Nucleic Acids Research, 2025, **53**, D1692–D1696
https://doi.org/10.1093/nar/gkae961
Advance access publication date: 30 October 2024
Database issue

mVOC 4.0: a database of microbial volatiles

Emanuel Kemmler[1], Marie Chantal Lemfack[2], Andrean Goede[1], Kathleen Gallo[1], Serge M.T. Toguem[2], Waqar Ahmed[3], Iris Millberg[2], Saskia Preissner[1], Birgit Piechulla[2] and Robert Preissner[1,*]

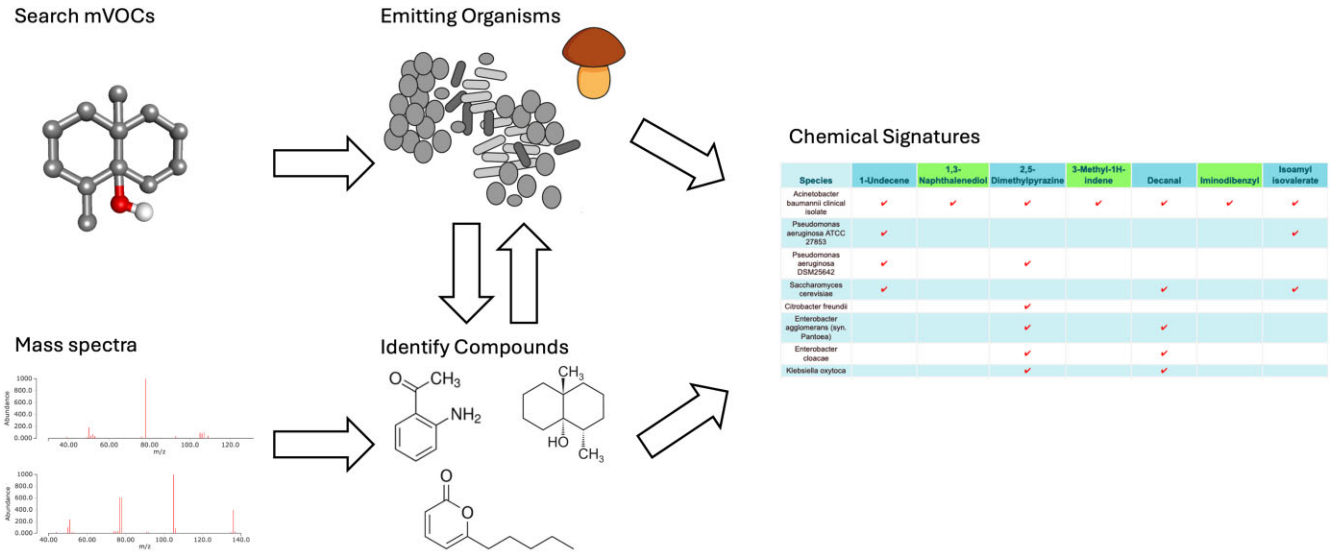
[1]Institute for Physiology & Science-IT, Charité – University Medicine Berlin, 10115 Berlin, Germany
[2]Institute of Biological Sciences, University of Rostock, Albert-Einstein-Straße 3, 18059 Rostock, Germany
[3]Division of Immunology, Immunity to Infection and Respiratory Medicine, School of Biological Sciences, The University of Manchester, Manchester, UK

*To whom correspondence should be addressed. Tel: +49 30 450 528 495; Email: robert.preissner@charite.de

Abstract

Metabolomic microbiome research has become an important topic for understanding agricultural, ecological as well as health correlations. Only the determination of both the non-volatile and the volatile organic compound (mVOC) production by microorganisms allows a holistic view for understanding the complete potential of metabolomes and metabolic capabilities of bacteria. In the recent past, more and more bacterial headspaces and culture media were analyzed, leading to an accumulation of about 3500 mVOCs in the updated mVOC 4.0 database, including compounds synthesized by the newly discovered non-canonical terpene pathway. Approximately 10% of all mVOCs can be assigned with a biological function, some mVOCs have the potential to impact agriculture in the future (e.g. eco-friendly pesticides) or animal and human health care. mVOC 4.0 offers various options for exploring extensively annotated mVOC data from different perspectives, including improved mass spectrometry matching. The mVOC 4.0 database includes literature searches with additional relevant keywords, making it the most up-to-date and comprehensive publicly available mVOC platform at: http://bioinformatics.charite.de/mvoc.

Graphical abstract



Introduction

With an estimated 10^{12} species, microorganisms are the largest and most diverse organismal kingdom on Earth (1,2). Numerous biochemical pathways present in one or the other microbial genus or species enable them to grow in moderate as well as extreme habitats and under distinct environmental conditions. Due to their diverse metabolic activities, microorganisms can also produce many different metabolites. Metabolomic studies in the past focussed primarily on the wealth and diversity of non-volatile compounds, while volatile organic metabolites remained largely unexplored (3). To fully understand their ecological and biological functions, as well

Received: September 13, 2024. **Revised:** October 7, 2024. **Editorial Decision:** October 8, 2024. **Accepted:** October 10, 2024

as their metabolic potential, both types of compounds need to be discovered and their structure elucidated (4).

A revival in the interest in microbial volatile organic compounds (mVOCs) took place in the first decade of the 21st century when, in dual culture systems, e.g. in divided Petri dishes, the influence of volatile compounds on other organisms was impressively demonstrated (5–8). Due to their volatility, mVOCs are considered important infochemicals which can quickly spread in the atmosphere to contact and interact with receiver organisms. Similar to known plant scent compounds or insect pheromones, the biological and ecological functions of mVOCs are considered to act as key molecules in attraction, defense, and information exchange. These mVOCs derive from different biosynthetic pathways and are grouped into aromatic compounds, fatty acid derivatives, terpenoids, and nitrogen- and sulfur-containing compounds, to name the most abundant (9). Besides the respectable structural diversity of mVOCs, it has to be noted that only about 10% of mVOCs a defined function could be assigned to (10,11).

In addition to the biological relevance of individual as well as mixtures of mVOCs, increasing attention was given to the potential of applications in agriculture as sustainable and eco-friendly alternatives to chemically synthesized pesticides and fertilizers (12,13) and animal and human health care as non-invasive diagnostic tools (14–17). Furthermore, the implementation of mVOC detection devices is used to monitor and control, e.g. foodstuff processing (18,19) or hardware effluvium (e.g. seat covers in cars (20)) as well as mVOC perspiration in buildings (21).

A new mVOC research direction appeared when unusual and unique compound structures were isolated and elucidated in microbial headspaces. Sodorifen was the first C16-derived non-canonical terpene that subsequently led to the discovery of a new biosynthetic route that is exclusively present in bacteria (22–25).

Taken together, mVOCs represent an attractive research field that still discloses several urgent questions. With over 570 citations in total, the previously published mVOC database has proven to be a well-cited and useful tool for many scientists with different research interests. Here, we introduce mVOC 4.0, which covers the literature search till July 2024. Furthermore, the webserver offers enhanced options for exploring extensively annotated mVOC data from different perspectives, including improved mass spectrometry matching. With about 3500 compounds, mVOC 4.0 now offers an expanded collection of mVOCs, featuring a total of 1349 species and isolates.

Materials and methods

Data collection

The data was collected using a combination of manual curation and text mining of scientific publications.

The manual curation was carried out by experts in the field. Scientific publications were manually selected and examined for microbial volatile organic compounds (mVOCs) and extracted together with related information like the corresponding producing organisms.

For the text mining process, lists of potential volatiles, species, and search criteria were constructed. The full-text search was performed on all available open-access publications from 2017 using the Konstanz Information Miner (KNIME) in combination with the European PubMed Central Advanced Search API (https://europepmc.org/advancesearch). The abstracts of all PubMed IDs were then checked against the list of species using a flexible matching method to account for misspellings and transformation errors. All in all, about 600 PubMed requests referencing volatiles were conducted, resulting in a list of about 30 000 publications with at least one microorganism mentioned within the abstract. The manuscripts, along with their supplementary material, were downloaded and cross-checked with a list of approximately 450 000 names and synonyms of potential volatile compounds. The number of volatiles identified in each publication was assessed. Subsequently, the publications were systematically reviewed and confirmed mVOCs were manually extracted.

Data processing and annotation

The resulting data set was cleaned and filtered to exclude erroneous and inconclusive entries. Names of compounds and organisms were standardized, and duplications were removed. Compounds were matched to the PubChem (26) database and annotated with IUPAC name, SMILES notation, and the International Chemical Identifier (InChI) according to their PubChem ID.

The annotation of the data with taxonomic lineage and physicochemical information like vapor pressure and boiling point was performed using the R programming language (27), utilising multiple web scraping techniques, like XML extraction. The taxonomic information was matched to the organisms using the NCBI taxonomy database (28), and physicochemical information was gathered from PubChem, ChemSynthesis (https://www.chemsynthesis.com/), and the Human Metabolome Database (HMDB) (29).

Software implementation

The data is maintained in a relational MySQL database (http://www.mysql.com/), hosted on an Ubuntu system within the Charité Berlin IT infrastructure. The website back-end operates on a lab-based LAMP server (Linux/Apache/MySQL/PHP), with PHP as the server-side backend language. The database connection is facilitated through the MySQL interface, while front-end data is delivered using a combination of HTML and JavaScript from submission responses and PHP requests.

Database functionalities

Search for mass spectra

The platform allows for the identification of the three most intense peaks in mass spectra, which is particularly useful for researchers working with experimental data.

In contrast to mVOC 2.0, mVOC 4.0 operates using the open database Massbank (30), which enables the user to choose a tolerance for the m/z value, and to check mass spectra within the six most intense peaks irrespective of their order. This is crucial because m/z values, peak intensity, and the order of peaks can differ for the same compound based on the age and quality of the MS instruments.

Results of the search are ordered according to the maximal peak number. For an identified compound, only the best-fitting mass spectrum is shown.

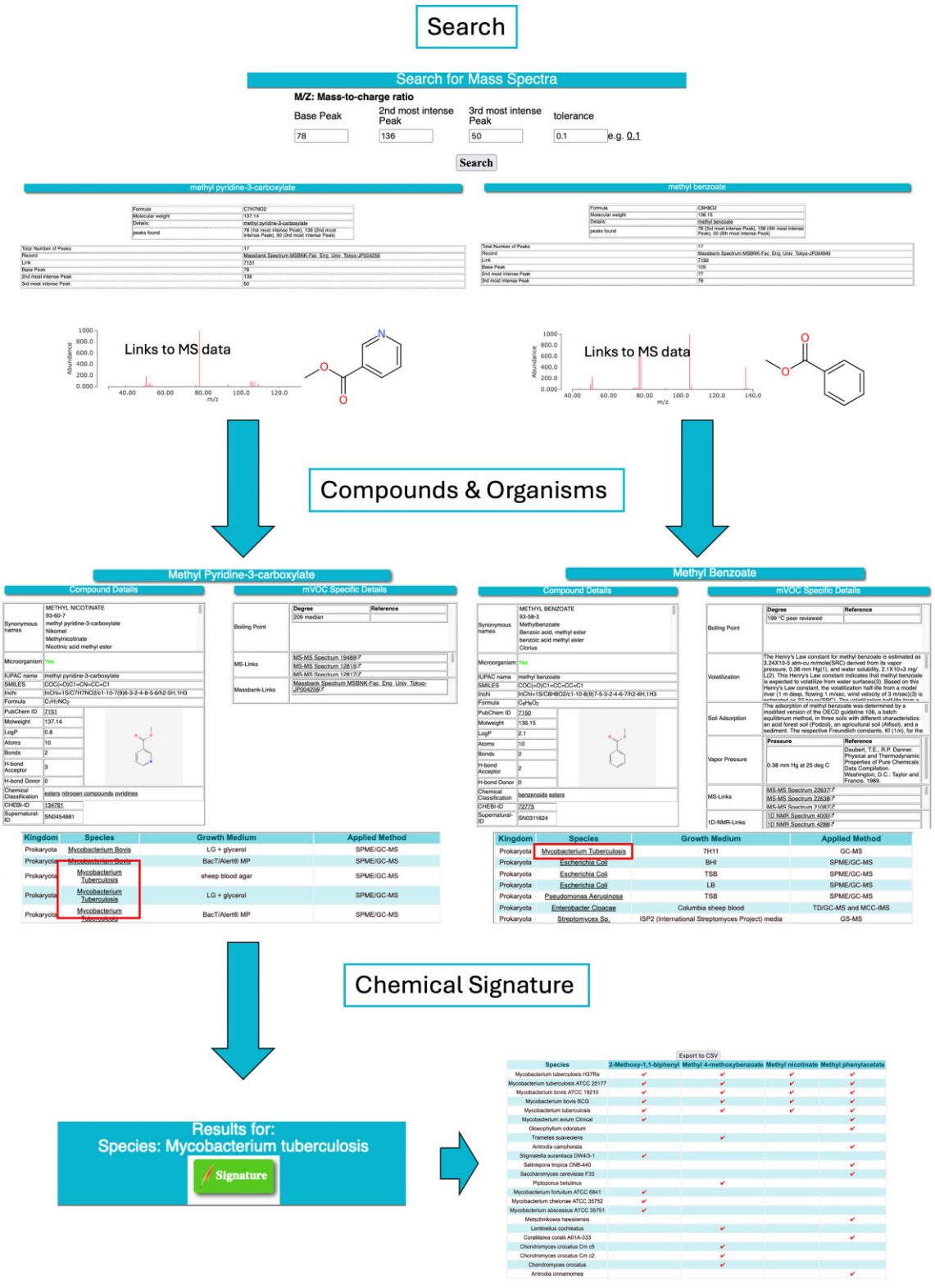


Figure 1. Use case of mVOC 4.0 database. A query of mass peaks was executed, corresponding compounds and emitting species were identified, and a chemical signature was found. The data hints to a potential contamination with *Mycobacterium tuberculosis*, an organism known to cause tuberculosis.

Search mVOCs

mVOC 4.0 offers a wide range of search options designed to facilitate in-depth exploration of microbial volatile organic compounds (mVOCs). Users can conduct targeted searches for specific mVOCs by applying various criteria, including compound name, molecular formula, and PubChem ID. Moreover, it is possible to search mVOCs by specifying molecular weight ranges or filtering by chemical class, enabling precise identification of relevant compounds. Additionally, the platform allows users to select a specific organism, retrieve all associated volatiles, and to search by emitting and receiving species for a selected mVOC. These versatile search functionalities ensure that users can efficiently retrieve and analyze relevant information, supporting a wide array of scientific investigations into the roles and applications of mVOCs.

Search results

Search results on the mVOC 4.0 website are presented in a table, making it easy for users to quickly browse relevant details. Each entry typically includes general information about the compound, the organisms that emit it, and specific mVOC details.

General information encompasses identifiers such as PubChem ID, SMILES and the IUPAC name, along with chemical details like molecular weight, bond count and LogP. A list of synonymous names is also given. The boiling point and vapour pressure, as well as information about volatilization and soil adsorption, are listed under mVOC-specific details.

Additionally, the results provide information on the biological source of the mVOC, including the specific organism or group of organisms known to emit the compound. This is particularly useful for researchers investigating the ecological roles or potential applications of specific volatiles.

For users conducting searches based on mass spectrometry data, the results also include information on the most intense peaks, aiding in the identification and comparison of compounds.

Use case

The mVOC 4.0 database enables users to explore the data from multiple different perspectives. Chemical structures can be searched via compound name, Pubchem-ID or chemical properties.

Another method for identifying VOCs involves analyzing specific high or low peaks in mass spectra or determining all mVOCs released by a particular species of microorganisms or fungi.

As described in Piechulla and Lemfack (31), mVOCs can be used for phenotyping microorganisms. This is especially helpful for the identification of microorganisms growing on hardware, e.g. indoor walls, that are indicators for contaminants and pollutants with potential consequences for human health. Similarly, mVOCs can also be used as diagnostic tools in medicine by the discovery and identification of potentially disease-causing bacteria and fungi (32). Syhre and Chambers (33,34) proposed a method to identify tuberculosis patients at an early stage of the infection using mVOCs.

As shown in Figure 1, mVOC 4.0 could be utilized to search for mass spectra peaks generated by mass spectrometry of breath samples from a patient. Using the 'Search mass spectra' subpage, querying the mass peaks (base peak of 78, 2nd most intense peak of 136, and 3rd most intense peak of 50) reveals methyl pyridine-3-carboxylate and methyl benzoate. Clicking on the compound names provides detailed information about their chemical properties, emitting species, growth medium, and extraction method. The list of synonyms shows that methyl nicotinate is the common name for methyl pyridine-3-carboxylate, which can be helpful for further research. In the list of emitting organisms, *Mycobacterium tuberculosis* is listed for both compounds, a bacteria known to cause tuberculosis. By selecting the organism name, 19 more compounds emitted by this organism are listed. Additionally, it is possible to retrieve a table of signature compounds using the signature button. This page shows a downloadable table with four different strains of *M. tuberculosis*, along with their corresponding signature molecules and other organisms that also emit at least one of these molecules. In a clinical setting, this combined information could be used to further explore in a non-invasive manner whether a patient is truly at risk of developing tuberculosis, for example, by examining the presence of the other signature molecules.

Conclusion and perspective

mVOC 2.0 has been proven to be a well-cited and useful tool for many scientists with different perspectives and research interests. Over the last years, we constantly developed the mVOC database, especially improving the data basis, leading to the 4th iteration of the web server, mVOC 4.0. By incorporating a comprehensive collection of literature, enhanced keyword searches, and improved mass spectrometry matching, mVOC 4.0 provides researchers with the most current and exhaustive resource available for exploring the ecological, biological, and metabolic roles of mVOCs.

Looking ahead, the mVOC 4.0 database has the potential to drive further research and innovation in multiple fields, including microbiology, environmental science, and medical biotechnology. As more data becomes available, the mVOC database will continue to evolve, remaining a valuable resource for researchers in the future.

Data availability

The mVOC 4.0 database is publicly available at https://bioinformatics.charite.de/mvoc without the need for login or registration. Results are displayed immediately on the website without the need to provide an email address.

Acknowledgements

The authors thank Andrea Mellin for data acquisition.

Funding

German Research Foundation DFG [Pi153/36-1 and 2]; University of Rostock; German Research Foundation DFG as part of the clinical research unit [CRU339]: Food allergy and tolerance (FOOD@) [428445448, 428447634].

Conflict of interest statement

None declared.

References

1. Locey,K.J. and Lennon,J.T. (2016) Scaling laws predict global microbial diversity. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, 5970–5975.
2. Horner-Devine,M.C., Leibold,M.A., Smith,V.H. and Bohannan,B.J.M. (2003) Bacterial diversity patterns along a gradient of primary productivity. *Ecol. Lett.*, **6**, 613–622.
3. Honeker,L.K., Graves,K.R., Tfaily,M.M., Krechmer,J.E. and Meredith,L.K. (2021) The volatilome: a vital piece of the complete soil metabolome. *Front. Environ. Sci.*, **9**, 649905.
4. Meredith,L.K. and Tfaily,M.M. (2022) Capturing the microbial volatilome: an oft overlooked 'ome'. *Trends Microbiol.*, **30**, 622–631.
5. Ryu,C.M., Farag,M.A., Hu,C.H., Reddy,M.S., Wei,H.X., Paré,P.W. and Kloepper,J.W. (2003) Bacterial volatiles promote growth in Arabidopsis [published correction appears in Proc Natl Acad Sci U S A, 100(14):8607]. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 4927–4932.
6. Kai,M., Effmert,U., Berg,G. and Piechulla,B. (2007) Volatiles of bacterial antagonists inhibit mycelial growth of the plant pathogen Rhizoctonia solani. *Arch. Microbiol.*, **187**, 351–360.
7. Vespermann,A., Kai,M. and Piechulla,B. (2007) Rhizobacterial volatiles affect the growth of fungi and Arabidopsis thaliana. *Appl. Environ. Microb.*, **73**, 5639–5641.
8. Kai,M., Haustein,M., Molina,F., Petri,A., Scholz,B. and Piechulla,B. (2009) Bacterial volatiles and their action potential. *Appl. Microbiol. Biotechnol.*, **81**, 1001–1012.
9. Lemfack,M.C., Gohlke,B.O., Toguem,S.M.T., Preissner,S., Piechulla,B. and Preissner,R. (2018) mVOC 2.0: A database of microbial volatiles. *Nucleic Acids Res.*, **46**, D1261–D1265.
10. Piechulla,B., Lemfack,M.C. and Kai,M. (2017) Effects of discrete bioactive microbial volatiles on plants and fungi. *Plant Cell Environ.*, **40**, 2042–2067.
11. Piechulla,B., Lemfack,M.C. and Magnus,N. (2020) Bioactive bacterial volatile organic compounds- an overview and critical comments. In: Ryu,C., Weisskopf,L. and Piechulla,B. (eds.) *Bacterial Volatile Compounds as Mediators of Airborne Interactions*. Springer Nature Singapore Pte Ltd., Singapore. pp. 39–92.
12. de Boer,W., Li,X., Meisner,A. and Garbeva,P. (2019) Pathogen suppression by microbial volatile organic compounds in soils. *FEMS Microbiol. Ecol.*, **95**, fiz105.
13. Thomas,G., Withall,D. and Birkett,M. (2020) Harnessing microbial volatiles to replace pesticides and fertilizers. *Microb. Biotechnol.*, **13**, 1366–1376.
14. Hertel,M., Preissner,R., Gillissen,B., Schmidt-Westhausen,A.M., Paris,S. and Preissner,S. (2016) Detection of signature volatiles for cariogenic microorganisms. *Eur. J. Clin. Microbiol. Infect. Dis.*, **35**, 235–244.
15. Hartwig,S., Raguse,J.D., Pfitzner,D., Preissner,R., Paris,S. and Preissner,S. (2017) Volatile Organic Compounds in the Breath of Oral Squamous Cell Carcinoma Patients: A Pilot Study. *Otolaryngol. Head Neck Surg.*, **157**, 981–987.
16. Elmassry,M.M. and Piechulla,B. (2020) Volatilomes of Bacterial Infections in Humans. *Front. Neurosci.*, **14**, 257.
17. Fitzgerald,S., Holland,L., Ahmed,W., Piechulla,B., Fowler,S.J. and Morrin,A. (2024) Volatilomes of human infection. *Anal Bioanal Chem*, **416**, 37–53.
18. Lorenzo,J.M., Bedia,M. and Bañón,S. (2013) Relationship between flavour deterioration and the volatile compound profile of semi-ripened sausage. *Meat Sci.*, **93**, 614–620.
19. Bhadra,S., Narvaez,C., Thomson,D.J. and Bridges,G.E. (2015) Non-destructive detection of fish spoilage using a wireless basic volatile sensor. *Talanta*, **134**, 718–723.
20. Diekmann,N., Burghartz,M., Remus,L., Kaufholz,A.L., Nawrath,T., Rohde,M., Schulz,S., Roselius,L., Schaper,J., Mamber,O., *et al.* (2013) Microbial communities related to volatile organic compound emission in automobile air conditioning units. *Appl. Microbiol. Biotechnol.*, **97**, 8777–8793.
21. Garbacz,M., Malec,A., Duda-Saternus,S., Suchorab,Z., Guz,Ł. and Łagód,G. (2020) Methods for early detection of microbiological infestation of buildings based on gas sensor technologies. *Chemosensors*, **8**, 7.
22. Domik,D., Thürmer,A., Weise,T., Brandt,W., Daniel,R. and Piechulla,B. (2016) A terpene synthase is involved in the synthesis of the volatile organic compound sodorifen of Serratia plymuthica 4Rx13. *Front. Microbiol.*, **7**, 737.
23. Domik,D., Magnus,N. and Piechulla,B. (2016) Analysis of a new cluster of genes involved in the synthesis of the unique volatile organic compound sodorifen of Serratia plymuthica 4Rx13. *FEMS Microbiol. Lett.*, **363**, fnw139.
24. Magnus,N., von Reuss,S.H., Braack,F., Zhang,C., Baer,K., Koch,A., Hampe,P.L., Sutour,S., Chen,F. and Piechulla,B. (2023) Non-canonical biosynthesis of the brexanet-type bishomosesquiterpene chlororaphen through two consecutive methylation steps in Pseudomonas chlororaphis O6 and Variovorax boronicumulans PHE5-4. *Angew. Chem. Int. Ed Engl.*, **62**, e202303692.
25. Duan,Y.T., Koutsaviti,A., Harizani,M., Ignea,C., Roussis,V., Zhao,Y., Ioannou,E. and Kampranis,S.C. (2023) Widespread biosynthesis of 16-carbon terpenoids in bacteria. *Nat. Chem. Biol.*, **19**, 1532–1539.
26. Kim,S., Chen,J., Cheng,T., Gindulyte,A., He,J., He,S., Li,Q., Shoemaker,B.A., Thiessen,P.A., Yu,B., *et al.* (2023) PubChem 2023 update. *Nucleic Acids Res.*, **51**, D1373–D1380.
27. R. Core Team (2021) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
28. Schoch,C.L., Ciufo,S., Domrachev,M., Hotton,C.L., Kannan,S., Khovanskaya,R., Leipe,D., Mcveigh,R., O'Neill,K., Robbertse,B., *et al.* (2020) NCBI Taxonomy: A comprehensive update on curation, resources and tools. *Database (Oxford).*, **2020**, baaa062.
29. Wishart,D.S., Guo,A., Oler,E., Wang,F., Anjum,A., Peters,H., Dizon,R., Sayeeda,Z., Tian,S., Lee,B.L., *et al.* (2022) HMDB 5.0: the Human Metabolome Database for 2022. *Nucleic Acids Res.*, **50**, D622–D631.
30. Horai,H., Arita,M., Kanaya,S., Nihei,Y., Ikeda,T., Suwa,K., Ojima,Y., Tanaka,K., Tanaka,S., Aoshima,K., *et al.* (2010) MassBank: A public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.*, **45**, 703–714.
31. Piechulla,B. and Lemfack,M.C. (2016) Microbial volatiles and their biotechnological applications. In: *Plant Specialized Metabolism*. CRC Press, pp. 251–252.
32. Mentel,S., Gallo,K., Wagendorf,O., Preissner,R., Nahles,S., Heiland,M. and Preissner,S. (2021) Prediction of oral squamous cell carcinoma based on machine learning of breath samples: a prospective controlled study. *BMC Oral Health*, **21**, 500.
33. Syhre,M. and Chambers,S.T. (2008) The scent of Mycobacterium tuberculosis. *Tuberculosis (Edinb.)*, **88**, 317–323.
34. Syhre,M., Manning,L., Phuanukoonnon,S., Harino,P. and Chambers,S.T. (2009) The scent of Mycobacterium tuberculosis–part II breath. *Tuberculosis (Edinb.)*, **89**, 263–266.

Received: September 13, 2024. **Revised:** October 7, 2024. **Editorial Decision:** October 8, 2024. **Accepted:** October 10, 2024