissanen, T. Kurtén, *J. Phys. Chem. A* **2017**, *121*, 2172.

[183] T. Kurtén, T. Petäjä, J. Smith, I. K. Ortega, M. Sipilä, H. Junninen, M. Ehn, H. Vehkamäki, L. Mauldin, D. R. Worsnop, M. Kulmala, *Atmos. Chem. Phys.* **2011**, *11*, 3007.

[184] M. Riva, M. Brüggemann, D. Li, S. Perrier, C. George, H. Herrmann, T. Berndt, *Anal. Chem.* **2020**, *92*, 8142.

[185] M. Passananti, E. Zapadinsky, T. Zanca, J. Kangasluoma, N. Myllys, M. P. Rissanen, T. Kurtén, M. Ehn, M. Attoui, H. Vehkamäki, *Chem. Commun.* **2019**, *55*, 5946.

[186] L. Himanen, A. Geurts, A. S. Foster, P. Rinke, *Adv. Sci.* **2019**, *6*, 1900808.

[187] L. D. Yee, G. Isaacman-VanWertz, R. A. Wernis, M. Meng, V. Rivera, N. M. Kreisberg, S. V. Hering, M. S. Bering, M. Glasius, M. A. Upshur, A. G. Bé, R. J. Thomson, F. M. Geiger, J. H. Offenberg, M. Lewandowski, I. Kourtchev, M. Kalberer, S. D. Sá, S. T. Martin, M. L. Alexander, B. B. Palm, W. Hu, P. Campuzano-Jost, D. A. Day, J. L. Jimenez, Y. Liu, K. A. McKinney, P. Artaxo, J. Viegas, A. Manzi, et al., *Atmos. Chem. Phys.* **2018**, *18*, 10433.

[188] C. N. Jen, L. E. Hatch, V. Selimovic, R. J. Yokelson, R. Weber, A. E. Fernandez, N. M. Kreisberg, K. C. Barsanti, A. H. Goldstein, *Atmos. Chem. Phys.* **2019**, *19*, 1013.

[189] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. t Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, et al., *Scientific Data* **2016**, *3*, 160018.