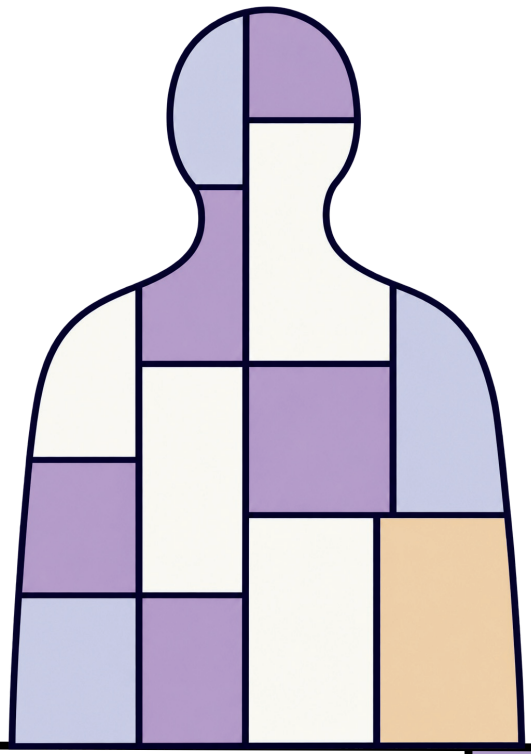
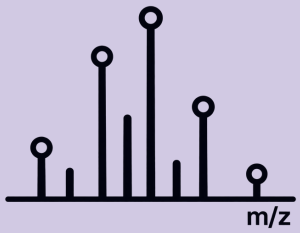
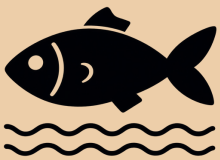
Building Blocks for Exposome Characterization

Representation, Extraction, and Interpretation of Chemical Information in LC–HRMS-Based Non-Target Analysis







Viktoriia Turkina

Building Blocks for Exposome Characterization

Representation, Extraction, and Interpretation of Chemical Information in LC–HRMS-Based Non-Target Analysis

Viktoriia Turkina

Building Blocks for Exposome Characterization
Representation, Extraction, and Interpretation of Chemical Information in LC–HRMS-Based Non-Target Analysis

PhD Dissertation, University of Amsterdam, the Netherlands
ISBN 978-94-6534-580-2

Cover, layout, and design: Viktoriia Turkina
Paranymphs: Anna Chernysheva & Lapo Renai

Copyright © 2026 by Viktoriia Turkina

Building Blocks for Exposome Characterization

Representation, Extraction, and Interpretation of Chemical Information in LC–HRMS-Based Non-Target Analysis

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op maandag 9 november 2026, te 13:00 uur

door Viktoriia Turkina
geboren te Vizinga

Promotiecommissie

Promotores:	prof. dr. G. L. Corthals	Universiteit van Amsterdam
Copromotores:	dr. S. Samanipour	Universiteit van Amsterdam
	dr. ir. A. Petrignani	Universiteit van Amsterdam
Overige leden:	dr. F.M. Béen	Vrije Universiteit Amsterdam
	prof. dr. P.D. Ferguson	Universiteit van Amsterdam
	prof. dr. A.C. van Asten	Universiteit van Amsterdam
	prof. dr. K. Keune	Universiteit van Amsterdam
	prof. dr. ir. B. Ensing	Universiteit van Amsterdam
	prof. dr. J. Hollender	Eawag
	dr. A.M. Brunner	TNO

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

The work described in this thesis has been carried out within the Environmental Modeling & Computational Mass Spectrometry (EMCMS) Group of the University of Amsterdam.

Table of contents

1 General introduction 11

1.1 Exposome Chemical Space Characterization 11
1.2 Scope of the thesis 15
1.3 References 17

2 Critical assessment of the chemical space covered by LC-HRMS non-targeted analysis 23

2.1 Introduction 25
2.2 Methods 28
2.3 Discussion 32
2.4 Recommendations and Outlook 44
2.5 Supporting Information 45
2.6 References 47

3 Enhancing Predictive Modeling with Molecular Fingerprint Fusion Strategies 55

3.1 Introduction 57
3.2 Methods 60
3.3 Results and discussion 66
3.4 Conclusions 76
3.5 Supporting Information 77
3.6 References 79

4 Prioritization of unknown LC-HRMS features based on predicted toxicity categories 85

4.1 Introduction 87
4.2 Methods 89
4.3 Results and Discussion 97
4.4 Implications 105
4.5 Supporting Information 106
4.6 References 108

5 From Spectra to Local Networks: Evaluating MS^2 Similarity Metrics 115

5.1 Introduction 117
5.2 Experimental Section 119
5.3 Results and Discussions 124
5.4 Conclusions 135
5.5 Supporting Information 137
5.6 References 139

6 Quantifying Fragment Ambiguity in LC-HRMS/MS Spectral Databases 143

6.1 Introduction 145
6.2 Experimental Section 147
6.3 Results 152
6.4 Conclusions 162
6.5 Supporting Information 164
6.6 References 168

7 Conclusion and outlook 173

Summary 179

Nederlandse samenvatting 183

Publications 187

Acknowledgements 193

“What we know is a drop; what we don’t know is an ocean.”

— Isaac Newton

1

Introduction and scope of the thesis

1.1 Exposome Chemical Space Characterization

The concept of chemical space has been introduced to describe the universe of all possible chemical structures. [1–3] A biologically and environmentally relevant subset of this space is represented by the human and environmental exposome chemical space, which comprises the multitude of chemicals to which humans and ecosystems are exposed over time. [4, 5] In recent decades, the number and diversity of chemicals produced and used worldwide have increased dramatically. Tens of thousands of substances are currently registered for commercial use. They, together with their generated transformation products, metabolites, impurities, and reaction products, are continuously released into the environment. [6, 7] Consequently, the chemical space of the exposome expands rapidly, meaning that humans and ecosystems are exposed to highly complex mixtures of chemicals, many of which remain poorly characterized with respect to their occurrence, fate, and potential adverse effects. [8, 9] Understanding these exposures and their implications for environmental and human health has therefore become one of the central challenges in environmental and human health research. [10, 11]

Despite substantial advances in analytical chemistry and toxicology, our current un-

derstanding of chemical exposure remains incomplete. [8,12] Regulatory monitoring programs traditionally focus on relatively small groups of predefined compounds selected on the basis of known occurrence, hazard, or regulatory relevance. [13, 14] While these targeted approaches provide highly reliable quantitative information, they inevitably overlook a large proportion of the chemicals present in environmental samples. [15] Consequently, comprehensive characterization of chemical exposure requires analytical strategies that are not restricted to predefined target lists and that can capture a broader range of chemical structures and physicochemical properties. [16]

Although a large fraction of the chemical exposome remains unexplored, a subset of chemicals has already been detected in environmental compartments and linked to adverse effects on human and environmental health. [17] Such compounds are commonly referred to as contaminants of emerging concern (CECs). A large proportion of CECs consists of semipolar and polar organic compounds. [18] For the analysis of such chemicals, liquid chromatography coupled to high-resolution tandem mass spectrometry (LC-HRMS/MS) has become one of the most powerful analytical platforms available. The combination of chromatographic separation, accurate mass measurements, isotopic patterns, and fragmentation information enables the detection of a broad range of compounds within a single analysis. [19] To reduce the bias associated with predefined analyte lists, LC-HRMS/MS has been increasingly employed in non-target analysis (NTA) workflows. Unlike targeted approaches, NTA does not require prior knowledge of the chemicals present in a sample and aims to detect all measurable compounds within the analytical window. [20–22]

The ability of NTA to expand the measurable fraction of chemical space has transformed environmental screening and exposomics. However, this increased analytical coverage comes at the cost of substantially more complex datasets. [23,24] Environmental samples routinely generate hundreds to thousands of detected features, typically defined by a combination of retention time, precursor ion mass-to-charge ratio, and chromatographic peak intensity. To process these data, numerous computational workflows and software platforms have been developed, including vendor-provided (e.g. Compound Discoverer [25]), commercial (e.g. mzmine [26]), and open-source solutions (e.g. MS-DIAL [27], OpenMS [28], patRoon [29], jHRMSToolBox [30]). [31] Nevertheless, despite continuous improvements in instrumentation and data processing, fewer than 5% of detected features can typically be assigned to chemical structures with high confidence. [32] The majority, therefore, remain unidentified, creating a substantial gap between chemical detection and chemical

understanding.

A fundamental challenge underlying NTA is that chemical structures are not observed directly. Instead, LC-HRMS/MS records properties related to the structure, such as retention time, accurate mass, isotopic patterns, and fragmentation spectra, from which structural information must be inferred. [19, 33] While accurate mass and isotopic patterns can provide information on elemental composition and molecular formula, they are often insufficient to distinguish among multiple candidate structures. Consequently, tandem mass spectrometry (MS^2), which provides information about molecular substructures through fragmentation behavior, has become one of the most important sources of structural information in NTA. Fragmentation spectra are used for spectral library matching, [34, 35] molecular networking, [36–38] suspect screening, [39] candidate ranking, [40] molecular fingerprint prediction, [41, 42] property estimation, [43] and toxicity assessment . [44–46] In these applications, spectral similarity is implicitly treated as a proxy for structural similarity, assuming that compounds producing similar fragmentation patterns are structurally related.

The relationship between molecular structure and MS^2 spectra is neither simple nor unique. Structurally related compounds may produce markedly different fragmentation patterns depending on experimental conditions, while structurally distinct compounds can occasionally generate highly similar spectra. [31] Consequently, the interpretation of MS^2 data depends strongly on how fragmentation information is represented, processed, and compared. [38, 47–49] Similar considerations apply to molecular structures themselves. In cheminformatics, chemical structures are transformed into machine-readable representations such as molecular descriptors and fingerprints, each capturing different aspects of molecular information. [50] The choice of representation directly influences the performance of predictive models, the estimation of physicochemical and toxicological properties, and the interpretation of structure–activity relationships. [51]

From this perspective, the characterization of chemicals in NTA can be viewed as a multilevel information-representation problem. At the molecular level, chemical structures must be encoded into representations suitable for computational analysis and predictive modeling. [51, 52] At the experimental level, LC–HRMS measurements provide alternative representations through chromatographic behavior, accurate masses, and fragmentation patterns. [31, 47] Both molecular fingerprints and MS^2 spectra capture only part of the information contained within a chemical structure, and their usefulness depends on how

effectively this information can be extracted and utilized.

A fundamental challenge here is the limited understanding of the information content available within molecular and spectral representations. Most annotation and identification workflows assume that MS^2 spectra contain sufficient information to distinguish between candidate structures. [53, 54] Yet the reliability of spectral matching depends not only on the similarity metric employed but also on the diagnostic level of individual fragments and the composition of the reference databases against which spectra are compared. Similar questions have been extensively investigated in fields such as proteomics and metabolomics through false discovery rate (FDR) estimation and confidence assessment frameworks. [55, 56] In environmental NTA, however, quantitative evaluations of the information content and reliability of MS^2 spectra remain comparatively limited.

A better understanding of how chemical information is encoded in both molecular structures and mass spectrometric data is critical because these representations form the foundation of many downstream applications. Evaluating their strengths, limitations, and information content is therefore essential for improving confidence in NTA-based chemical characterization. These considerations raise several fundamental questions. How much of the environmentally relevant chemical space is currently accessible through LC-HRMS-based NTA workflows? Which molecular and spectral representations are most effective for extracting chemically meaningful information? To what extent can chromatographic and fragmentation data be used to prioritize unknown chemicals before identification? How reliable are spectral similarity metrics as indicators of structural relatedness? And how diagnostic are the individual fragments that underpin spectral matching and annotation?

Addressing these questions requires investigation across multiple levels of chemical characterization, from chemical space coverage and molecular representation to feature prioritization and compound annotation. This thesis, therefore, examines the measurable chemical space of the exposome, explores strategies for representing and extracting chemical information, and evaluates the assumptions underlying MS^2-based identification workflows.

1.2 Scope of the thesis

This thesis investigates how chemical information can be extracted and utilized across different stages of LC–HRMS/MS NTA, from chemical space exploration and feature prioritization to compound annotation. Particular emphasis is placed on understanding the measurable chemical space of the exposome, developing strategies for extracting information from unknown features, and evaluating the information content underlying MS^2-based annotation approaches. To address these objectives, the thesis is divided into three interconnected parts. The first part focuses on assessing the extent to which current NTA workflows capture environmentally relevant chemical space. The second part investigates how molecular and experimental representations can be used to extract information from unknown features and support their prioritization. The third part examines the interpretation, limitations, and potential pitfalls of MS^2-based annotation approaches.

Chapter 2 evaluates the coverage of chemical space achieved by LC–HRMS-based NTA through a meta-analysis of studies published between 2017 and 2023. Experimental parameters, reporting practices, and reported compound identifications are systematically collected and analyzed to assess the extent to which current workflows explore the chemical exposome.

Chapter 3 investigates molecular representation as a means of extracting chemical information. Different strategies for combining non-hashed molecular fingerprints are explored and evaluated across multiple predictive modeling tasks in order to assess how complementary structural representations can improve the description of molecular properties and activities.

Chapter 4 focuses on the prioritization of unknown LC–HRMS features without requiring prior compound identification. Using chromatographic and fragmentation information, machine-learning models are developed to relate experimental measurements directly to aquatic toxicity categories.

Chapter 5 examines the role of spectral similarity in compound annotation. A range of MS^2 similarity metrics is evaluated using publicly available spectral libraries to investigate how effectively spectral similarity reflects structural relatedness and supports compound discrimination.

Chapter 6 investigates fragment-level ambiguity in tandem mass spectrometry data and develops a framework for quantifying the diagnostic value of individual fragment ions. The chapter examines how fragment specificity varies across spectral databases and what this implies for the interpretation of spectral matching results.

Together, these chapters progress from assessing what fraction of chemical space is currently observable, through methods for extracting information from unknown features, to investigating the information content underlying MS^2-based annotation approaches.

1.3 References

[1] Kirkpatrick P, Ellis C. Chemical space. Nature. 2004 12;432:823. doi:10.1038/432823a.

[2] Reymond J. L. The Chemical Space Project. Acc Chem Res. 2015 3;48(3):722-30. doi:10.1021/ar500432k.

[3] Reymond J. L. Chemical space as a unifying theme for chemistry. J Cheminform. 2025 1;17(1):6. doi:10.1186/S13321-025-00954-0.

[4] Miller G. W, Jones D. P. The Nature of Nurture: Refining the Definition of the Exposome. Toxicol Sci. 2014 1;137(1):1-2. doi:10.1093/toxsci/kft251.

[5] Villette C, Miller G. W, Sibilia J, Heintz D. Inviting ecosystems into the exposome framework. Exposome. 2026 1;6(1). doi:10.1093/exposome/osag007.

[6] Managing the systemic use of chemicals in Europe. European Environment Agency (EEA) (2026, June, 18);. https://www.eea.europa.eu/en/analysis/publications/managing-the-systemic-use-of-chemicals.

[7] Levi P. G, Cullen J. M. Mapping Global Flows of Chemicals: From Fossil Fuel Feedstocks to Chemical Products. Environ Sci Technol. 2018 2;52(4):1725-34. doi:10.1021/acs.est.7b04573.

[8] Samanipour S, Barron L. P, van Herwerden D, Praetorius A, Thomas K. V, O'Brien J. W. Exploring the Chemical Space of the Exposome: How Far Have We Gone? JACS Au. 2024 7;4(7):2412-25. doi:10.1021/jacsau.4c00220.

[9] Bernhardt E. S, Rosi E. J, Gessner M. O. Synthetic chemicals as agents of global change. Front Ecol Environ. 2017. doi:10.1002/fee.1450.

[10] Vermeulen R, Schymanski E. L, Barabási A. L, Miller G. W. The exposome and health: where chemistry meets biology. Science. 2020 1;367(6476):392. doi:10.1126/science.aay3164.

[11] DeBord D. G, Carreón T, Lentz T. J, Middendorf P. J, Hoover M. D, Schulte P. A. Use of the "Exposome" in the Practice of Epidemiology: A Primer on -Omic Technologies. Am J Epidemiol. 2016 8;184(4):302. doi:10.1093/aje/kwv325.

[12] Qiu W, Gago-Ferrero P, Hollender J, Leusch F. D. L, Richardson S. D, Samanipour S, et al. Advancing the Discovery of Emerging Contaminants: A Leap in Technology and Data. Environ Sci Technol. 2026 6;60(22):15530-43. doi:10.1021/acs.est.6c03027.

[13] Brack W, Dulio V, Ågerstrand M, Allan I, Altenburger R, Brinkmann M, et al. Towards the review of the European Union Water Framework Directive: Recommendations for more efficient assessment and management of chemical contamination in European surface water resources. Sci Total Environ. 2016 1;576:720. doi:10.1016/j.scitotenv.2016.10.104.

[14] Oldenkamp R, Hamers T, Wilkinson J, Slootweg J, Posthuma L. Regulatory Risk Assessment of Pharmaceuticals in the Environment: Current Practice and Future Priorities. Environ Toxicol Chem. 2024 3;43(3):611-22. doi:10.1002/etc.5535.

[15] Hollender J, van Bavel B, Dulio V, Farmen E, Furtmann K, Koschorreck J, et al. High resolution mass spectrometry-based non-target screening can support regulatory environmental monitoring and chemicals management. Environ Sci Eur. 2019 7;31(1):42. doi:10.1186/s12302-019-0225-x.

[16] Tkalec Z, Antignac J. P, Bandow N, Béen F. M, Belova L, Bessems J, et al. Innovative analytical methodologies for characterizing chemical exposure with a view to next-generation risk assessment. Environ Int. 2024 4;186:108585. doi:10.1016/j.envint.2024.108585.

[17] Tian Z, Zhao H, Peter K. T, Gonzalez M, Wetzel J, Wu C, et al. A ubiquitous tire rubber–derived chemical induces acute mortality in coho salmon. Science. 2021 1;371(6525):185-9. doi:10.1126/science.abd6951.

[18] Boahen E, Owusu L, Adjei-Anim S. O. A comprehensive review of emerging environmental contaminants of global concern. Discov Environ. 2025 9;3(1):144. doi:10.1007/s44274-025-00259-x.

[19] Schymanski E. L, Jeon J, Gulde R, Fenner K, Ruff M, Singer H. P, et al. Identifying small molecules via high resolution mass spectrometry: Communicating confidence. Environ Sci Technol. 2014 2;48:2097-8. doi:10.1021/es5002105.

[20] Menger F, Gago-Ferrero P, Wiberg K, Ahrens L. Wide-scope screening of polar contaminants of concern in water: A critical review of liquid chromatography-high resolution mass spectrometry-based strategies. Trends Environ Anal Chem. 2020 12;28:e00102. doi:10.1016/j.teac.2020.e00102.

[21] Hollender J, Schymanski E. L, Ahrens L, Alygizakis N, Béen F, Bijlsma L, et al. NORMAN guidance on suspect and non-target screening in environmental monitoring. Environ Sci Eur. 2023 9;35(1):75. doi:10.1186/s12302-023-00779-4.

[22] Ciccarelli D, Samanipour S, Rapp-Wright H, Bieber S, Letzel T, O'Brien J. W, et al. Bridging knowledge gaps in human chemical exposure via drinking water with non-target screening. Crit Rev Environ Sci Technol. 2024;55(3):190-214. doi:10.1080/10643389.2024.2396690.

[23] Schulze B, Heffernan A. L, Samanipour S, Gomez Ramos M. J, Veal C, Thomas K. V, et al. Is Nontarget Analysis Ready for Regulatory Application? Influence of Peak-Picking Algorithms on Data Analysis. Anal Chem. 2023 12;95(50):18361-9. doi:10.1021/acs.analchem.3C03003.

[24] Schmidt T. C, Renner G, Samanipour S. Computational mass spectrometry for exposomics in non-target screening. Anal Bioanal Chem. 2025 11;417(27):6031-2. doi:10.1007/s00216-025-06093-7.

[25] Cooper B, Yang R. An assessment of AcquireX and Compound Discoverer software 3.3 for non-targeted metabolomics. Sci Rep. 2024 2;14(1):4841. doi:10.1038/s41598-024-55356-3.

[26] Schmid R, Heuckeroth S, Korf A, Smirnov A, Myers O, Dyrlund T. S, et al. Integrative analysis of multimodal mass spectrometry data in MZmine 3. Nat Biotechnol. 2023 3;41(4):447-9. doi:10.1038/s41587-023-01690-2.

[27] Takeda H, Matsuzawa Y, Takeuchi M, Takahashi M, Nishida K, Harayama T, et al. MS-DIAL 5 multimodal mass spectrometry data mining unveils lipidome complexities. Nat Commun. 2024 11;15(1):9903. doi:10.1038/s41467-024-54137-w.

[28] Pfeuffer J, Bielow C, Wein S, Jeong K, Netz E, Walter A, et al. OpenMS 3 enables reproducible analysis of large-scale mass spectrometry data. Nat Methods. 2024 2;21(3):365-7. doi:10.1038/s41592-024-02197-7.

[29] Helmus R, Velde B. v. d, Brunner A. M, Laak T. L. t, Wezel A. P. v, Schymanski E. L. patRoon 2.0: Improved non-target analysis workflows including automated transformation product screening. J Open Source Softw. 2022 3;7(71):4029. doi:10.21105/joss.04029.

[30] van Herwerden D, Nikolopoulos A, Barron L. P, O'Brien J. W, Pirok B. W. J, Thomas K. V, et al. Exploring the chemical subspace of RPLC: A data driven approach. Anal Chim Acta. 2024 8;1317:342869. doi:10.1016/j.aca.2024.342869.

[31] Rutz A, Bittremieux W, Schmid R, Cailloux O, van der Hooft J. J. J, Beniddir M. A. Scalability of mass spectrometry-based metabolomics for natural extracts libraries exploration: current status, challenges, and opportunities. Nat Prod Rep. 2025. doi:10.1039/d5np00034c.

[32] Hulleman T, Turkina V, O'Brien J. W, Chojnacka A, Thomas K. V, Samanipour S. Critical Assessment of the Chemical Space Covered by LC-HRMS Non-Targeted Analysis. Environ Sci Technol. 2023 9;57(38):14101-12. doi:10.1021/acs.est.3c03606.

[33] Alygizakis N, Lestremau F, Gago-Ferrero P, Gil-Solsona R, Arturi K, Hollender J, et al. Towards a harmonized identification scoring system in LC-HRMS/MS based non-target screening (NTS) of emerging contaminants. Trends Anal Chem. 2023 2;159:116944. doi:10.1016/j.trac.2023.116944.

[34] Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, et al. MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom. 2010 7;45(7):703-14. doi:10.1002/jms.1777.

[35] Padilla-González F, Rizzo S, Dirks C, Bergkamp W, Rasker S, Aloisi I. Creation of an Open-Access High-Resolution Tandem Mass Spectral Library of 1000 Food Toxicants. Anal Chem. 2025 11;97(43):23822-30. doi:10.1021/acs.analchem.5C03020.

[36] Wang Y, Wang Y, Zhang Z, Xu K, Fang Q, Wu X, et al. Molecular networking: An efficient tool for discovering and identifying natural products. J Pharm Biomed Anal. 2025 7;259:116741. doi:10.1016/j.jpba.2025.116741.

[37] Schmid R, Petras D, Nothias L. F, Wang M, Aron A. T, Jagels A, et al. Ion identity molecular networking for mass spectrometry-based metabolomics in the GNPS environment. Nat Commun. 2021 6;12(1):1-12. doi:10.1038/s41467-021-23953-9.

[38] Coler E. A, Melnik A, Lotfi A, Boginski V, Semenov A, Correspondence A. A. A, et al. Ordering molecular diversity in untargeted metabolomics via molecular community networking. Cell Rep Methods. 2026 6;0(0):101468. doi:10.1016/j.crmrth.2026.101468.

[39] Manz K. E, Feerick A, Braun J. M, Feng Y. L, Hall A, Koelmel J, et al. Non-targeted analysis (NTA) and suspect screening analysis (SSA): a review of examining the chemical exposome. J Expo Sci Environ Epidemiol. 2023 6;33(4):524-36. doi:10.1038/s41370-023-00574-6.

[40] Ngan H. L, Turkina V, van Herwerden D, Yan H, Cai Z, Samanipour S. Machine Learning for Enhanced Identification Probability in RPLC/HRMS Nontargeted Workflows. Anal Chem. 2025 8;97(33):18028-35. doi:10.1021/acs.analchem.5c01873.

[41] Dührkop K, Fleischauer M, Ludwig M, Aksenov A. A, Melnik A. V, Meusel M, et al. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. Nat Methods. 2019 3;16(4):299-302. doi:10.1038/s41592-019-0344-8.

[42] Bushuiev R, Bushuiev A, Samusevich R, Brungs C, Sivic J, Pluskal T. Self-supervised learning of molecular representations from millions of tandem mass spectra using DreaMS. Nat Biotechnol. 2025 5;44(4):630-40. doi:10.1038/s41587-025-02663-3.

[43] Boelrijk J, van Herwerden D, Ensing B, Forré P, Samanipour S. Predicting RP-LC retention indices of structurally unknown chemicals from mass spectrometry data. J Cheminform. 2023 2;15(1):28. doi:10.1186/s13321-023-00699-8.

[44] Arturi K, Harris E. J, Gasser L, Escher B. I, Braun G, Bosshard R, et al. MLinvitroTox reloaded for high-throughput hazard-based prioritization of high-resolution mass spectrometry data. J Cheminform. 2025 12;17(1):14. doi:10.1186/s13321-025-00950-4.

[45] Peets P, Wang W. C, Macleod M, Breitholtz M, Martin J. W, Kruve A. MS2Tox Machine Learning Tool for Predicting the Ecotoxicity of Unidentified Chemicals in Water by Nontarget LC-HRMS. Environ Sci Technol. 2022 11;56(22):15508-17. doi:10.1021/acs.est.2c02536.

[46] Turkina V, Gringhuis J. T, Boot S, Petrignani A, Corthals G, Praetorius A, et al. Prioritization of Unknown LC-HRMS Features Based on Predicted Toxicity Categories. Environ Sci Technol. 2025 4;59(16):8004-15. doi:10.1021/acs.est.4c13026.

[47] Scheubert K, Hufsky F, Böcker S. Computational mass spectrometry for small molecules. J Cheminform. 2013 3;5(3):12. doi:10.1186/1758-2946-5-12.

[48] Tsugawa H, Cajka T, Kind T, Ma Y, Higgins B, Ikeda K, et al. MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. Nat Methods. 2015 5;12(6):523-6. doi:10.1038/nmeth.3393.

[49] Bittremieux W, Schmid R, Huber F, Van Der Hooft J. J. J, Wang M, Dorrestein P. C. Comparison of Cosine, Modified Cosine, and Neutral Loss Based Spectrum Alignment For Discovery of Structurally Related Molecules. J Am Soc Mass Spectrom. 2022 9;33(9):1733-44. doi:10.1021/jasms.2C00153.

[50] Carracedo-Reboredo P, Liñares-Blanco J, Rodríguez-Fernández N, Cedrón F, Novoa F. J, Carballal A, et al. A review on machine learning approaches and trends in drug discovery. Comput Struct Biotechnol J. 2021 1;19:4538-58. doi:10.1016/j.csbj.2021.08.011.

[51] Cherkasov A, Muratov E. N, Fourches D, Varnek A, Baskin I. I, Cronin M, et al. QSAR Modeling: Where Have You Been? Where Are You Going To? J Med Chem. 2014 6;57(12):4977-5010. doi:10.1021/jm4004285.

[52] Samanipour S, O'Brien J. W, Reid M. J, Thomas K. V, Praetorius A. From Molecular Descriptors to Intrinsic Fish Toxicity of Chemicals: An Alternative Approach to Chemical Prioritization. Environ Sci Technol. 2023 11;57(46):17950-8. doi:10.1021/acs.est.2c07353.

[53] Bach E, Schymanski E. L, Rousu J. Joint structural annotation of small molecules using liquid chromatography retention order and tandem mass spectrometry data. Nat Mach Intell. 2022 12;4(12):1224-37. doi:10.1038/s42256-022-00577-2.

[54] Witting M, Böcker S. Current status of retention time prediction in metabolite identification. J Sep Sci. 2020 5;43(9-10):1746-54. doi:10.1002/jssc.202000060.

[55] Scheubert K, Hufsky F, Petras D, Wang M, Nothias L. F, Dührkop K, et al. Significance estimation for large scale metabolomics annotations by spectral matching. Nat Commun. 2017 11;8(1):1494. doi:10.1038/s41467-017-01318-5.

[56] Lee S, Park H, Kim H. False discovery rate estimation using candidate peptides for each spectrum. BMC Bioinform. 2022 12;23(1):454. doi:10.1186/s12859-022-05002-4.

2

Critical assessment of the chemical space covered by LC-HRMS non-targeted analysis

Authors: Viktoriia Turkina*, Tobias Hulleman*, Jake W. O'Brien, Aleksandra Chojnacka, Kevin V. Thomas, Saer Samanipour (*equal contribution)

Published in: Environ. Sci. Technol. 2023, 57, 38, 14101–14112.

https://doi.org/10.1021/acs.est.3c03606

Abstract

Non-targeted analysis (NTA) has emerged as a valuable approach for comprehensive monitoring of chemicals of emerging concern (CECs) in the exposome. The NTA approach, theoretically, is able to identify compounds with diverse physicochemical properties and sources. Non-targeted analysis methods, even though generic and wide scoping, have been shown to have limitations in terms of their coverage of the chemical space, as the number of the identified chemicals in each sample is very low (e.g. $\leq 5\%$). Investigating the chemical space covered by each NTA assay is crucial for understanding the limitations and challenges associated with the workflow from experimental methods to the data acquisition and data processing. In this review, we examined recent NTA studies published between 2017 and 2023 that employed liquid chromatography-high resolution mass spectrometry. The parameters used in each study were documented and reported chemicals at the confidence level 1 and 2 were retrieved. The chosen experimental setups and the quality of reporting were critically evaluated and discussed. The findings revealed that only around 2% of the estimated chemical space was covered by the NTA studies investigated. Little to no trend was found between the experimental setup and the observed coverage, due to the generic and wide scope of NTA studies. The limited coverage of chemical space by the NTA studies highlights the necessity for a more comprehensive approach in experimental and data processing setups to enable the exploration of a broader range of chemical space, with the ultimate goal of protecting human and environmental health. Recommendations to further explore a wider range of the chemical space were given.

2.1 Introduction

The exposome is the measure of all the exposures, both chemical and non-chemical, of an individual in a lifetime and how those exposures relate to health. [1] The chemical space of exposome refers to the chemical space relevant to human and environmental exposure. [2–4] In contrast, the chemical space in general refers to all possible organic structures that are plausible from an organic chemistry point of view. [2, 5] Theoretical estimates of such structures suggests there are around 10^{60} unique structures with molecular weights less than 500 Da. [6, 7] This theoretical chemical space incorporates both known and unknown unknowns [2, 8] and may include structures that can cause adverse effects depending on their exposure levels. In fact, when looking at the known unknowns (i.e. structures recorded in the chemical databases, but not initially known to be present in the sample), several of them have been shown to have adverse effects on environmental and human health. [8–11]

Chemical prioritization has been one of the main means for dealing with the diversity of chemical space in the human and environmental exposome. [3, 12, 13] This consists of exploration of the literature for measured chemicals and their properties/toxicities as well as national/international chemical registries. [14, 15] A combination of predicted properties and toxicity is used to rank chemicals in the databases based on their potential impact on the environment and human health. [16] Chemicals with a high potential of such impact are considered as chemicals of emerging concern (CECs). [17,18] To facilitate chemical prioritization, several databases consisting of chemical structures, the associated physicochemical properties (both measured and predicted), and their biological activities have been made publicly available (e.g. PubChem, [19] NORMAN Databases, [20] and CompTox [14]). However, most of these known unknowns remain unmeasured in environmental and biological matrices due to difficulties associated with the inclusion of such a large number of chemicals in routine monitoring programs. [11, 13]

Non targeted analysis (NTA) combined with liquid chromatography coupled with high resolution mass spectrometry (LC-HRMS) is considered as one of the most comprehensive methods for the detection and identification of known and unknown unknowns in complex environmental and biological samples. [21, 22] This approach utilizes a generic and wide scope strategy for the sample preparation and analysis to maximize the coverage of

the chemical space of the sample. [2, 13, 21, 23–31] This typically results in very large and complex datasets (e.g. 5 GB per sample) that must be pre-processed prior to the identification workflow. [31–33] The NTA data processing workflows include several steps from data conversion to library search and the confidence assessment of the candidate spectra. [2, 23, 26–29] Due to the complexity of such datasets and sheer size of the chemical databases, the NTA workflows are not very sensitive and do not result in a high percentage of identified chromatographic features. [34, 35] A more sensitive but less comprehensive data processing alternative is suspect screening where the chemicals of interest are known prior to the data processing workflow. This approach is more sensitive in terms of limits of detection but is unable to detect unknown unknowns. [20, 29, 36] These two strategies are commonly employed together for the screening of complex environmental and biological samples. [23]

The NTA strategy, even though powerful, has not been widely accepted within the regulatory framework due to reproduciblity issues. [30, 34, 37] Recent studies have indicated that small changes in both experimental (e.g. data dependent vs data independent acquisition) and data processing parameters may result in different outcomes and thus conclusions. [34, 35] Additionally, a recent study has postulated the potential impact of different experimental parameters on the measured chemical space. [38] In fact, the aforementioned issues with NTA assays have sparked a debate in the scientific community and have given start to a new wave of data processing tools development. [25, 39, 40] Additionally, several efforts have been put into better defining the much needed quality control and assurance for such experiments to be successful in detection and identification of the known and unknown unknowns in complex environmental samples, thus better understanding the coverage of the analyzed chemical space. [23, 38, 41–44]

Several recently published reviews discuss in detail the impact of different steps on the chemical space coverage through different experimental approaches. [2, 23, 26, 27] They cover both data processing and experimental parameters including study scope, sampling and sample treatment, instrumental conditions, data processing and treatment, and reporting. However, none of these reviews attempted to assess (i.e. quantify) the coverage of the identified chemical space reached by the already conducted NTA environmental studies. Quantification of the coverage of chemical space by an analytical method is not a trivial task. Theoretically, it can be quantified as the number of identified compounds in the given sample divided by the number of all compounds present in the chemical subspace of the

sample. But practically, this calculation is impossible, due to the complex chemical nature of samples and the number of unknown constituents. Nevertheless, the investigation of experimentally explored chemical space is highly relevant for the researchers to be aware of the limited coverage of the associated chemical space.

In this review, we aim to quantify the coverage of the identified chemical space by recent environmental studies and investigate the relationship between the selected experimental parameters and the explored chemical space. To quantify the covered chemical space via NTA, we collected all recent studies that perform NTA (not suspect screening) and reported levels 1 and 2, in terms of identification, structures. [45] Additionally, we limited the scope of this study to semi-polar and polar chemicals analyzable with liquid chromatography coupled with high resolution mass spectrometry, resulting in a total of 61 papers. As an approximation of the chemical space the NORMAN SusDat database containing around 60k unique chemicals with available PubChem CIDs (compound ID number) was used (`https://www.NORMAN-network.com/nds/susdat/susdatSearchShow.php`). We collected a list of experimental and instrumental parameters, including sample preparation (i.e. storage and extraction conditions), chromatographic separation (e.g. eluents, gradient type, and injection volume), high resolution mass spectrometry settings (e.g. mass analyzer, data acquisition mode, and polarity), and data processing workflows (e.g. mass and retention time tolerance, retention time domain alignment and databases used for the search). We also noted any unreported parameters to identify the most commonly omitted settings. Furthermore, we extracted information on the scope of the studies and samples analyzed.

Finally, we estimated the coverage of chemical space explored by recent NTA studies by comparing the structures identified in these studies with the chemical space represented by the compounds in the NORMAN SusDat database, as shown in Figure 2.1. This figure provides an insight of the range of chemicals that may be present in environmental samples. To our knowledge, this is the first study "quantifying" the coverage of chemical space via NTA assays.

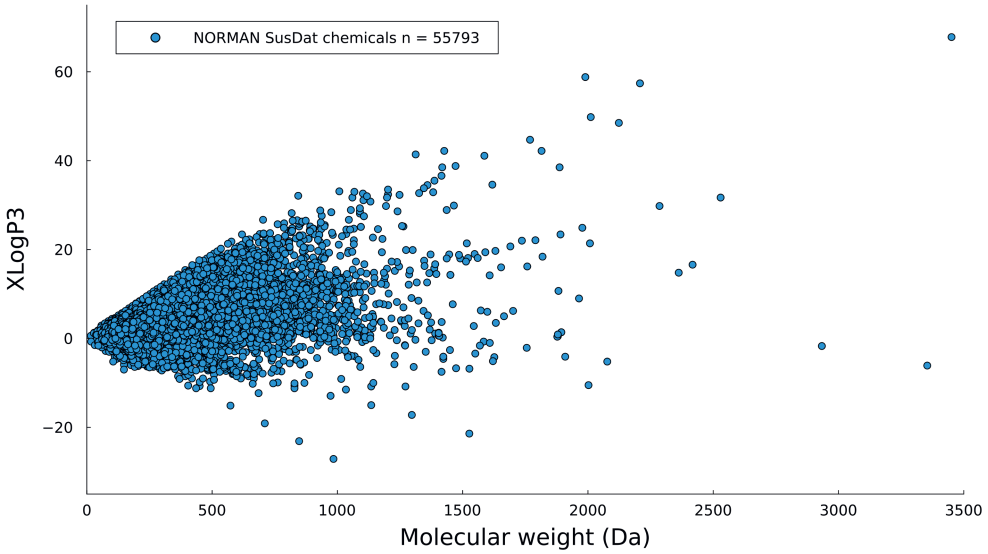


Figure 2.1: Distribution of all chemicals in the NORMAN SusDat database (n = 55793) based on their molecular weights (Da) and logP values.

2.2 Methods

2.2.1 Selection of NTA studies

This review is particularly focused on the development of the NTA approach in environmental studies, specifically after the discussions regarding reproducibility were initiated. [39] Thus, we used the citation database Scopus to search for relevant studies published from 2017 to 2023 in the field of non-target analysis (NTA) with a focus on environmental science. The search was limited to articles that contained the keywords "non targeted analysis", "non target analysis", "untargeted analysis", "untargeted screening", or "non-target screening" while excluding articles containing "metabolomics", "metabolic", or "gas chromatography". This initial search resulted in 377 publications adhering to the search parameters, which were then manually filtered to include only those that met a specific set of criteria.

The first criterion was that articles used non-target analysis to probe chemicals of emerging concern, preferably in environmental matrices. Secondly, the publications had to use a non-target workflow. Some articles included the desired keywords in the

title or abstract but were actually targeted studies with a very extensive list of target chemicals. For the same reason, the third criterion was that studies used LC-HRMS for sample analysis. Studies, which conducted GC-HRMS analysis were not included in the review, since such studies mainly employ suspect screening rather than non-targeted analysis. Furthermore, the recent development of NTA has been focused on LC rather than GC. Therefore, this review is focused on the coverage of the chemical space by NTA conducted via LC-HRMS. Additionally, direct infusion studies, studies that used rare setups, or heavily modified setups were excluded. Finally, review articles and studies which did not perform any identification were excluded as they did not contribute any additional methods or identified compounds. The search for relevant studies meeting these criteria was completed on March 1st, 2023, resulting in the inclusion of 61 studies in this review. [46]

2.2.2 Collection of instrumental parameters

To capture the impact of each step of the NTA workflow on chemical space coverage, we extracted specific parameters used in the studies we reviewed. Sample preparation, chromatographic separation, data acquisition, and data processing were the four main steps where parameters were identified. Sample preparation parameters included the sample matrix, storage conditions, pre-storage modifications, extraction methods, and extraction conditions where applicable. Chromatographic separation parameters included the column used, eluent composition, gradient complexity, number of column volumes, column temperature, and injection volume. Gradients were classified as linear, semi-linear, or complex based on their complexity. The number of column volumes refers to the volume of solvent that passes through a chromatography column relative to the volume of the column itself. The calculation was performed using the equation 2.1.

$$Column\ volumes = \frac{F \times T_{run}}{\pi \times (\frac{dc}{2})^2 \times L} \tag{2.1}$$

Where *F* is the flow rate (mL/min), T_{run} is the total run time of the method (min) - excluding equilibration time, *dc* is the internal diameter of the column (cm), and *L* is the length of the column in (cm). HRMS instrumental parameters included the mass analyzer,

sampling rate (in the case of Q-TOF), resolution (in the case of Orbitrap), data acquisition mode, polarity, and mass range. Data processing parameters included mass tolerance, time domain alignment, mass calibration, retention time tolerance, databases used, and total database size (labeled as small if $\leq$1000 compounds or large if $>$1000 compounds). A summary of the collected parameters can be found in Figure 2.2. Furthermore, we made note of parameters that were not reported in order to identify which settings were commonly omitted. Lastly, we gathered information on the scope of the studies. The collected parameters along with the list of the publications are publicly available through this link `https://doi.org/10.5281/zenodo.8249216`. [46]

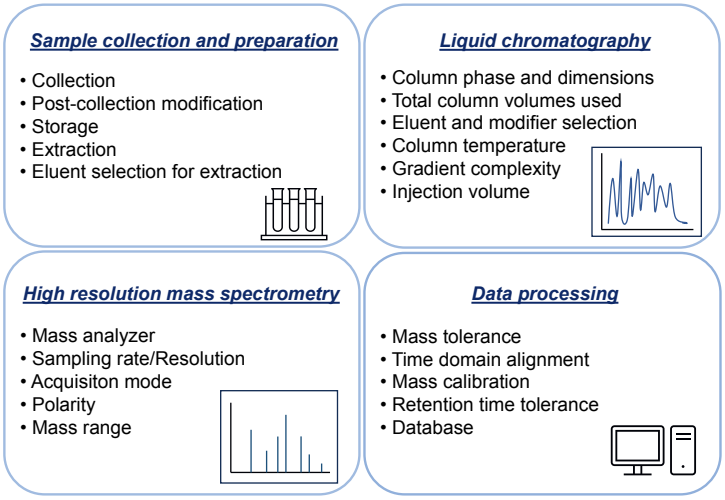


Figure 2.2: Summary of the main instrumental parameters collected from the reviewed NTA studies

2.2.3 Collection of reported structures

To assess the extent of chemical space coverage by recent NTA studies, we extracted the reported structures. To ensure the reliability and accuracy of our analysis, we only included structures identified with a high level of confidence (i.e levels one and two on the Schymanski scale), which are less susceptible to false positive identifications. [45] For each compound, SMILES, IUPAC name, and the regular names provided by the authors were extracted. Finally, we excluded articles from our chemical space coverage assessment if the authors did not specify the identification level, did not include the identified compounds in either the article or supplementary materials, or only reported compounds

within their target list.

2.2.4 Data processing

The list of the collected compounds was stored in CSV format, and Julia version 1.7 was used to import and process the data. A modified version of the PubChemCrawler.jl package was employed to retrieve chemical data such as XLogP3 and MW of the compounds from the PubChem database by using their available identifiers (SMILES, IUPAC, InChIKey, or regular name). [47] logP values extracted from PubChem are generated using XLogP3 with an additive model starting from a reference compound. [48] Retrieved data along with the collected experimental parameters were combined into a dataset that included PubChem CIDs corresponding to the compounds, their logP values, molecular weights, and experimental parameters.

For the evaluation of the chemical space coverage, we additionally calculated elemental mass defects (EMD) of six elemental ratios (CO, CCl, CN, CS, CF, and CH) for each collected compound and the ones included in the NORMAN SusDat database. [49] EMD values were used to cluster structurally similar compounds together and separate others, as they incorporate structural information and are used to compare compounds based on their elemental composition. [50] The combination of logP, MW, and EMDs was used for principal component analysis (PCA), which is an unsupervised algorithm for dimensional reduction combining variables into principal components. [51] This approach is able to identify trends and clusters in the data sets. Prior to the analysis, the data were mean-centred and scaled to keep the initial weight of all variables comparable. PCA was performed using the ScikitLearn.jl julia package, and in total, three principal components were utilized.

The NORMAN SusDat database was used for the approximation of the chemical space of environmental samples. While the chemical space comprises both known and unknown compounds, it is practically impossible to include the latter in our approximations. The Norman SusDat database includes CECs that have either been detected in various environmental compartments or have been identified as potential CECs, providing a comprehensive set of chemicals with a wide coverage of physical and chemical properties and structures. [20]

Finally, the classes of the collected compounds were defined to illustrate the frequency

of identification of specific classes. To obtain the class of each CEC, the corresponding InChIKey was used to generate information on superclasses, classes, and sub-classes of each compound via ClassyFire. ClassyFire divides a given chemical compound into classes based on its structural features (i.e. functional groups). [52]

2.3 Discussion

In this review, we estimated the coverage of the chemical space of environmental samples by investigating recent NTA studies. To evaluate the impact of selected workflow parameters on the coverage of chemical space, we collected information on these parameters (e.g. mass analyzer, data acquisition mode, ionization mode and size of the database used) from the studies. [46] The identified compounds were categorized into classes and their relative frequency of occurrence was determined. XLogP3, MW, and EMDs were used to represent the vastness of the chemical space, approximated with the NORMAN SusDat Database. PCA was employed to illustrate the coverage of the space of chemicals detected in recent environmental studies.

2.3.1 Overview of the studies

In total, 61 studies were collected, with 55 of them published since 2020. Only studies using NTA were included, while those using screening or targeted approaches but claiming to be untargeted were excluded. This indicates that $\approx 90\%$ of the reviewed studies were published in the last three years, yielding an average of more than 15 studies per year. In contrast, during the period from 2017 to 2020, only six of the selected studies were published. Therefore, the significant increase in the number of such studies in recent years reflects the successful applications of NTA workflows in exposome analysis. The scope of these studies varies, with 30 studies focusing on a wide range of chemicals and another 21 studies specifically targeting groups among which are per- and polyfluoroalkyl substances (PFAS), pesticides, pharmaceuticals, and illicit drugs. Such prior prioritization influences the choice of experimental setup. The remaining 10 studies focused on NTA workflow development, indicating a growing interest and the need for further advancements in this field.

2.3.2 Overview of selected parameters

Sample collection and preparation

The collection and preparation of samples in the non-targeted analysis (NTA) workflow can introduce potential sources of loss of chemical information. Issues such as ensuring sample representativeness (e.g. selecting appropriate grab or passive sampling techniques), addressing potential sample contamination, accounting for matrix effects, optimizing extraction methods for selectivity, and avoiding bias towards specific chemical groups are important considerations in NTA. [2,23,26] These challenges may impact the accuracy and reliability of NTA results, potentially affecting the comprehensiveness and quality of the chemical information obtained from the analysis. Therefore, careful attention to sample collection and preparation steps are essential to minimize potential sources of bias and ensure robust and reliable NTA outcomes.

The majority of the collected studies (67%) analyzed water samples (n = 41). Other matrices that were investigated include biota (n = 5), dust (n = 3), urine (n = 3), atmospheric particulate matter (n = 2), paper (n = 2), serum (n = 2), blood (n = 1), human hair (n = 1), ovarian follicular fluid (n = 1), sewage sludge (n = 1), snow (n = 1), and surface soil (n = 1), turtle tissue (n = 1).

To prevent microbiological growth, the studies on water samples reported a conservation step, which involved either adding an acid or storing the sample at a temperature of -20°C or 4°C. Out of the 41 water studies, 7 studies either did not include a step to stop microbiological growth or did not report it. If this step was omitted, it could significantly alter the sample's final composition when it is eventually analyzed in the laboratory. [53, 54]

Around 54% of publications analyzing water included a sample filtering step prior to analysis. This step is a compromise to preserve the LC system and column but may lead to the loss of the chemicals adsorbed to the particle's surface. Approximately 67% of studies included solid phase extraction (SPE) in their sample preparation, out of which 73% used reversed-phase hydrophilic-lipophilic balance (HLB) SPE. However, only 29% of studies with SPE used acidic and/or basic modifiers in the extraction eluents. That implies that most studies using only HLB SPE are potentially leaving ionizable compounds on the sorbent and may exclude them from the analysis. The remaining studies employed alternative preptreatment techniques among which vacuum-assisted evaporation, centrifu-

gation, liquid-liquid extraction (LLE), ultrasonic extraction as well as their combination. These choices are mostly dictated by the sample nature/matrix. There were three studies that performed no sample extraction and injected directly into the LC-MS with a higher injection volume. [55–57] While this protocol minimizes sample adulteration and keeps the sampling of chemical space more comprehensive, it can also pose a challenge to detection sensitivity due to the low analyte concentration. [23]

Overall the sample collection and preparation section is well reported in the selected studies. However, many of the studies focused on analyzing a wide range of chemicals do not explore alternative extraction methods to ensure a more comprehensive coverage of the chemical space. This could result in a bias towards specific compounds, rather than capturing a more diverse set of chemicals.

Liquid chromatography

Chromatographic separation is employed to minimize sample complexity by spreading analytes across the time axis. This helps to reduce ion suppression (matrix effect) and provides additional information (retention time) for the identification of the analytes. The chemistry of the stationary phase along with the elution conditions affects the quality of separation and the type of analytes being retained. Thus, the selection of chromatographic conditions heavily influences the coverage of the chemical space of the sample. [23]

The majority of NTA studies use conventional reverse-phase separation with a generic C18 column. Optimization of the separation includes proper selection of eluents and modifiers, including suitable elution power and gradient setup, to avoid co-elution and excessive or insufficient retention of chemicals. [58] A simple linear gradient of an aqueous phase and methanol or acetonitrile from low to high percentage is most widely accepted for the wide scope screening. This method proved its reproducibility across different scopes of the studies. [26] However, this strategy focuses on polar to semipolar compounds, potentially excluding very polar (i.e logP smaller than -2) and very hydrophobic substances (i.e. logP larger 6) from the comprehensive investigation of the chemical composition of samples. [59] To cover the polar part of the chemical space, orthogonal methods such as hydrophilic interaction chromatography (HILIC) become more popular, while for hydrophobic volatile chemicals, GC is a widely used technique. [55, 60, 61] Finally, to ensure the reproducibility and reliability of the studies parameters such as injection

volume and column temperature, should be properly reported. [62]

More than 90% of the collected studies used a C18 column for the separation, among which almost all were endcapped with a column length of 50 mm (20%), 100 mm (49%), or 150 mm (30%). Column diameters were either 2.1 mm (78% of the studies), 3 mm (16%), or 2 mm (3%) with the particle diameter under 3.5 μL. Additionally, two different studies reported 4.6 mm and 0.05 mm column diameters. Although applying a simple gradient ensures higher reproducibility of the method, only half of the studies ($\approx 51\%$) used a linear gradient, while around 32% used a semi-linear gradient, and the remaining (18%) used a more complex type of gradient.

The median number of column volumes eluted in the studies is 15.9, with an interquartile range of 15.9. The use of a sufficient number of column volumes should ensure the complete elution of most hydrophobic compounds (high logP and MW) and the absence of carryover. The optimal number depends on the stationary phase, eluent power, and the analytes themselves. [63] Nevertheless, the widely accepted hypothesis is that there is a linear relationship between logP and retention/number of column volumes used. The hypothesis is applied for the reverse phase mode with comparable C18 selectivity, similar gradients, and eluent composition. [64, 65] However, our results do not indicate the presence of a linear relationship between the number of column volumes and logP of the chemicals, since no clear linear pattern could be identified between these parameters (Section 2.5 Figure S1).

In addition, the column temperatures used were all slightly above room temperature which is favorable for repeatability and reproducibility. [62] 31% of publications used 40°C, 16% used 35°C, 11% used 30°C, two studies held the column at 25°C, one at 20°C, one at 45°C and one at 50°C. About 29% of papers did not report the column temperature, which hinders the reproducibility of experiments.

Finally, 18% of the studies did not report the injection volume used. Injection volume should not have a large effect on the final observed chemical space as they depend on the extraction method and efficiencies. Nevertheless, the success of the method's transfer depends on it. Most of the studies used either 5 (n = 17) or 10 μL (n = 13) injection volume, which is adequate when using SPE extraction. The remaining were spread across 1, 3, 4, 7, 20, 100, 140, and 660 μL.

To conclude, despite the rising discussion about reporting quality, [23, 66–68] chro-

matographic separation parameters in the collected studies were not always properly reported. Proper harmonized reporting ensures successful method transfer, whereas inconsistent reporting raises questions related to the reproducibility of the study, reliability of the results, and the possibility of retrospective studies. While the majority of the studies seek to comprehensively investigate the chemical composition of the samples, only approximately 10% employ an alternative to the conventional approach to analyze the samples. Lastly, the hypothetical linear trend between logP and retention was not confirmed, indicating the need for more sophisticated strategies for method development and optimization.

High resolution mass spectrometry

The Orbitrap and the quadrupole time of flight (QTOF) equipped with electrospray ionization (ESI) are the two most commonly used HRMS instruments in liquid chromatography-based (LC) NTA experiments. For complimentary analysis, it is recommended to perform separate experiments in both positive and negative modes. [69] The mass resolution of Orbitrap mass analyzers is generally higher than that of QTOF, but both can provide high-resolution mass spectra (Resolution $\geq 30,000$). [70]

In QTOF, resolution is determined by the architecture of the mass analyzer, [71] while for Orbitrap, the resolving power depends on a user specified resolution. In the case of Orbitrap, the speed of scans is directly related to the spectral resolution. However, the increase in mass resolution is limited by the time required for scanning operations. For QTOF, a crucial parameter for data quality is the sampling speed, which is reported as spectra per second in Hz. If the scan rate is too high, fewer ions are sampled, which can lead to a sensitivity issue. Conversely, if the scan rate is too low, fewer data points on the time axis are recorded, potentially causing missed detection of analytes eluting in a narrow time range. [72]

MS^2 spectra for structure elucidation are recorded using either data-dependent acquisition (DDA) or data-independent acquisition (DIA). DDA mode records fragments of preselected precursor ions (which are chosen based on their abundance or via an inclusion list) while DIA mode fragments all precursor ions within a certain mass range. The latter is preferable for comprehensive investigations of complex samples. However, DDA mode is currently the preferred choice in environmental studies, partly due to the limited availability of processing tools for DIA files and also because the DIA experimental setup is not com-

monly employed with Orbitrap mass analyzers. [26] QTOF analyzers are more commonly used for DIA due to higher data acquisition rates.

Roughly, half of the collected studies (n = 31) utilized an Orbitrap mass analyzer, while the other half employed a QTOF mass analyzer. However, a significant proportion (approximately 74%) of the studies reported using DDA, which inherently limits their results to predefined ions. The scan rate for QTOF analyzers was mostly set at 4 Hz, although some studies operated at lower rates of 3, 2, or 1 Hz. Many studies using Orbitrap analyzers operated at a resolution of 70,000, while some studies used lower resolutions with a minimum of 35,000 and higher with a maximum of 240,000. Approximately 22% of the studies did not report either resolution or scan rate.

Less than half of the studies (around 42%) conducted separate experiments in positive and negative modes, utilizing multiple injections, different modifiers, and sometimes different columns, which is considered a more suitable scenario for achieving comprehensive coverage of chemical space. In approximately 30% of the studies, MS was operated only in positive mode. There were eleven publications where the analysis was reported in both modes, but the details were insufficient to determine if the experiment was performed simultaneously or separately in both modes. In three other studies, an exclusively negative mode was used to prioritize a specific group of compounds of interest, such as PFAS, [73–75] deliberately narrowing down the investigated chemical space. Finally, two of the reviewed studies employed simultaneous positive and negative ionization modes with formic acid as a modifier. This approach is not preferable for NTA given that acidic additives are not always optimal for a negative ionization mode. Additionally, the acquired data becomes extremely complex and lacks quality for reliable and robust processing.

The selected mass range in the collected studies is between 50 and 1200 m/z, which is based on approximated chemical space covering the largest part. However, some studies set their maximum m/z at 1000 or lower, which leads to the exclusion of the part of chemical space with higher MW.

To conclude, despite recent advancements in DIA technology, DDA remains the predominant choice in the reviewed studies. However, the recommended approach for improved reproducibility and reliability of NTA studies, and to enhance coverage of chemical space in environmental and metabolomics research is to acquire data in DIA mode for initial screening and then continue with DDA for individual feature identification. [23,76–78]

Finally, in terms of reproducibility, the lack of comprehensively reported information hinders method transfer and therefore it warrants actions towards a harmonized reporting strategy. [66–68]

Data processing

Data processing is considered a major bottleneck in NTA workflows. It refers to a series of procedures that starts with the data conversion and ends with the feature identification. [23] One of the steps for reliable processing is the mass calibration, either external or internal. During this step the measured m/z values of known structures are compared against theoretical m/z values. These shifts/correction factors are applied to all mass channels, depending on the instrumental setup. This step ensures the quality of the spectra in terms of accurate mass measurement. [79] An inadequate mass calibration may result in false positive and negative detections during the identification. [80]

One of the last steps of CEC identification is the use of a database to relate the MS output to a known chemical structure. To proceed with the identification, experimental data undergo pre-processing steps: data compression, to remove noise and blank peaks, feature detection, to find features in 3-dimensional data, componentization, to group fragments and isotopologues belonging to the same compound, and feature prioritization to reduce the number of irrelevant features. [81] Since most of the collected studies used vendor software for the latter four steps, which makes it almost impossible to retrieve the information of algorithms utilized, these parameters cannot be adequately discussed for their influence on the coverage of chemical space. For the identification of known unknowns, pre-processed data is compared with chemical databases and matched against references from available spectral libraries, utilizing a combination of features, retention time, accurate mass, and fragmentation pattern. [45] The mass tolerance is the initial parameter used for the candidates' list compilation. This parameter, along with the database used, heavily affects the results of the candidate search. The number of chemicals included in databases used in the evaluated studies differs from a few hundred structures in in-house libraries [82] to tens and hundreds of thousands in publicly available libraries [27] such as NORMAN, [20] MassBank, [83] or PubChem. [19] These search algorithms result in a set of candidate structures that ultimately must be confirmed via either a reference standard and/or an orthogonal method. [45] The retention tolerance is applied for level 1 confirma-

tion employing either predicted or measured retention times. [23]

For the transparency and reproducibility of the method, proper reporting of applied setups for each data processing step is essential. Nevertheless, a significant part of the studies did not provide sufficient information to reproduce the results. Specifically, approximately 39% did not mention anything about mass calibration, while 25% reported that they performed calibration but did not describe the procedure. Only about 36% included a report on the mass calibration procedure. A large number of the papers (43%) also did not report whether a retention alignment was performed. 34% did report the fact that a retention alignment was done but did not specify the algorithm that was used or provide the details on the parameters used. The remaining 23% of publications did report both the fact that one was performed and which algorithm was used.

In contrast, mass tolerance applied for the search was reported in almost all studies, around 95%. Among which around 76% used a mass tolerance for the database query of 5 ppm, which is highly common in the NTA database search workflows. There were also studies that used a relatively high mass tolerance of 20 ppm, 17 ppm, or 10 ppm and some studies that used mass tolerances lower than 5 at 3 ppm, 2 ppm, and even 1 ppm. Generally, the studies that were using the lower mass tolerances for the database search reported a higher resolution of the mass analyzer. However, there was no clear indication of whether the mass tolerance applied to formula assignment or structural identification. On the other hand, retention tolerances had much lower reporting rates as 45% of the studies did not include this information. The remaining studies used tolerances in a range between 0.1 min and 0.5 min. However, there are a few publications that used a wider tolerance, up to 1.8 min, which may result in a high false positive rate. Finally, approximately 9% did not report the databases used or referred to the software but not the databases that the software was using. The majority, 82%, used a total database size containing more than 5000 compounds, while only 5 studies used databases with less than one thousand compounds.

The data processing step is one of the main bottlenecks for the NTA approach and thus requires greater attention within the community. Nevertheless, the reporting quality needs improvement. Furthermore, it was found that around 56% of the identified unique chemicals are available in MassBank EU. That means that roughly 44% of the HRMS spectra acquired for the identified compounds have not been deposited in public databases such as MassBank. For NTA to reach its full potential, the expansion of publicly available

spectral databases is vital for the improvement of the coverage of chemical space at the identification step.

2.3.3 Explored chemical space

The studies yielded a total of 2657 compounds reported in the identification level 1 up to 2b. The contribution to the total number from each study varies between 1 and 370, with a mean and median of 50 and 30, respectively (Section 2.5 Figure S2). Among these, 1606 compounds were identified as unique structures, accounting for $\approx 60\%$ of the total number of retrieved compounds. This finding implies that around half of the overall variety of compounds were detected more than once in various environmental compartments. However, in 8 studies, there was no report of either identification level, or any identifiers, which hinders the retrieval of the compounds from these studies. The class of each collected CEC was obtained and displayed in Figure 2.3. The most commonly found compounds were benzoids, followed by organocyclic compounds and then organic acids and derivatives. The latter category, along with organohalogen compounds, constitutes PFAS, which have been of particular interest in recent years. The median molecular weights of compounds from SusDat were 239 Da and 261 Da for those collected from the studies, with a median XLogP3 of 3.2 for SusDat and 2.2 for collected compounds. Based on histograms in Figure 2.4, compounds with the most frequently occurring properties are being identified in recent NTA studies, which can be partially explained by the generalized experimental workflows with reverse phase C18 columns.

Most of the compounds detected in the studies clustered closely together, with only a few compounds found further away from this main cluster (Figure 2.5). The collected compounds were analyzed in relation to their properties and plotted on a chemical space approximation represented by the NORMAN SusDat database. Figure 2.5 shows the plot in dimensions of molecular weight (MW) and XLogP3, which emphasizes the limited space that is currently explored using current non-target analysis workflows. To examine the effect of some of the mass spectrometry (MS) parameters used on the explored chemical space, all compounds were plotted and clustered based on factors such as the mass analyzer used, acquisition mode, ionization mode, and the total database size used (Section 2.5 Figures S3-S6). However, neither of these parameters showed an unambiguous influence on the coverage of the chemical space. It should be noted that the representation in

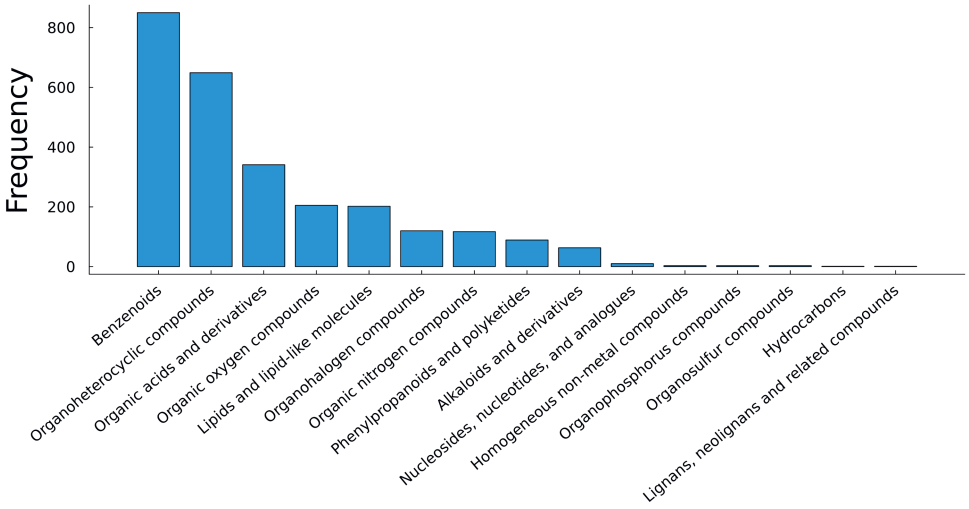


Figure 2.3: Histogram of all of the classes obtained from the Classyfire search for the detected CECs in reviewed studies

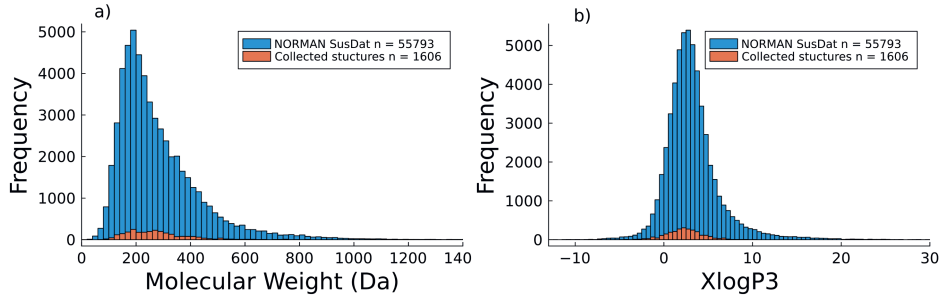


Figure 2.4: Molecular weights (a) and logP (b) distributions for the collected compounds (orange) and ones included in NORMAN SusDat database (blue).

MW and logP dimensions does not provide information about the elemental composition of compounds or their classes, which may result in an over-representation of the covered chemical space. Therefore, it is important to consider other parameters beyond MW and logP when evaluating the coverage of the chemical space by the collected structures.

The PCA scores plot in Figure 2.6 reveals that many regions of the chemical space are unexplored. The PCA was applied to the dataset combining the collected compounds with the ones from the NORMAN SusDat, with MW, XLogP3, and the EMDs as input variables. The first two principal components in the analysis were found to be primarily influenced by the elemental mass differences (EMDs) associated with compounds containing chlorine (Cl), fluorine (F), cyanide (CN), and sulfur (S). These EMDs represent the high variability in the elemental composition of the compounds and were identified as the most important variables in the PCA. This indicates that fewer compounds in the dataset contain halogens, nitrogen (N), and sulfur, while hydrogen (H), which is present in nearly every compound, does not contribute significantly to the variability in the data. The third principal component is primarily influenced by MW and XLogP3 (Section 2.5 Figure S7). In total, the first three principal components explain 74% of the variance (Section 2.5 Figure S8). In Section 2.5 Figure S9-S11, the coverage of chemical space by different compound classes is displayed. Section 2.5 Figure S11 specifically highlights the coverage by organic acids and derivatives as well as organohalogen compounds. The majority of PFAS, not exclusively, fall into these classes. The figures reveal that the distribution of compound classes across the chemical space is not homogeneous, suggesting an over-representation of certain classes of compounds. This observation can be attributed to the prior prioritization of specific classes, which may bias the identification towards those classes of compounds.

Overall only around 2% of the estimated chemical space was covered by NTA studies investigated in this review. The coverage was defined as the number of unique structures retrieved from the reviewed studies versus the number of structures in NORMAN SusDat database as an approximation of chemical space. We used NORMAN SusDat as an approximation of chemical space as it contains a set of highly relevant and curated structures for environmental and human exposome. It should be noted that this is a small subspace of the total chemical space and serves as means of approximation for the true chemical space of exposome. No clear relationship between experimental conditions and coverage of the chemical space was discovered, which may indicate that the used experimental approaches are generic enough for the NTA assays. On the other hand, this may be caused by

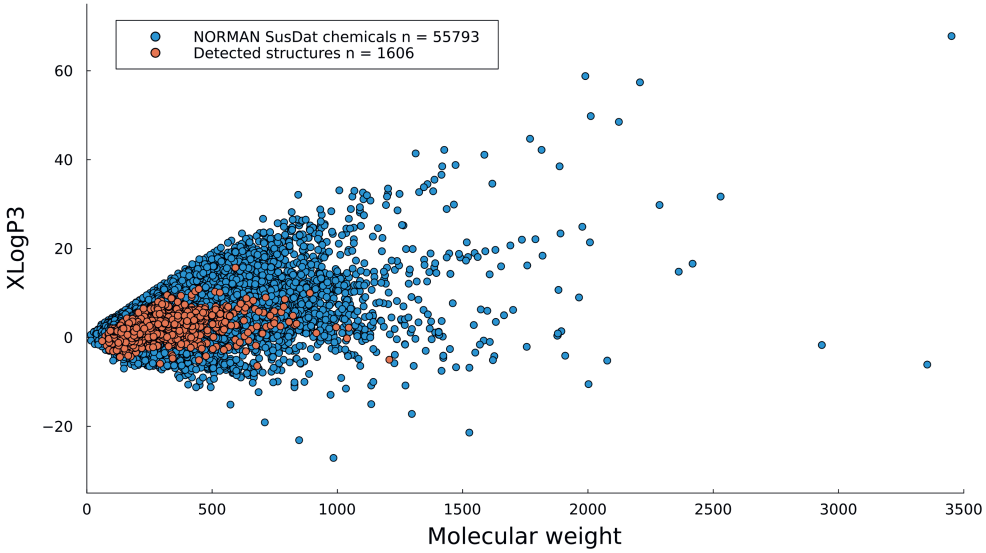


Figure 2.5: Distribution of all chemicals found in the reviewed articles at level 1 to 2b (orange) overlaid on NORMAN SusDat database chemicals (blue) based on their molecular weights and XLogP3 value

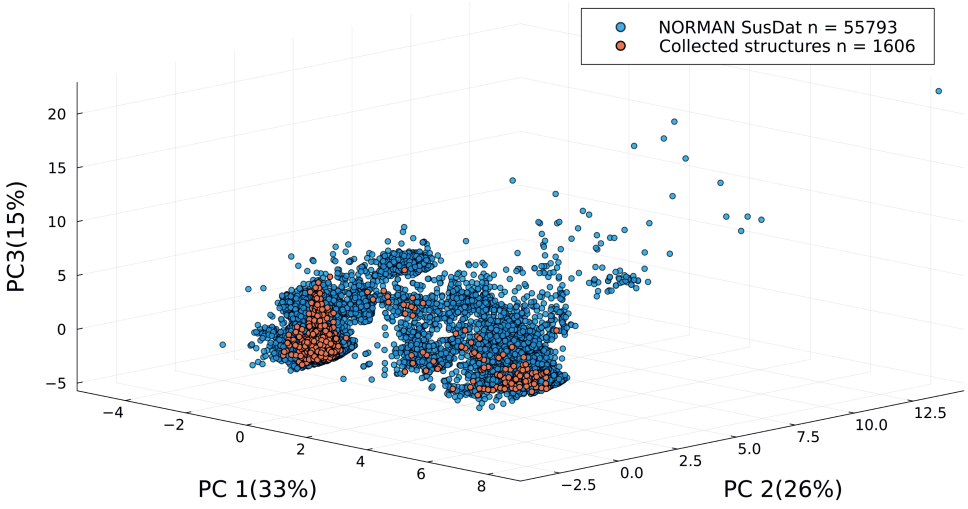


Figure 2.6: Scores plot of three principal components of the NORMAN SusDat database (blue) and the collected structures (orange).

the lack of detailed and standardized reporting of the experimental conditions. Therefore, a more rigorous investigation of the parameters and standardization of reporting criteria has to be designed and performed. Although the most widely accepted properties of compounds such as logP and MW are widely used while discussing chemical space, [23] in this study we showed that they may not be the most relevant markers for assessing the coverage of chemical space. Finally, such a low coverage emphasizes the need for more comprehensive approaches to experimental and data processing workflows in order to explore a broader range of the chemical space and ultimately protect human and environmental health.

2.4 Recommendations and Outlook

Despite the ability of NTA to provide holistic information about the chemical composition of the samples, their true coverage of the chemical space has not been investigated. Furthermore, the NTA studies have suffered from issues related to their reproducibility, due to the complexity of both experimental and computational approaches employed in NTA assays. One of the main bottlenecks for a more reproducible NTA assay is the lack of standardization of the reporting criteria (including the experimental conditions). Our detailed investigation of the previously published NTA studies further suggests the need for such criteria. Minimum accepted experimental criteria and data processing parameters should be reported to ensure the transparency and reliability of the results. The utilization of harmonized reporting tools such as BP4NTA SRT or NORMAN suspect screening reporting tools can help the reproduciblity and transparency of future NTA studies. [68, 77, 78] This will potentially lead to the acceptance of the NTA approach by the regulatory bodies.

The potential coverage of the chemical space should be assessed during the design of the experimental setups. Most of the recent studies focused their experimental setups based on the conventional workflow including HLB SPE for sample preparation, reverse phase separation with C18 columns, and DDA acquisition mode, without considering alternative approaches. The best practice would be an application of alternative extraction methods, implementation of orthogonal techniques (e.g. RPLC and HILIC), DIA acquisition mode as the first screening approach, and the application of reliable/robust data processing tools, preferably open source/access. For the identification part of the workflow, the sharing of experimental mass spectra of identified compounds along with their acquisi-

tion conditions is vital to the progress of the community. Additionally, archiving the raw data in public repositories for both the retrospective analysis as well as data processing tool development is highly essential.

To our knowledge, no other study has evaluated the coverage of the chemical space via NTA studies in such detail. However, due to the lack of standardized reporting criteria, the direct impact of different experimental choices on the covered chemical space could not be established. Also, our study is limited to the works published after 2017 and we only included studies with clear level 1 and 2 identification reporting. Furthermore, we excluded the suspect screening studies, which may result in an underestimation of the coverage of NTA studies. However, our study, even though limited, clearly shows the shortcomings of the current NTA practices and the need for further development in different areas - including experimental setup.

2.5 Supporting Information

The Supporting Information with figures (S1 - S11) showing the relationship between experimental parameters and the covered chemical space is available at `https://doi.org/10.1021/acs.est.3c03606`.

Notes

Information retrieved in this study can be found at `https://doi.org/10.5281/zenodo.8249216`. References to the reviewed studies and collected experimental parameters are in All experimental parameters.xlsx. The script to perform the calculations is available at `https://github.com//tobihul//Code-for-Critical-assessment-of-covered-chemical-space-with-LC-HRMS-non-targeted-analysis`. PubChemCrawler package is available at `https://github.com/JuliaHealth/PubChemCrawler.jl`.

Acknowledgments

The authors are thankful to the members of Environmental Modeling & Computational Mass Spectrometry (www.emcms.info). S.S. and V.T. are thankful to the UvA Data Science Center and ChemistryNL TKI for the financial support (projects Edified and SCOPE). The authors thank Denice van Herwerden for her help in setting up the calculations of elemental mass defects. J.W.O. is the recipient of a National Health and Medical Research Council (NHMRC) Investigator Grant (EL12009209) funded by the Australian Government. K.V.T, and S.S. also acknowledge funding support from the Australian National Health and Medical Research Council (NHMRC; APP1185347). The Queensland Alliance for Environmental Health Sciences (QAEHS), The University of Queensland (UQ), gratefully acknowledges the financial support of Queensland Health. This research was funded in whole or in part by NHMRC Emerging Leadership Fellowship EL1 2009209. For the purposes of open access, the author has applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising from this submission.

2.6 References

[1] Wild C. P. Complementing the Genome with an "Exposome": The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology. Cancer Epidemiol Biomark Prev. 2005 8;14(8):1847-50. doi:10.1158/1055-9965.epi-05-0456.

[2] Milman B. L, Zhurkovich I. K. The chemical space for non-target analysis. Trends Analyt Chem. 2017 12;97:179-87. doi:10.1016/j.trac.2017.09.013.

[3] Wambaugh J. F, Bare J. C, Carignan C. C, Dionisio K. L, Dodson R. E, Jolliet O, et al. New approach methodologies for exposure science. Curr Opin Toxicol. 2019 6;15:76-92. doi:10.1016/j.cotox.2019.07.001.

[4] Escher B. I, Stapleton H. M, Schymanski E. L. Tracking complex mixtures of chemicals in our changing environment. Science. 2020;367(6476):388-92. doi:10.1126/science.aay6636.

[5] Reymond J. L. The Chemical Space Project. Acc Chem Res. 2015 3;48(3):722-30. doi:10.1021/ar500432k.

[6] Van Deursen R, Reymond J. L. Chemical space travel. Chem Med Chem. 2007 5;2(5):636-40. doi:10.1002/cmdc.200700021.

[7] Reymond J. L, Awale M. Exploring chemical space for drug discovery using the chemical universe database. ACS Chem Neurosci. 2012 9;3(9):649-57. doi:10.1021/cn3000422.

[8] McEachran A. D, Sobus J. R, Williams A. J. Identifying known unknowns using the US EPA's CompTox Chemistry Dashboard. Anal Bioanal Chem. 2017 3;409:1729-35. doi:10.1007/s00216-016-0139-z.

[9] Vermeulen R, Schymanski E. L, Barabási A. L, Miller G. W. The exposome and health: where chemistry meets biology. Science. 2020 1;367(6476):392. doi:10.1126/science.aay3164.

[10] Landrigan P. J, Fuller R, Acosta N. J. R, Adeyi O, Arnold R, Basu N. N, et al. The Lancet Commission on pollution and health. Lancet. 2018 2;391(10119):462-512. doi:10.1016/s0140-6736(17)32345-0.

[11] Petrie B, Barden R, Kasprzyk-Hordern B. A review on emerging contaminants in wastewaters and the environment: Current knowledge, understudied areas and recommendations for future monitoring. Water Res. 2015 4;72:3-27. doi:10.1016/j.watres.2014.08.053.

[12] Samanipour S, O'Brien J. W, Reid M. J, Thomas K. V, Praetorius A. From Molecular Descriptors to Intrinsic Fish Toxicity of Chemicals: An Alternative Approach to Chemical Prioritization. Environ Sci Technol. 2022. doi:10.1021/acs.est.2c07353.

[13] Dulio V, van Bavel B, Brorström-Lundén E, Harmsen J, Hollender J, Schlabach M, et al. Emerging pollutants in the EU: 10 years of NORMAN in support of environmental policies and regulations. Environ Sci Eur. 2018 12;30(5). doi:10.1186/s12302-018-0135-3.

[14] Williams A. J, Grulke C. M, Edwards J, McEachran A. D, Mansouri K, Baker N. C, et al. The CompTox Chemistry Dashboard: A community data resource for environmental chemistry. J Cheminform. 2017 11;9(1). doi:10.1186/s13321-017-0247-6.

[15] Richard A. M, Judson R. S, Houck K. A, Grulke C. M, Volarath P, Thillainadarajah I, et al. ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. Chem Res Toxicol. 2016 8;29(8):1225-51. doi:0.1021/acs.chemrestox.6b00135.

[16] Tian Z, Zhao H, Peter K. T, Gonzalez M, Wetzel J, Wu C, et al. A ubiquitous tire rubber–derived chemical induces acute mortality in coho salmon. Science. 2021 1;371(6525):185-9. doi:10.1126/science.abd6951.

[17] Sauvé S, Desrosiers M. A review of what is an emerging contaminant. Chem Cent J. 2014;8:15. doi:10.1186/1752-153x-8-15.

[18] Maddela N. R, Ramakrishnan B, Kakarla D, Venkateswarlu K, Megharaj M. Major contaminants of emerging concern in soils: a perspective on potential health risks. RSC Adv. 2022 4;12(20):12396-415. doi:10.1039/d1ra09072k.

[19] Wang Y, Xiao J, Suzek T. O, Zhang J, Wang J, Bryant S. H. PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res. 2009;37(2):623-33. doi:10.1093/nar/gkp456.

[20] Mohammed Taha H, Aalizadeh R, Alygizakis N, Antignac J. P, Arp H. P. H, Bade R, et al. The NORMAN Suspect List Exchange (NORMAN-SLE): facilitating European and worldwide collaboration on suspect screening in high resolution mass spectrometry. Environ Sci Eur. 2022 12;34(104). doi:10.1186/s12302-022-00680-6.

[21] Aceña J, Stampachiacchiere S, Pérez S, Barceló D. Advances in liquid chromatography - High-resolution mass spectrometry for quantitative and qualitative environmental analysis. Anal Bioanal Chem. 2015 8;407:6289-99. doi:10.1007/s00216-015-8852-6.

[22] Picó Y, Barceló D. Transformation products of emerging contaminants in the environment and high-resolution mass spectrometry: A new horizon. Anal Bioanal Chem. 2015 8;407:6257-73. doi:10.1007/S00216-015-8739-6.

[23] Schulze B, Jeon Y, Kaserzon S, Heffernan A. L, Dewapriya P, O'Brien J, et al. An assessment of quality assurance/quality control efforts in high resolution mass spectrometry non-target workflows for analysis of environmental samples. Trends Anal Chem. 2020 12;133:116063. doi:10.1016/j.trac.2020.116063.

[24] Muir D. C. G, Howard P. H. Are there other persistent organic pollutants? A challenge for environmental chemists. Environ Sci Technol. 2006 12;40:7157-66. doi:10.1021/es061677a.

[25] Samanipour S, Martin J. W, Lamoree M. H, Reid M. J, Thomas K. V. Letter to the Editor: Optimism for Nontarget Analysis in Environmental Chemistry. Environ Sci Technol. 2019 5;53:5529-30. doi:10.1021/acs.est.9b01476.

[26] Menger F, Gago-Ferrero P, Wiberg K, Ahrens L. Wide-scope screening of polar contaminants of concern in water: A critical review of liquid chromatography-high resolution mass spectrometry-based strategies. Trends Environ Anal Chem. 2020 12;28:e00102. doi:10.1016/j.teac.2020.e00102.

[27] Hollender J, Schymanski E. L, Singer H. P, Ferguson P. L. Nontarget Screening with High Resolution Mass Spectrometry in the Environment: Ready to Go? Environ Sci Technol. 2017 10;51:11505-12. doi:10.1021/acs.est.7b02184.

[28] Zedda M, Zwiener C. Is nontarget screening of emerging contaminants by LC-HRMS successful? A plea for compound libraries and computer tools. Anal Bioanal Chem. 2012 7;403:2493-502. doi:10.1007/s00216-012-5893-y.

[29] Krauss M, Singer H, Hollender J. LC-high resolution MS in environmental analysis: From target screening to the identification of unknowns. Anal Bioanal Chem. 2010 6;397:943-51. doi:10.1007/s00216-010-3608-9.

[30] McCord J. P, Groff L. C, Sobus J. R. Quantitative non-targeted analysis: Bridging the gap between contaminant discovery and risk characterization. Environ Int. 2022 1;158:107011. doi:10.1016/j.envint.2021.107011.

[31] Schmidt T. C. Recent trends in water analysis triggering future monitoring of organic micropollutants. Anal Bioanal Chem. 2018 7;410:3933-41. doi:10.1007/s00216-018-1015-9.

[32] Samanipour S, Reid M. J, Thomas K. V. Statistical Variable Selection: An Alternative Prioritization Strategy during the Nontarget Analysis of LC-HR-MS Data. Anal Chem. 2017 5;89:5585-91. doi:10.1021/acs.analchem.7b00743.

[33] van Herwerden D, O'Brien J. W, Lege S, Pirok B. W. J, Thomas K. V, Samanipour S. Cumulative Neutral Loss Model for Fragment Deconvolution in Electrospray Ionization High-Resolution Mass Spectrometry Data. Anal Chem. 2023 8;95(33):12247-55. doi:10.1021/acs.analchem.3c00896.

[34] Hohrenk L. L, Itzel F, Baetz N, Tuerk J, Vosough M, Schmidt T. C. Comparison of Software Tools for Liquid Chromatography-High-Resolution Mass Spectrometry Data Processing in Nontarget Screening of Environmental Samples. Anal Chem. 2020 1;92:1898-907. doi:10.1021/acs.analchem.9b04095.

[35] Samanipour S, Baz-Lomba J. A, Alygizakis N. A, Reid M. J, Thomaidis N. S, Thomas K. V. Two stage algorithm vs commonly used approaches for the suspect screening of complex environmental samples analyzed via liquid chromatography high resolution time of flight mass spectroscopy: A test study. J Chromatograph A. 2017 6;1501:68-78. doi:10.1016/j.chroma.2017.04.040.

[36] Hernández F, Bakker J, Bijlsma L, de Boer J, Botero-Coy A. M, Bruinen de Bruin Y, et al. The role of analytical chemistry in exposure science: Focus on the aquatic environment. Chemosphere. 2019 5;222:564-83. doi:10.1016/j.chemosphere.2019.01.118.

[37] Brack W, Hollender J, de Alda M. L, Müller C, Schulze T, Schymanski E, et al. High-resolution mass spectrometry to complement monitoring and track emerging chemicals and pollution trends in European water resources. Environ Sci Eur. 2019 12;31:62. doi:10.1186/s12302-019-0230-0.

[38] Black G, Lowe C, Anumol T, Bade J, Favela K, Feng Y. L, et al. Exploring chemical space in non-targeted analysis: a proposed ChemSpace tool. Anal Bioanal Chem. 2023 1;415:35-44. doi:10.1007/s00216-022-04434-4.

[39] Hites R. A, Jobst K. J. Is Nontargeted Screening Reproducible? Environ Sci Technol. 2018 11;52:11975-6. doi:10.1021/acs.est.8b05671.

[40] Ulrich E. M, Sobus J. R, Grulke C. M, Richard A. M, Newton S. R, Strynar M. J, et al. EPA's non-targeted analysis collaborative trial (ENTACT): genesis, design, and initial findings. Anal Bioanal Chem. 2019 2;411:853-66. doi:10.1007/s00216-018-1435-6.

[41] Fisher C. M, Peter K. T, Newton S. R, Schaub A. J, Sobus J. R. Approaches for assessing performance of high-resolution mass spectrometry–based non-targeted analysis methods. Anal Bioanal Chem. 2022 9;414:6455-71. doi:10.1007/s00216-022-04203-3.

[42] Knolhoff A. M, Premo J. H, Fisher C. M. A proposed quality control standard mixture and its uses for evaluating nontargeted and suspect screening LC/HR-MS method performance. Anal Chem. 2021 1;93(3):1596-603. doi:10.1021/acs.analchem.0c04036.

[43] Singh R. R, Chao A, Phillips K. A, Xia X. R, Shea D, Sobus J. R, et al. Expanded coverage of non-targeted LC-HRMS using atmospheric pressure chemical ionization: a case study with ENTACT mixtures. Anal Bioanal Chem. 2020 8;412:4931-9. doi:10.1007/s00216-020-02716-3.

[44] Newton S. R, Sobus J. R, Ulrich E. M, Singh R. R, Chao A, McCord J, et al. Examining NTA performance and potential using fortified and reference house dust as part of EPA's Non-Targeted Analysis Collaborative Trial (ENTACT). Anal Bioanal Chem. 2020 7;412:4221-33. doi:10.1007/s00216-020-02658-w.

[45] Schymanski E. L, Jeon J, Gulde R, Fenner K, Ruff M, Singer H. P, et al. Identifying small molecules via high resolution mass spectrometry: Communicating confidence. Environ Sci Technol. 2014 2;48:2097-8. doi:10.1021/es5002105.

[46] Hulleman T, Turkina V, O'Brien J. W, Chojnacka A, Thomas K. V, Samanipour S. Data files for: Critical assessment of the chemical space covered by LC-HRMS non-targeted analysis; 2023. doi:10.5281/zenodo.8249216.

[47] Hulleman T. Code for: Critical assessment of covered chemical space with LC-HRMS non-targeted analysis.. GitHub; 2023. https://github.com/tobihul/Code-for-Critical-assessment-of-covered-chemical-space-with-LC-HRMS-non-targeted-analysis.

[48] Cheng T, Zhao Y, Li X, Lin F, Xu Y, Zhang X, et al. Computation of octanol-water partition coefficients by guiding an additive model with knowledge. J Chem Inf Model. 2007;47:2140-8. doi:10.1021/ci700257y.

[49] van Herwerden D, O'Brien J. W, Choi P. M, Thomas K. V, Schoenmakers P. J, Samanipour S. Naive Bayes classification model for isotopologue detection in LC-HRMS data. Chemom Intell Lab Syst. 2022 4;223:104515. doi:10.1016/j.chemolab.2022.104515.

[50] Sleno L. The use of mass defect in modern mass spectrometry. J Mass Spectrom. 2012 2;47:226-36. doi:10.1002/jms.2953.

[51] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning Data Mining, Inference, and Prediction. Springer; 2009. doi:10.1007/b94608.

[52] Djoumbou Feunang Y, Eisner R, Knox C, Chepelev L, Hastings J, Owen G, et al. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. J Cheminform. 2016 11;8:61. doi:10.1186/s13321-016-0174-y.

[53] Kruve A. Semi-quantitative non-target analysis of water with liquid chromatography/high-resolution mass spectrometry: How far are we? Rapid Commun Mass Spectrom. 2019;33(S3):54-63. doi:10.1002/rcm.8208.

[54] Lyytikäinen M, Kukkonen J. V. K, Lydy M. J. Analysis of pesticides in water and sediment under different storage conditions using gas chromatography. Arch Environ Contam Toxicol. 2003 5;44:437-44. doi:10.1007/s00244-002-2168-1.

[55] Been F, Kruve A, Vughs D, Meekel N, Reus A, Zwartsen A, et al. Risk-based prioritization of suspects detected in riverine water using complementary chromatographic techniques. Water Res. 2021 10;204:117612. doi:10.1016/j.watres.2021.117612.

[56] Hu L. X, Olaitan O. J, Li Z, Yang Y. Y, Chimezie A, Adepoju-Bello A. A, et al. What is in Nigerian waters? Target and non-target screening analysis for organic chemicals. Chemosphere. 2021 12;284:131546. doi:10.1016/j.chemosphere.2021.131546.

[57] Köppe T, Jewell K. S, Dietrich C, Wick A, Ternes T. A. Application of a non-target workflow for the identification of specific contaminants using the example of the Nidda river basin. Water Res. 2020 7;178:115703. doi:10.1016/j.watres.2020.115703.

[58] Kunzelmann M, Winter M, Åberg M, Hellenäs K. E, Rosén J. Non-targeted analysis of unexpected food contaminants using LC-HRMS. Anal Bioanal Chem. 2018 9;410:5593-602. doi:10.1007/s00216-018-1028-4.

[59] Reemtsma T, Berger U, Arp H. P. H, Gallard H, Knepper T. P, Neumann M, et al. Mind the Gap: Persistent and Mobile Organic Compounds - Water Contaminants That Slip Through. Environ Sci Technol. 2016 10;50:10308-15. doi:10.1021/acs.est.6b03338.

[60] Brüggen S, Schmitz O. J. A New Concept for Regulatory Water Monitoring Via High-Performance Liquid Chromatography Coupled to High-Resolution Mass Spectrometry. J Anal Test. 2018 10;2:342-51. doi:10.1007/s41664-018-0081-5.

[61] Badea S. L, Geana E. I, Niculescu V. C, Ionete R. E. Recent progresses in analytical GC and LC mass spectrometric based-methods for the detection of emerging chlorinated and brominated contaminants and their transformation products in aquatic environment. Sci Total Environ. 2020 6;722:137914. doi:10.1016/j.scitotenv.2020.137914.

[62] Greibrokk T, Andersen T. High-temperature liquid chromatography. J Chromatogr A. 2003 6;1000(1-2):743-55. doi:10.1016/s0021-9673(02)01963-5.

[63] Snyder L. R, Kirkland J. J, Dolan J. W. Introduction to Modern Liquid Chromatography. John Wiley and Sons; 2010. doi:10.1002/9780470508183.

[64] Bade R, Bijlsma L, Sancho J. V, Hernández F. Critical evaluation of a simple retention time predictor based on LogKow as a complementary tool in the identification of emerging contaminants in water. Talanta. 2015 7;139:143-9. doi:10.1016/j.talanta.2015.02.055.

[65] Kaliszan R, Haber P, Tomasz B, Siluk D, Valko K. Lipophilicity and pKa estimates from gradient high-performance liquid chromatography. J Chromatogr A. 2002 8;965:117-27. doi:10.1016/s0021-9673(02)00114-0.

[66] Phillips A. L, Peter K. T, Sobus J. R, Fisher C. M, Manzano C. A, McEachran A. D, et al. Standardizing non-targeted analysis reporting to advance exposure science and environmental epidemiology. J Expo Sci Environ Epidemiol. 2023 2. doi:10.1038/s41370-022-00490-1.

[67] Peter K. T, Phillips A. L, Knolhoff A. M, Gardinali P. R, Manzano C. A, Miller K. E, et al. Nontargeted Analysis Study Reporting Tool: A Framework to Improve Research Transparency and Reproducibility. Anal Chem. 2021 10;93(41):13870-9. doi:10.1021/acs.analchem.1c02621.

[68] Place B. J, Ulrich E. M, Challis J. K, Chao A, Du B, Favela K, et al. An Introduction to the Benchmarking and Publications for Non-Targeted Analysis Working Group. Anal Chem. 2021 12;93(49):16289-96. doi:10.1021/acs.analchem.1c02660.

[69] Malm L, Palm E, Souihi A, Plassmann M, Liigand J, Kruve A. Guide to Semi-Quantitative Non-Targeted Screening Using LC/ESI/HRMS. Molecules. 2021 6;26(12):3524. doi:10.3390/molecules26123524.

[70] Zubarev R. A, Makarov A. Orbitrap mass spectrometry. Anal Chem. 2013 6;85(11):5288-96. doi:10.1021/ac4001223.

[71] Boesl U. Time-of-flight mass spectrometry: Introduction to the basics. Mass Spectrom Rev. 2017 1;36(1):86-109. doi:10.1002/mas.21520.

[72] Gosetti F, Mazzucco E, Gennaro M. C, Marengo E. Contaminants in water: non-target UHPLC/MS analysis. Environ Chem Lett. 2016 3;14:51-65. doi:10.1007/s10311-015-0527-1.

[73] Jeong Y, Da Silva K. M, Iturrospe E, Fuiji Y, Boogaerts T, van Nuijs A. L. N, et al. Occurrence and contamination profile of legacy and emerging per- and polyfluoroalkyl substances (PFAS) in Belgian wastewater using target, suspect and non-target screening approaches. J Hazard Mater. 2022 9;437:129378. doi:10.1016/j.jhazmat.2022.129378.

[74] Xia X, Zheng Y, Tang X, Zhao N, Wang B, Lin H, et al. Nontarget Identification of Novel Per- and Polyfluoroalkyl Substances in Cord Blood Samples. Environ Sci Technol. 2022 12;56(23):17061-9. doi:10.1021/acs.est.2c04820.

[75] Yu N, Wen H, Wang X, Yamazaki E, Taniyasu S, Yamashita N, et al. Nontarget Discovery of Per-And Polyfluoroalkyl Substances in Atmospheric Particulate Matter and Gaseous Phase Using Cryogenic Air Sampler. Environ Sci Technol. 2020 3;54(6):3103-13. doi:10.1021/acs.est.9b05457.

[76] Alygizakis N. A, Oswald P, Thomaidis N. S, Schymanski E. L, Aalizadeh R, Schulze T, et al. NORMAN digital sample freezing platform: A European virtual platform to exchange liquid chromatography high resolution-mass spectrometry data and screen suspects in "digitally frozen" environmental samples. Trends Anal Chem. 2019 6;115:129-37. doi:10.1016/j.trac.2019.04.008.

[77] Rostkowski P, Haglund P, Aalizadeh R, Alygizakis N, Thomaidis N, Arandes J. B, et al. The strength in numbers: comprehensive characterization of house dust using complementary mass spectrometric techniques. Anal Bioanal Chem. 2019 4;411:1957-77. doi:10.1007/s00216-019-01615-6.

[78] Alygizakis N. A, Samanipour S, Hollender J, Ibáñez M, Kaserzon S, Kokkali V, et al. Exploring the Potential of a Global Emerging Contaminant Early Warning Network through the Use of Retrospective Suspect Screening with High-Resolution Mass Spectrometry. Environ Sci Technol. 2018 5;52:5135-44. doi:10.1021/acs.est.8b00365.

[79] Podwojski K, Fritsch A, Chamrad D. C, Paul W, Sitek B, Stühler K, et al. Retention time alignment algorithms for LC/MS data must consider non-linear shifts. Bioinformatics. 2009 3;25(6):758-64. doi:10.1093/bioinformatics/btp052.

[80] Clauser K. R, Baker P, Burlingame A. L. Role of accurate mass measurement (±10 ppm) in protein identification strategies employing MS or MS/MS and database searching. Anal Chem. 1999 7;71(14):2871-82. doi:10.1021/ac9810516.

[81] Minkus S, Bieber S, Letzel T. Spotlight on mass spectrometric non-target screening analysis: Advanced data processing methods recently communicated for extracting, prioritizing and quantifying features. Anal Sci Adv. 2022 4;3:103-12. doi:10.1002/ansa.202200001.

[82] Kleis J. N, Hess C, Germerott T, Roehrich J. Sensitive Screening of New Psychoactive Substances in Serum Using Liquid Chromatography-Quadrupole Time-of-Flight Mass Spectrometry. J Anal Toxicol. 2022 7;46(6):592-9. doi:10.1093/jat/bkab072.

[83] Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, et al. MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom. 2010 7;45(7):703-14. doi:10.1002/jms.1777.

3

Enhancing Predictive Modeling with Molecular Fingerprint Fusion Strategies

Authors: Viktoriia Turkina, Melanie R.W. Messih, Etienne Kant, Jelle T. Gringhuis, Annemieke Petrignani, Garry Corthals, Jake W. O'Brien, Saer Samanipour

Published in: Digital Discovery, 2026,5, 2232-2242.

https://doi.org/10.1039/D5DD00302D

Abstract

A large number of chemicals remain poorly characterized in terms of their physicochemical properties, biological activity, and environmental fate. Quantitative structure-activity relationship (QSAR) models have become indispensable tools for predicting these properties, especially for compounds that lack comprehensive experimental data. The choice of structural representation as an input to such models plays a critical role in ensuring high predictive performance and in identifying molecular features that strongly contribute to activity prediction. Both hashed and non-hashed molecular fingerprints are widely employed as inputs in QSAR modeling across various domains. While some studies have explored combining multiple fingerprints to improve molecular representation, comprehensive investigations into different fingerprint fusion strategies and the generalizability of a fused fingerprint across diverse prediction tasks remain limited. In this study, we applied low-, mid-, and high-levels fusion strategies to combine six non-hashed fingerprints and evaluated model performance across six publicly available datasets, including three regression and three classification tasks. Our results demonstrate that mid-level fusion, where fingerprint bits are selectively combined based on their importance within individual models, consistently improves predictive accuracy, as assessed by RMSE and R^2 for regression, and F1-score and ROC-AUC for classification. The algorithm developed for molecular fingerprints fusion is methodologically general and can be applied to a wide range of predictive modeling problems or other non-hashed molecular fingerprints.

3.1 Introduction

The increase in variety and global scale of chemical production has resulted in thousands of new chemicals being introduced into the market and eventually into the environment. [1, 2] A number of these chemicals have been shown to have adverse impact on environmental and human health. [3,4] Due to limited resources, experimental assessment of activities and properties of such a large number of chemicals has become unfeasible. According to the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) regulation, among thousands of new chemicals produced in the European Union annually, only substances with production exceeding one tonne per year are required to be registered while the rest remain uncharacterized. [5, 6] To address this gap, modeling approaches become indispensable in investigating and estimating chemical properties and behavior. [7–11]

A major part of the existing modeling approaches employ QSAR, establishing a relationship, linear or non-linear, between a chemical's structure and its activity. [10] For such models, a chemical's structure can be represented in many ways, for example by molecular descriptors, molecular fingerprints (FPs), or graph-based molecular representations. [10–16] Molecular descriptors are abstract numerical values that capture different properties of a molecule. Molecular fingerprints simplify the molecular structure into a series of bits or counts in a vector. [17, 18] In graph-based representations, each node (atom) carries certain features (such as atom type, charge, or hybridization state), and these features are aggregated and processed by computational models, such as graph neural networks (GNNs), to predict molecular properties or activities. [19]

Among these representations, molecular fingerprints (FPs) deserve particular attention because of their interpretability and broad applicability across machine learning models. FPs provide a standardized way to encode molecular structure into a format that is easy to process while still retaining important chemical information. Moreover, they have become a foundational tool in many cheminformatics applications, including virtual screening, [17, 20, 21] similarity searching, and predictive modeling.

Generally, there are two ways of encoding a molecular structure into FPs: non-hashed, which uses a predefined list of molecular features, and hashed, where the structure is converted with the help of a hash function. [17, 22, 23] The hashed algorithms are specifically

tailored to capture intricate details of the structure. [24] However, depending on the complexity of the structure, multiple molecular features may be represented with the same bit. This leads to a loss of one-to-one correspondence between bits and structural information. [18] Such clarity in interpretation is crucial for establishing structural rules for potential hazard identification. [25] While non-hashed FPs avoid this issue, the initial predefined list does not always include a comprehensive set of structural features necessary to effectively distinguish relevant variations in structures. [18] This limitation can lead to the exclusion of information essential for accurate modeling.

There are several hashed and non-hashed algorithms to calculate FPs with different levels of structural information included. For example, topological FPs describe molecular features limited to a certain distance, while structure-based FPs depict different aspects of the molecular substructures. [17]

Although different FPs are extensively used in various fields, it has also been demonstrated that a single way to encode structure into FPs may not capture all the necessary structural information for accurate modeling. [26] To address this, several studies have focused on fusing multiple types of representations to enhance predictive performance. For instance, in the context of ligand-based virtual screening, Hert et al. showed that combining similarity scores or ranked lists calculated based multiple fingerprints or reference structures improved retrieval performance and scaffold hopping. [27] Fingerprint fusion has also been adopted within supervised predictive modeling frameworks. In case of the most straightforward approach, low-level fusion, multiple FPs are concatenated into a single feature vector. Xie et al., for example, combined MACCS and ECFP fingerprints through horizontal concatenation to improve predictions of logP and protein–ligand binding affinities. [28] However, this method increases feature dimensionality and may introduce redundancy and sparsity, which can challenge model training, especially with large chemical datasets.

To mitigate such issues, mid-level fusion approaches apply transformation or selection techniques to individual representations before combining them. Srisongkram et al. demonstrated improved performance on KRASG12C ligand affinity prediction by selecting 13 highly target-correlated FP bits from a conjoint PubChem and SSC representation. [29] Similarly, Shen et al. applied normalization and principal component analysis (PCA) independently to ECFP4, MACCS, and 208 descriptors, then the transformed features were concatenated for classifying USP7 inhibitors. [30] FP-GNN represents another mid-level

strategy, integrating molecular fingerprints with graph-based embeddings in a learned latent space. [31] However, such representations usually require extensive datasets, which are mostly unavailable for environmental QSAR modeling. [32]

At the most abstract level, high-level fusion combines predictions from multiple models trained on distinct FPs. Matsuyama and Ishida implemented this strategy using a stacking ensemble of eight different FP types, including Pharmacophore, AP2DC, Daylight, ECFP4, ECFP6, MACCS, Topotor, and Chemdesc, to predict activity across 14 PubChem targets. [33]

While combining multiple FPs can help capture complementary structural information, data fusion is not without trade-offs. In particular, low-level fusion can significantly expand feature space, leading to high memory usage and reduced scalability. For example, training a model on the full CompTox database [34] using several concatenated FPs would require gigabytes of RAM, making it impractical for many QSAR pipelines. These challenges underscore the importance of carefully selecting fusion levels to balance informativeness, model complexity, and resource constraints.

Although fingerprint fusion was explored in several QSAR studies, the main motivation was finding the most effective structural representation to predict specific targets. [28, 29, 35] Systematic investigations into the generalizability of fused fingerprints across multiple fusion levels and prediction tasks remain unexplored.

In this study, we systematically evaluated the effect of fingerprint fusion strategies on QSAR model performance. We selected six publicly available datasets including three for the regression and three for the classification problems. To keep the interpretability of molecular representation, we focused on a set of six non-hashed FPs: Estate, PubChem, MACCS, AtomPairCount, Substructure Count, and Klekotha Roth as structural input for QSAR models. We fused calculated fingerprints on low-, mid-, and high-levels and compared the QSAR performance based on root-mean-square error (RMSE), coefficient of determination (R^2) for regression and F1-score, the area under the receiver operating characteristic curve (ROC-AUC) for classification tasks.

3.2 Methods

3.2.1 Overall workflow

The overall workflow begins with the conversion of canonical SMILES strings [36] into a set of non-hashed FPs. For each individual FP, we trained and optimized either a regression or classification model using the Random Forest (RF) algorithm, depending on the target. We then applied low-, mid-, and high-levels fusion strategies to combine the individual fingerprints and assessed the performance of models trained on both individual and fused FPs. These models were evaluated across a range of biological, ecological, and physiochemical prediction tasks (Figure 3.1).

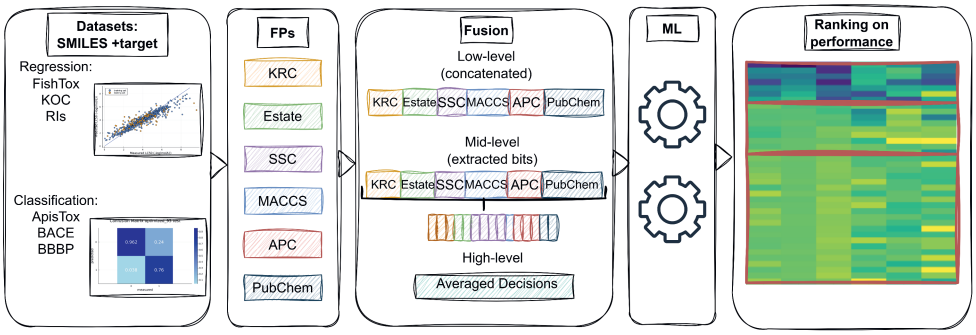


Figure 3.1: The overall workflow of the study. SMILES strings were converted into set of non-hashed fingerprints. For each FP type, an RF model was trained and optimized for either regression or classification, depending on the prediction task. Subsequently, low-, mid-, and high-levels fusion strategies were employed to combine the individual FPs. Performance of the models was evaluated across various biological, ecological, and physiochemical endpoints.

3.2.2 Molecular Fingerprints

We selected a set of various non-hashed topological and structure-based FPs including Atom Pair 2D Count (AP2DC), [37] Electrotopological state (E-state), [38] Klekotha-Roth Count (KRC), [39] Molecular Access Systems (MACCS), [40] PubChem, [41] and Substructure Keys Count (SSC) FPs. [42] E-state, KRC, MACCS, PubChem, and SSC FPs are substructure-based FPs and AP2DC is a topological one. Each bit within this set corresponds to an interpretable molecular feature, aiding in understanding the significance of structural information related to the targeted activity (Section 5.5 Table S1).

Various FP algorithms encompass different aspects and details of molecular structure, meaning they contain different sets of substructures or molecular features in a various number of bits. For instance, AP2DC FP is a 780-bit topological FP defined by the atomic environment and the distance between atomic pairs. [37] The shortest among the selected FPs is the 79-bit E-state FP, which is based on both the electronic and topological characteristics of molecules. [38] In contrast, KRC FP is the largest one encoded in 4860 bits. It was originally developed to predict certain types of bioactivity through substructure screening and consists of relevant SMART strings. [39] The MACCS FP, one of the most frequently used, is constructed using SMART patterns and optimized for substructure searching, resulting in 166 bits. [40] Another widely adopted FP, PubChem, covers a wide range of different substructures and features recorded in the PubChem database in a 881-bit binary string. [41] Finally, SSC FP consists of a count vector of 307 SMART patterns selected for functional group classification by Christian Laggner. [42]

The canonical SMILES were used to calculate six non-hashed FPs. All FPs except E-state were computed using the PaDEL software package, implemented via a Python 3 wrapper called PaDELPy (version 0.1.13). [42] The E-state FP was calculated using the RDKit software package (version 2022.9.5), which includes Python bindings. [43] SMILES for which fingerprint generation failed (e.g., due to parsing errors or unsupported structures) were excluded from subsequent analysis. Across all datasets, the fraction of removed molecules was < 1.3% (Section 3.5 Table S2). Generated fingerprints were further inspected for invalid entries. Any NaN values, missing fingerprint bits, or extreme values arising during fingerprint calculation were replaced with zero to ensure numerical stability and consistency across datasets. No imputation based on label information was performed, and all preprocessing steps were applied prior to data splitting, meaning the preprocessing steps were applied uniformly to all datasets before training, cross-validation, and test set evaluation.

3.2.3 Datasets

We collected six datasets from different sources: two from ecotoxicology, two related to physiological parameters, and two to physicochemical properties. Three of the datasets correspond to regression problems, FishTox, [44] KOC, [45] and RI, [46] while the other three, ApisTox, [47] BBBP, [48] and BACE, [49] to classification tasks.

FishTox dataset contains experimentally determined acute fish toxicity LC50[-LOG(mg/L)]

for 907 chemicals. The toxicity values are defined as the concentration of the chemicals that resulted in the death of 50% of feathed minnows (Pimephales promelas) throughout 96 h (96 h LC50). These values were collected from three databases: OASIS, ECOTOX, and EAT5, and were supplied by Cassotti et al. [50] The chemicals in this data set were derived from different origins, including pharmaceuticals, pesticides, conventional persistent organic pollutants, and industrial chemicals. [44]

KOC dataset contains experimentally measured logarithmic values of the soil organic carbon-normalized (KOC[LOG (L kg^{-1})]) for 824 organic compounds representing 31 different chemical classes. The data was compiled from EPI Suite (Version 4.1) and the literature sources and was provided by Wang et al. [45] The values are measured by the standard testing guidelines, e.g. the Organization for Economic Cooperation and Development (OECD) guidelines for estimation of KOC using HPLC and the batch equilibrium method.

RIs dataset contains Retention Indices (RI) for reversed-phase liquid chromatography (RPLC), based on synthetic and naturally occurring homologous series of amphiphilic cocamide diethanolamine surfactants (C(n = 0-23)-DEA). It includes 3018 emerging pollutants, 2290 measured in positive electrospray ionization (+ESI) mode and 728 in negative (−ESI) mode. The RI values were obtained under a variety of liquid chromatography (LC) conditions, including variations in pH, column types, temperatures, flow rates, mobile phase compositions, and gradient elution programs and were supplied by Aalizadeh et al. [46]

ApisTox consists of 1035 compounds with their experimentally measured acute bee toxicity. The values were collected from the ECOTOX, PPDB, and BPDB databases. The dataset contains 296 toxic and 739 non-toxic molecules. U.S. EPA guidelines for honey bees were used to define toxic/non-toxic labels. Compounds with an LD50 less than or equal to a threshold 11 μg/org were classified as toxic. The dataset was compiled and provided by Adamczyk et al. [47]

BBBP dataset consists of 2,053 molecules compiled from several publications focused on blood-brain barrier (BBB) permeability. Compounds were divided into BBB+ (able to cross the BBB) and BBB- (unable to cross) classes. Chemicals were assigned to the BBB+ class if blood-brain partition coefficient value (logBB) is greater than or equal to -1, and to the BBB- class if logBB is less than -1. Overall, the dataset provided by Martins et al. [48] contains 1,570 BBB+ and 483 BBB- molecules.

Finally, *BACE dataset* includes 1547 synthetic inhibitors of human β-secretase 1 (BACE-1). The dataset uses binary labels to indicate compound activity. Compounds were classified as active if their IC50 values, the concentration at which a compound inhibits 50% of its target activity, were ≤ 100 nM, and inactive otherwise. This dataset was compiled and provided by Subramanian et al. [49]

3.2.4 Modeling

For modeling, a Random Forest Regression (RFR) and Random Forest Classification (RFC) algorithm were implemented in Julia with scikit-learn. [51] Random Forest is a supervised algorithm that constructs several unique decision trees from bootstrap data. After model development, the predictions resulting from the individual trees are averaged to produce the final RFR model prediction and from the major voting of predictions to produce the final RFC model prediction. [52] The major benefit of employing the RF modeling approach lies in its ability to handle nonlinearity and noncontinuity within the data while keeping the interpretability of the results. There are several other widely applied modeling algorithms, such as Support Vector Machine (SVM) or Convolutional Neural Networks (CNNs), for this purpose. However, these algorithms tend to be very data-intensive and have low transparency. Therefore, the RF algorithm was selected to ensure the robustness and transparency of the process. Also, RF was successfully applied previously for QSAR modeling and showed accurate and robust performance. [44, 53–55] While this study focuses on RF models to ensure a controlled comparison of fusion strategies, the proposed fingerprint fusion framework is not tied to a specific learning algorithm and can, in principle, be combined with other commonly used algorithms, although the magnitude of performance gains may vary.

To train individual fingerprint-based models, the datasets were randomly split into a training set (90%) and a test set (10%). The split was selected to maintain an independent set for testing performance while preserving sufficient training data for stable feature importance estimation, which is critical for mid-level fusion. Since such design limits the ability to quantify uncertainty in performance metrics, the observed performance should be interpreted within this limitation. The training set was employed to develop, optimize, and validate the model, while the test set was applied to evaluate the performance of the optimized models. The hyperparameters were optimized using a 3D grid with the number

of trees varying from 100 to 600, the number of variables to consider being sqrt or log2, and the minimum number of samples in each leaf ranging from 2 to 8. For classification tasks, class imbalance was handled using cost-sensitive learning by setting *class_weight* to "balanced" in the Random Forest classifier. The performance evaluation of the models was based on initial estimation using a training set, internal validation with 3-fold cross-validation, and external validation using a test set.

The regression tasks were evaluated by RMSE and R^2, while the classification tasks were evaluated by F1-score and ROC-AUC (Section 3.5 Equations S1 - S5). Model performance reported in the Results section refers exclusively to the external test set. Training and 3-fold cross-validation were used only for hyperparameter optimization and internal validation.

3.2.5 PCA

To visualize the chemical space coverage with different fingerprints, we used Principal Component Analysis (PCA) which is an unsupervised method, enabling an unbiased evaluation of the underlying trends in the data. For each dataset, we constructed a PCA model using the horizontally concatenated set of non-hashed fingerprints, representing the maximal structural information available from these fingerprints. By analyzing the resulting scores and loadings, we assessed the chemical space captured by individual fingerprints within each dataset. The algorithm was implemented in Julia using scikit-learn. [51]

3.2.6 HCA

To investigate the similarity and relationships between trained models based on their predictive performance, we applied Hierarchical Cluster Analysis (HCA), an unsupervised clustering technique. Models were clustered according to normalized performance metrics (R^2 and F1-score) across datasets. For each dataset, these scores, obtained from both individual and fused FPs models, were normalized and used as input for clustering. Euclidean distance was employed as the similarity metric, and the average linkage method was used to build the hierarchical clusters. HCA was implemented in Julia using the Clustering.jl package.

3.2.7 Molecular fingerprints' fusion

We evaluated the impact of low-, mid-, and high-levels data fusion strategies on model performance. Importantly, we did not apply any explicit correlation filtering or dimensionality reduction prior to fusion or modelling. This decision was made because of models we used in this study, RF and PCA, are inherently robust to multicollinearity.

For *low-level* fusion, we horizontally concatenated all the bits from six selected non-hashed fingerprints. In addition, we trained models using a reduced version of this fused fingerprint by retaining only those features whose cumulative importance accounted for at least 55% and up to 95% of the total, in 5% steps. The importance was calculated based on impurity-based variable importance. This variable importance metric reflects the reduction of impurity achieved by the splits on each variable during the model training process. This metric indicates the contribution of each feature to explaining the variance in the training data.

For *mid-level* fusion, we first trained, optimized, validated, and tested Random Forest (RF) models using each individual fingerprint. From each model, we extracted important variables based on two selection strategies: cumulative and individual importance. Specifically, we selected features with a cumulative importance between 55% and 95% (in 5% steps) or with an individual importance between 0.2% and 2.0% (in 0.2% steps). The selected features from each model were then combined to create a fused fingerprint, which served as input for final model development.

Finally, for *high-level* fusion, we implemented an ensemble approach by averaging the outputs of the six individual fingerprint models. In the case of classification, the final class was assigned based on the averaged class probabilities from these models.

3.2.8 Calculations and data availability

All calculations were performed on a personal computer (PC) with an Intel Core i7-1260P central processing unit and 32 GB of RAM operating Windows 10 Education version 22H2. All data processing and statistical analyses were performed using Julia language version 1.10.6. All datasets as well as calculated FPs in this study, and interpretation of the optimized FP can be found at `https://doi.org/10.5281/zenodo.15791757`. The algorithm

for a molecular fingerprint fusion was converted into an open-source open-access Julia package and can be found at `https://bitbucket.org/viktoriiaturkina/fp_optimization.jl/src/master/`.

3.3 Results and discussion

3.3.1 Individual FPs on PCA-space

To explore the chemical information captured by various FPs, we performed PCA for each dataset individually (Section 3.5 Figures S1 - S5). Our goal was to evaluate how effectively individual FPs, as well as their combinations, represent the structural diversity of the target chemical space. If an individual FP encodes information that largely overlaps with the concatenated set of all FPs, its PCA projection should closely resemble that of the full fusion. Conversely, if the FPs capture complementary structural features, the combined coverage of their individual PCA spaces should approximate the coverage observed when all FPs are concatenated, indicating that each contributes unique and non-redundant information.

We observed that concatenated FP, combining all six FPs, resulted in the most comprehensive structural coverage, while individual fingerprints captured more limited, specific subspaces. On average, the first two principal components accounted for 64% of the total variance across datasets (PC1 = 46%, PC2 = 18%). Interestingly, we observed a consistent pattern across datasets in terms of how individual FPs contributed to the coverage of PCA space. Specifically, the FPs tended to align with similar directions in PCA space regardless of the dataset, suggesting a stable, dataset-independent contribution pattern. This consistency suggests that different fingerprints systematically capture distinct structural features. Moreover, complementarity across FPs underscores the advantage of their fusion: no single fingerprint captures the full breadth of structural variability, and integrating them can yield a more holistic molecular representation.

For example, in the FishTox dataset (Figure 3.2), the first two principal components explained 53.37% of the variance (PC1 = 33.89%, PC2 = 19.48%). The PCA loadings show strong orthogonal contributions from APC and KRC fingerprints along PC1 and PC2, respectively. In contrast, SSC and Estate fingerprints contributed less, while PubChem and MACCS fingerprints had the lowest influence. Notably, PubChem and MACCS are the only bit-vector

fingerprints among the set, whereas the others are count-based. This difference in vector type may explain their lower variance contribution. Additionally, we did not observe a consistent relationship between fingerprint vector length and chemical space coverage.

These six molecular FPs differ in how they encode chemical structure information. MACCS, PubChem, and SSC FPs focus on the presence of predefined chemical features or functional groups, such as rings or specific atoms. However, they complement each other because they use different sets of predefined substructures and encoding schemes, leading to a varied coverage of structural features. Atom Pair fingerprints capture the relationships between atom types and the topological distances between them, thus encoding aspects of molecular shape and connectivity. Klekota-Roth fingerprints use a much larger and more detailed set of structural fragments, offering a more comprehensive representation of molecular substructure diversity. In contrast, EState fingerprints differ fundamentally by integrating both electronic and topological information, characterizing atoms based on their electronic environment within the molecule. Together, these fingerprints provide diverse and complementary ways to capture chemical structure.

Because PCA is unsupervised and does not depend on the prediction target, it provides an unbiased view of how different FPs encode chemical information. The fact that individual FPs showed consistent loading directions across multiple datasets indicates that each fingerprint systematically captures specific and complementary features of chemical structure. This consistency across datasets indicates that different FPs capture largely non-overlapping aspects of molecular structure, and their combination provides a more complete and structured representation of chemical space.

3.3.2 Individual fingerprint-based models

To evaluate the target-specific predictive performance of individual FPs, independent from their PCA space coverage, we trained, optimized, validated, and tested six RF models per dataset, each using a different FP as input. Model performance was assessed using the coefficient of determination (R^2) and root mean squared error (RMSE) for regression tasks, and area under the curve (AUC) and F1-score for classification (Table 3.1).

In contrast to the unsupervised PCA analysis, which provided insight into chemical space coverage, the supervised RF models did not reveal a consistent trend in predictive

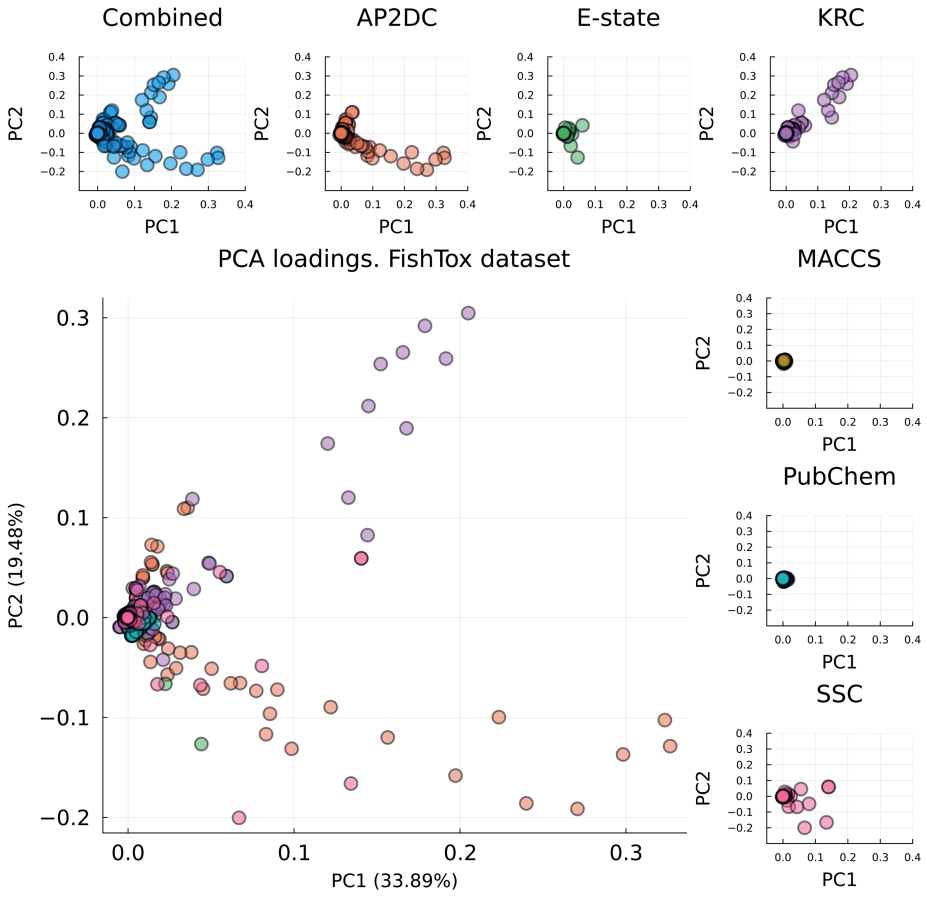


Figure 3.2: PCA loadings plot of combined fingerprint for FishTox dataset and contribution of each individual fingerprint.

performance based on the type of FP (bit vector vs. count vector). PubChem FP, a bit vector, outperformed the rest of the individual non-hashed FPs for FishTox and Koc datasets. For RI dataset, the best performance was achieved with the SSC and APC, count vector FPs. For ApisTox, strong performance came from KRC, SSC, count vectors, and MACCS, a short 166 bit vector. For the BACE dataset, E-state, the shortest count vector, and SSC showed the best predictive performance. Finally, for BBBP the best results were archived with APC, a count vector, and PubChem, a bit vector. However, it is important to note that overall predictive performance remained modest across all models. These results indicate that the level of structural detail or FP format, count or bit vectors, encoded does not consistently translate to better predictive performance. Furthermore, no association was found between fingerprint length and model performance, implying that not all features encoded by a fingerprint are relevant to specific activity predictions. Instead, irrelevant or redundant structural features may introduce noise into the model, ultimately reducing its predictive power.

Moreover, no single FP is sufficient to fully represent molecular structure, as each captures distinct aspects of chemical information. However, the features identified as important across different FPs are often complementary. For example, the APC FP encodes atom pair occurrences (e.g., C-C, C-X, C-Cl, C-O, C-N, N-O, O-O) at various distances, but does not capture bond types. In contrast, the MACCS fingerprint emphasizes bond types rather than atom identity. The predictive power of the E-State fingerprint primarily stems from its encoding of electronic and topological features related to aromaticity, which are less emphasized in other FPs. KRC highlights diverse oxygen-containing substructures, while SSC-derived models rely heavily on features such as the number of chiral centres and rotatable bonds.

It is important to note that all the FPs used in this study are based on 2D molecular representations and thus do not capture 3D conformational or stereochemical information. This limitation may constrain their ability to model QSAR for targets sensitive to stereochemistry.

3.3.3 Fingerprints fusion performance

We investigated how different fusion strategies (low-, mid-, and high-levels) affect the predictive power of models built using non-hashed fingerprints (Figure 3.3). Performance dif-

Table 3.1: Performance assessment of the individual fingerprint-based models evaluated on the test set.

Dataset	Metric	Fingerprints					
		KRC	Estate	SSC	MACCS	APC	PubChem
Regression							
FishTox	R2	0.396	0.494	0.484	0.418	0.482	**0.527**
	RMSE	1.291	1.182	1.194	1.267	1.196	**1.142**
Koc	R2	0.528	0.700	0.679	0.695	0.748	**0.770**
	RMSE	0.910	0.726	0.752	0.732	0.664	**0.635**
RIs	R2	0.656	0.654	**0.692**	0.651	**0.695**	0.672
	RMSE	189.3	189.8	**179.0**	190.7	**178.1**	184.6
Classification							
ApisTox	F1	**0.723**	0.711	**0.720**	**0.723**	0.622	0.681
	AUC	**0.830**	0.841	**0.868**	**0.867**	0.820	0.854
BACE	F1	0.840	**0.872**	**0.867**	0.814	0.861	0.794
	AUC	0.916	**0.945**	**0.930**	0.900	0.939	0.911
BBBP	F1	0.935	0.936	0.932	0.929	**0.954**	**0.942**
	AUC	0.947	0.943	0.950	0.949	**0.947**	**0.958**

ferences were interpreted in terms of consistency and relative trends rather than statistical significance.

Mid-level fusion, particularly with features selected based on 95% cumulative importance, yielded the best average performance across tasks. Low-level fusion showed improvements over individual FPs in several datasets. However, its performance was on average lower than that of the best mid-level fused models. This may be due to the inclusion of redundant features, as no feature selection was applied prior to concatenation. In contrast, high-level fusion, which aggregates predictions from individual models, was generally less effective. On average, it was outperformed by individual FPs such as APC and PubChem. Since high-level fusion works at the decision level, this completely misses potential synergies or interactions between features across different FPs. Thus, the simple averaging form of high-level fusion likely underperformed because it could not exploit cross-FP feature interactions, but also treated weak models equally.

In regression tasks, fused FPs showed clear advantages. For instance, in the FishTox dataset, PubChem reached an R^2 of 0.53, while the 95% mid-level fused FP achieved 0.62. In the Koc dataset, the top individual R^2 was 0.77 (PubChem), compared to 0.81 for a fused FP using a 0.4% individual importance threshold. Similarly, for the RI dataset, APC yielded 0.70, while the fused model reached 0.79.

Classification tasks showed a similar trend. In ApisTox, the best individual F1-score was 0.72 (KRC and MACCS), while the fused 95% model reached 0.81. For BACE, Estate and SSC achieved F1 scores of 0.87, while the fused 1.8% model improved to 0.90. In the BBBP dataset, where individual models already performed strongly (e.g., APC with an F1 of 0.95), fusion strategies offered only marginal improvements, with F1-score ranging from 0.94 to 0.96. Such improvements should be interpreted cautiously, as they may fall within the expected variability of model training and data splitting. In some cases, fused models matched or slightly underperformed compared to the top individual FPs. For example, in ApisTox, fused variants with 1.4%, 70%, 90%, and 95% importance matched the best F1-score of 0.72, while others (1.8%, 2.0%) were slightly lower.

HCA of normalized metrics of models performance reveals two main clusters (Figure 3.3). The first one is a group comprising the better-performing fused models specifically, including mid-level fused optimized FP from 75% and 95% cumulative importance. While the second is a broader cluster, including individual FPs, high-level fusion, and low-importance threshold optimized models.

Among all models, the optimized 95% fused FP performed best on average across the six datasets (Figure 3.3). While low- and high-levels fusion methods slightly improved over individual FPs, mid-level fusion proved to be the most effective approach, improving model performance in most cases. These findings highlight the value of combining complementary features, particularly through selective fusion strategies.

However, no single fusion strategy proved universally optimal across all datasets, which shows the need for case specific optimization approach. Furthermore, while RF algorithm was selected to ensure interpretability and enable controlled comparison of fusion strategies, the outcome of the fused FPs is specific to tree-based QSAR workflows and may not be generalizable to other machine learning algorithms (e.g., SVM, neural networks).

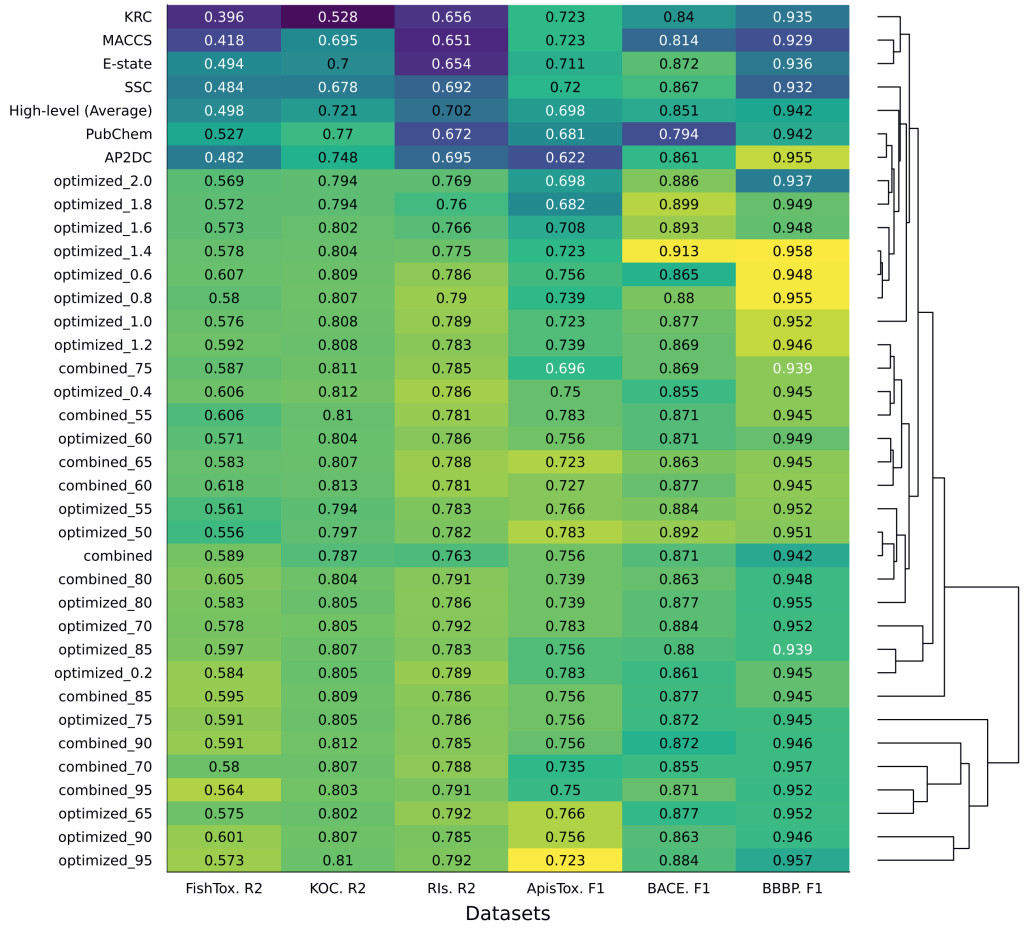


Figure 3.3: Test set performance of models trained and optimized on individual and fused molecular fingerprints. "High-level (Average)" represents high-level fusion, where predictions from individual models are averaged. "combined$_{55-95}$" refers to low-level fusion, where fingerprints are concatenated and features are selected based on cumulative feature importance between 55% and 95%. "optimized$_{55-95}$" corresponds to mid-level fusion, using features extracted from individually trained models based on the same 55 - 95% importance threshold. "optimized$_{0.2-2}$" also represents mid-level fusion, using features selected from individual models with importance scores ranging from 0.2% to 2%. Results clustered by HCA across six datasets, based on normalized R^2 (for regression tasks) and F1-score (for classification tasks). The heatmap gradient is based on normalized performance metrics for each column independently, allowing for relative performance comparison within each dataset.

3.3.4 Feature Composition of Optimized Fingerprints

Despite the variability in performance across individual tasks, a mid-level fused FP, composed of features cumulatively accounting for 95% of model importance, achieved the best average performance across datasets. While it was not the top-performing strategy for every case, its overall robustness suggests it can serve as a strong baseline for fusion design. We further investigated its feature composition across different prediction tasks. The goal was to examine how shared and task-specific features contribute to predictive performance, and to identify structural patterns that are consistently informative across different prediction tasks.

From the total number of 7,800 features from a cumulative set of six FPs, a subset of 444 important bits were selected for each target: 50 from APC, 12 from E-State, 149 from KRFPC, 68 from MACCS, 142 from PubChem, and 23 from SubFPs. Among these, 96 bits specifically describe different aspects of the carbon backbone. These include 8 bits from APC, 7 from E-State, 45 from KRFPC, 16 from MACCS, 14 from PubChem, and 6 from SubFPs. Each individual FP contributed unique representations of carbon-based features. For example, KRFPC encodes different carbon chain lengths with specific bits such as KRFPC3224 ("CC"), KRFPC3640 ("CCC"), and KRFPC3692 ("CCCC"). In contrast, MACCS and SubFPs represent carbon carbon branching characteristics: SubFPC4 for primary carbons and SubFPC2 for secondary carbons. PubChem fingerprints capture information with bits like PubChemFP696, which represents long continuous carbon chains such as "C-C-C-C-C-C-C-C". E-State descriptors provide information on electronic environments, such as bit 19 for ssssC (quaternary carbon) and bit 13 for sssCH (tertiary carbon with one hydrogen). APC captures the presence of C-C bonds at various topological distances, specifically distances 1 through 5 and 9. The remaining bits describe features involving aromaticity and heteroatoms, such as oxygen, nitrogen, chlorine, sulfur, and fluorine in a variety of substructural contexts. These include aromatic and saturated ring systems, heterocycles, and functional groups. Some fingerprints, such as MACCS, include more generalized representations. For instance, MACCSFP107 ("XA(A)A") describes the structure where "X" represents any halogen atom (F, Cl, Br, I), and "A" denotes any atom type.

On the other hand, the extracted bits which correspond to prediction of specific target focus more on specific functional groups or heteroatoms. For example, the FishTox

model highlights halogenated features such as C–Cl, O–Cl, and Cl–Cl, along with sulfur-containing atoms and aromatic chlorine or bromine substructures, as well as nitro groups, phenols, diaryl ethers, and contributions from long alkyl chains. The KOC model emphasizes halogenation: particularly C–Br and polychlorinated aromatic rings along with large aromatic systems, carboxylic acids, and ethers. The ApisTox model shows signals for phosphorus-based substructures (e.g., C-P, N-P, O-P), as well as N-O and N-S linkages. The RI model is influenced by nitro groups, amines, and long alkyl chains, with features including N-O, O-O, and C-N bonds, high hydrogen and carbon content, and aromatic amines. The BACE model highlights fluorinated and nitrogen-rich substructures such as C-F, N-F, N-N, and N-O, along with amines and fluorinated aromatics. Finally, the BBBP model shares several features with the BACE model, particularly nitrogen- and oxygen-containing substructures (e.g., C-N, N-O, C-S), but also includes peptide-like and carboxylic acid substructures. Overall, these models reflect distinctive substructure patterns associated with their respective prediction tasks.

To further evaluate the impact of individually selected fingerprint bits, we trained models using a shared subset of common bits selected across all tasks, referred to as common bits. Their performance was then compared to that of models trained on task-specific fingerprints optimized at the 95% cumulative importance level. Across the six datasets, the optimized 95% fingerprint consistently demonstrated either improved or comparable predictive performance (Section 3.5 Table S3). The most notable gain was observed in the ApisTox dataset, where the F1-score increased from 0.696 to 0.809. FishTox also exhibited modest improvements, while performance for KOC, RIs, and BBBP remained almost unchanged. These results indicate that the addition of specific to individual tasks fingerprint bits can enhance predictive accuracy. Importantly, the common fingerprint bits were selected based on shared importance across all six datasets, suggesting that this representation captures structural features broadly relevant to diverse prediction tasks.

3.3.5 Limitations

Although we combined molecular FPs on different fusion level, certain important structural information for target prediction could not be fully captured. For example, most existing fingerprinting algorithms do not encode three-dimensional (3D) molecular characteristics, such as chirality. This limits the ability of QSAR models to differentiate the activ-

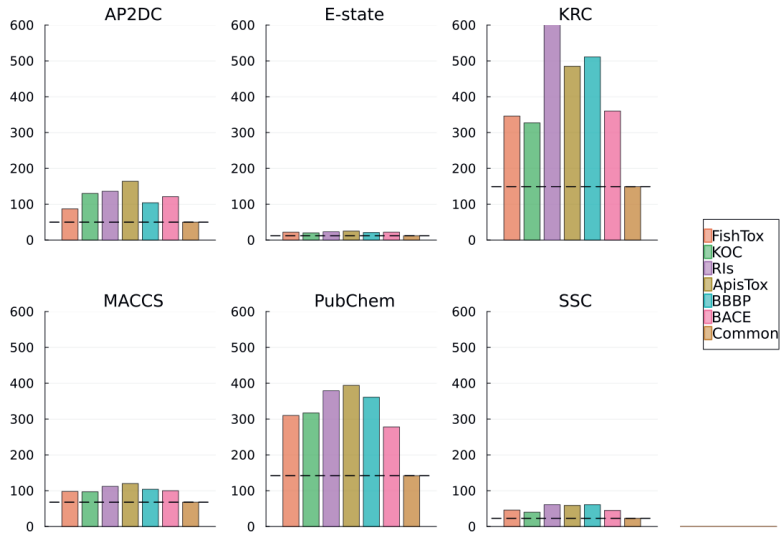


Figure 3.4: Number of bits from each individual fingerprint contributing to the optimized 95% fingerprint (per dataset), and number of shared bits from each fingerprint present in all six optimized 95% datasets.

ity of stereoisomers. On the other hand, features like charge distribution are not explicitly represented in standard FPs, although they may be indirectly captured through correlated structural elements.

Furthermore, even the most carefully curated combination of fingerprint keys has limited structural coverage. As a result, highly specific or uncommon substructures associated with niche activity classes may still be underrepresented. This implies that expanding the diversity of the training dataset could enrich the final fused FP, allowing for broader coverage of molecular features relevant to specific targets. Alternatively, ML-based FPs (e.g. MolBERT [56]), for instance, the ones generated by graph neural networks, can automatically learn relevant structural representations from the data itself, potentially capturing rare or complex features missed by predefined fingerprints. However, this advantage comes at the cost of limited interpretability, often making it difficult or even impossible to trace specific features back to identifiable chemical substructure.

In this study, we focused exclusively on non-hashed fingerprints to preserve interpretability and ensure consistent bit-to-substructure alignment across molecules. This alignment is critical for traditional machine learning models such as Random Forests and SVMs, where each feature (bit) is assumed to have a consistent and interpretable meaning.

Hashed fingerprints, in contrast, compress structural information into fixed-length vectors using hashing functions, which can lead to bit collisions—i.e., different substructures mapping to the same bit. This results in a loss of alignment and can introduce noise, reducing model reliability and interpretability. As a result, hashed fingerprints are often better suited for deep learning architectures, which can learn hierarchical representations even when individual features lack clear chemical meaning. Nevertheless, alternative molecular representations, such as circular fingerprints (e.g., ECFP), physicochemical descriptors, or graph-based encodings may still offer complementary information. Integrating these with our current approach could improve predictive performance and offer new opportunities for multi-level fingerprint fusion.

Although our study included six diverse datasets and targets, results may vary for other tasks or datasets with different chemical distributions. Therefore, the conclusions regarding fusion strategies should be interpreted with respect to the data and endpoints investigated here. Moreover, like many widely used benchmarks the selected datasets may have their limitations. Importantly, in this study we do not propose a new state-of-the-art predictor any of the targets, but rather analyze the relative behavior of fingerprint fusion strategies across a range of diverse supervised predictive tasks.

Lastly, the study was limited to the use of RF models to ensure interpretability and a controlled comparison of fusion strategies. Although the proposed fingerprint fusion framework is not algorithm-specific and can be combined and applied to other learning algorithms, the observed performance gains may not directly generalize across different approaches.

3.4 Conclusions

Molecular representations play a crucial role in achieving the best accuracy of predictive modeling. Non-hashed molecular fingerprints, in particular, offer consistent, human-readable, and computationally efficient structural descriptors, making them valuable tools for assessing chemical behavior. However, individual fingerprint representations may not fully capture all relevant structural information needed for robust QSAR performance. To address this, we explored the impact of various fingerprint fusion strategies, including low-, mid-, and high-level, on QSAR model performance.

Our assessment demonstrated that all fusion approaches, except high-level fusion, outperformed models trained on individual fingerprints. This confirms that combining complementary structural features enhances predictive capability. Among the tested strategies, mid-level fusion resulted in the best average performance across six diverse datasets and tasks. In this approach, fingerprint bits with cumulative importance up to 95% were selected from each individual fingerprint and then concatenated to form a conjoint representation. This strategy not only preserved the interpretability of features but also ensured that the most informative structural elements were retained. The mid-level fusion algorithm successfully captured relevant molecular features, particularly for predicting toxicity-related endpoints. Moreover, its flexibility allows application to a wide range of case studies, although the magnitude of performance improvement remains dataset- and endpoint-dependent. By focusing on interpretability and predictive strength, the proposed algorithm represents a practical and generalizable solution for enhancing QSAR modeling through optimized fingerprint fusion.

3.5 Supporting Information

The Supporting Information with additional details on selected non-hashed molecular fingerprints (Table S1), evaluation metrics (Equations S1–S5), PCA loadings and score plots for each dataset (Figures S1–S5), and the performance of the model trained only on common bits (Table S2) is available at `https://doi.org/10.1039/d5dd00302d`.

Notes

All datasets as well as calculated FPs, in this study, and interpretation of the optimized FP can be found at `https://doi.org/10.5281/zenodo.15791757`. The algorithm for a molecular fingerprint fusion was converted into an open-source open-access Julia package and can be found at `https://bitbucket.org/viktoriiaturkina/fp_optimization.jl/src/master/`

Acknowledgements

The authors express their gratitude to all the members of Environmental Modeling & Computational Mass Spectrometry (www.emcms.info). They are also thankful to Marco Federici, Denice van Herwerden, and Adan Rotteveel for their kind help and contribution to this study. S.S., A.P., and V.T. thank the ChemistryNL TKI and the UvA Data Science Centre for their funding support (projects EDIFIED and SCOPE). J.W.O.B. is the recipient of a National Health and Medical Research Council (NHMRC) Investigator Grant (EL12009209) funded by the Australian Government. S.S. also acknowledges financial support from the Australian National Health and Medical Research Council (NHMRC; APP1185347). The Queensland Alliance for Environmental Health Sciences (QAEHS), The University of Queensland (UQ), gratefully acknowledges the financial support of Queensland Health. For the purposes of open access, the author has applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising from this submission.

3.6 References

[1] Arp H. P. H, Aurich D, Schymanski E. L, Sims K, Hale S. E. Avoiding the Next Silent Spring: Our Chemical Past, Present, and Future. Environ Sci Technol. 2023 4;57(16):6355-9. doi:10.1021/acs.est.3c01735.

[2] Wang Z, Walker G. W, Muir D. C. G, Nagatani-Yoshida K. Toward a Global Understanding of Chemical Pollution: A First Comprehensive Analysis of National and Regional Chemical Inventories. Environ Sci Technol. 2020 3;54(5):2575-84. doi:10.1021/acs.est.9b06379.

[3] Escher B. I, Stapleton H. M, Schymanski E. L. Tracking complex mixtures of chemicals in our changing environment. Science. 2020;367(6476):388-92. doi:10.1126/science.aay6636.

[4] Vermeulen R, Schymanski E. L, Barabási A. L, Miller G. W. The exposome and health: where chemistry meets biology. Science. 2020 1;367(6476):392. doi:10.1126/science.aay3164.

[5] Rudén C, Hansson S. O. Registration, Evaluation, and Authorization of Chemicals (REACH) Is but the First Step–How Far Will It Take Us? Six Further Steps to Improve the European Chemicals Legislation. Environ Health Perspect. 2010 1;118(1):6. doi:10.1289/ehp.0901157.

[6] Muir D. C. G, Howard P. H. Are there other persistent organic pollutants? A challenge for environmental chemists. Environ Sci Technol. 2006 12;40(23):7157-66. doi:10.1021/es061677a.

[7] Cronin M. T. D, Jaworska J. S, Walker J. D, Comber M. H. I, Watts C. D, Worth A. P. Use of QSARs in international decision-making frameworks to predict health effects of chemical substances. Environ Health Perspect. 2003 8;111(10):1391. doi:10.1289/ehp.5760.

[8] Howard P. H, Muir D. C. G. Identifying new persistent and bioaccumulative organics among chemicals in commerce II: Pharmaceuticals. Environ Sci Technol. 2011 8;45(16):6938-46. doi:10.1021/es201196x.

[9] Butler K. T, Davies D. W, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. Nature 2018 559:7715. 2018 7;559(7715):547-55. doi:10.1038/s41586-018-0337-2.

[10] Sigurnjak Bureš M, Cvetnić M, Miloloža M, Kučić Grgić D, Markić M, Kušić H, et al. Modeling the toxicity of pollutants mixtures for risk assessment: a review. Environ Chem Lett. 2021 1;19(2):1629-55. doi:10.1007/s10311-020-01107-5.

[11] Roy K, Das R. N, Popelier P. L. A. Quantitative structure–activity relationship for toxicity of ionic liquids to Daphnia magna: Aromaticity vs. lipophilicity. Chemosphere. 2014 10;112:120-7. doi:10.1016/j.chemosphere.2014.04.002.

[12] Karelson M, Lobanov V. S, Katritzky A. R. Quantum-chemical descriptors in QSAR/QSPR studies. Chem Rev. 1996;96(3):1027-43. doi:10.1021/cr950202r.

[13] Liu H, Papa E, Gramatica P. QSAR prediction of estrogen activity for a large set of diverse chemicals under the guidance of OECD principles. Chem Res Toxicol. 2006 11;19(11):1540-8. doi:10.1021/tx0601509.

[14] Gedeck P, Rohde B, Bartels C. QSAR - How good is it in practice? Comparison of descriptor sets on an unbiased cross section of corporate data sets. J Chem Inf Model. 2006;46(5):1924-36. doi:10.1021/ci050413p.

[15] Melnikov F, Kostal J, Voutchkova-Kostal A, Zimmerman J. B, Anastas P. T. Assessment of predictive models for estimating the acute aquatic toxicity of organic chemicals. Green Chem. 2016 8;18(16):4432-45. doi:10.1039/c6gc00720a.

[16] Mao J, Akhtar J, Zhang X, Sun L, Guan S, Li X, et al. Comprehensive strategies of machine-learning-based quantitative structure-activity relationship models. iScience. 2021 9;24(9):103052. doi:10.1016/j.isci.2021.103052.

[17] Cereto-Massagué A, Ojeda M. J, Valls C, Mulero M, Garcia-Vallvé S, Pujadas G. Molecular fingerprint similarity search in virtual screening. Methods. 2015;(71):58-63. doi:10.1016/j.ymeth.2014.08.005.

[18] Grisoni F, Ballabio D, Todeschini R, Consonni V. Molecular descriptors for structure–activity applications: A hands-on approach. In: Methods Mol. Biol.. vol. 1800. Humana Press Inc.; 2018. p. 3 53. doi:10.1007/978-1-4939-7899-1_1.

[19] Boulougouri M, Vandergheynst P, Probst D. Molecular set representation learning. Nat Mach Intell. 2024 7;6(7):754-63. doi:10.1038/s42256-024-00856-0.

[20] Muegge I, Mukherjee P. An overview of molecular fingerprint similarity search in virtual screening. Expert Opin Drug Discov. 2016 2;11(2):137-48. doi:10.1517/17460441.2016.1117070.

[21] Czarnecki W. M, Podlewska S, Bojarski A. J. Robust optimization of SVM hyperparameters in the classification of bioactive compounds. J Cheminform. 2015 8;7(1):1-15. doi:10.1186/s13321-015-0088-0.

[22] Huang G, Li J, Zhao C. Computational Prediction and Analysis of Associations between Small Molecules and Binding-Associated S-Nitrosylation Sites. Molecules 2018, Vol 23, Page 954. 2018 4;23(4):954. doi:10.3390/molecules23040954.

[23] Warszycki D, Struski L, Smieja M, Kafel R, Kurczab R. Pharmacoprint: A Combination of a Pharmacophore Fingerprint and Artificial Intelligence as a Tool for Computer-Aided Drug Design. J Chem Inf Model. 2021 10;61(10):5054-65. doi:10.1021/acs.jcim.1c00589.

[24] Rogers D, Hahn M. Extended-Connectivity Fingerprints. J Chem Inf Model. 2010 5;50(5):742-54. doi:10.1021/ci100050t.

[25] Cronin M. T. D, Bauer F. J, Bonnell M, Campos B, Ebbrell D. J, Firman J. W, et al. A scheme to evaluate structural alerts to predict toxicity – Assessing confidence by characterising uncertainties. Regul Toxicol Pharmacol. 2022 11;135:105249. doi:10.1016/j.yrtph.2022.105249.

[26] Tseng Y. J, Hopfinger A. J, Esposito E. X. The great descriptor melting pot: Mixing descriptors for the common good of QSAR models. J Comput Aided Mol Des. 2012 12;26(1):39-43. doi:10.1007/s10822-011-9511-4.

[27] Hert J, Willett P, Wilton D. J, Acklin P, Azzaoui K, Jacoby E, et al. New Methods for Ligand-Based Virtual Screening: Use of Data Fusion and Machine Learning to Enhance the Effectiveness of Similarity Searching. J Chem Inf Model. 2006;46(2):462-70. doi:10.1021/ci050348j.

[28] Xie L, Xu L, Kong R, Chang S, Xu X. Improvement of Prediction Performance With Conjoint Molecular Fingerprint in Deep Learning. Front Pharmacol. 2020 12;11:606668. doi:10.3389/fphar.2020.606668.

[29] Srisongkram T, Khamtang P, Weerapreeyakul N. Prediction of KRASG12C inhibitors using conjoint fingerprint and machine learning-based QSAR models. J Mol Graph Model. 2023 7;122:108466. doi:10.1016/j.jmgm.2023.108466.

[30] Shen W. f, Tang H. w, Li J. b, Li X, Chen S. Multimodal data fusion for supervised learning-based identification of USP7 inhibitors: a systematic comparison. J Cheminform. 2023 12;15(1):1-16. doi:10.1186/s13321-022-00675-8.

[31] Cai H, Zhang H, Zhao D, Wu J, Wang L. FP-GNN: a versatile deep learning architecture for enhanced molecular property prediction. Brief Bioinform. 2022 11;23(6):1-12. doi:10.1093/bib/bbac408.

[32] Chuang K. V, Gunsalus L. M, Keiser M. J. Learning Molecular Representations for Medicinal Chemistry. J Med Chem. 2020 8;63(16):8705-22. doi:10.1021/acs.jmedchem.0c00385.

[33] Matsuyama Y, Ishida T. Stacking Multiple Molecular Fingerprints for Improving Ligand-Based Virtual Screening. In: Lect. Notes Comput. Sci.. vol. 10955 LNCS. Springer Verlag; 2018. p. 279-88. doi:10.1007/978-3-319-95933-7_35.

[34] Williams A. J, Grulke C. M, Edwards J, McEachran A. D, Mansouri K, Baker N. C, et al. The CompTox Chemistry Dashboard: A community data resource for environmental chemistry. J Cheminform. 2017 11;9(1):1-27. doi:10.1186/s13321-017-0247-6.

[35] Ding W, Nan Y, Wu J, Han C, Xin X, Li S, et al. Combining multi-dimensional molecular fingerprints to predict the hERG cardiotoxicity of compounds. Comput Biol Med. 2022 5;144:105390. doi:10.1016/j.compbiomed.2022.105390.

[36] Weininger D. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. J Chem Inf Comput Sci. 1988 2;28(1):31-6. doi:10.1021/ci00057a005.

[37] Smith D. H, Carhart R. E, Venkataraghavan R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. J Chem Inf Comput. 1985 5;25(2):64-73. doi:10.1021/ci00046a002.

[38] Hall L. H, Kier L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. J Chem Inf Comput. 1995;35(6):1039-45. doi:10.1021/ci00028a014.

[39] Klekota J, Roth F. P. Chemical substructures that enrich for biological activity. Bioinformatics. 2008;24(21):2518-25. doi:10.1093/bioinformatics/btn479.

[40] Durant J. L, Leland B. A, Henry D. R, Nourse J. G. Reoptimization of MDL keys for use in drug discovery. J Chem Inf Comput. 2002 11;42(6):1273-80. doi:10.1021/ci010132r.

[41] Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res. 2021 1;49(D1):D1388-95. doi:10.1093/nae/gkaa971.

[42] Yap C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. J Comput Chem. 2011 5;32(7):1466-74. doi:10.1002/jcc.21707.

[43] RDKit: Open-source cheminformatics Software (accessed 23/01/2024); 2024. doi:https://doi.org/10.5281/zenodo.10633624.

[44] Samanipour S, O'Brien J. W, Reid M. J, Thomas K. V, Praetorius A. From Molecular Descriptors to Intrinsic Fish Toxicity of Chemicals: An Alternative Approach to Chemical Prioritization. Environ Sci Technol. 2023 11;57(46):17950-8. doi:10.1021/acs.est.2c07353.

[45] Wang Y, Chen J, Yang X, Lyakurwa F, Li X, Qiao X. In silico model for predicting soil organic carbon normalized sorption coefficient (KOC) of organic chemicals. Chemosphere. 2015 1;119:438-44. doi:10.1016/j.chemosphere.2014.07.007.

[46] Aalizadeh R, Nikolopoulou V, Thomaidis N. S. Development of Liquid Chromatographic Retention Index Based on Cocamide Diethanolamine Homologous Series (C(n)-DEA). Anal Chem. 2022 11;94(46):15987-96. doi:10.1021/acs.analchem.2c02893.

[47] Adamczyk J, Poziemski J, Siedlecki P. ApisTox: a new benchmark dataset for the classification of small molecules toxicity on honey bees. Sci Data. 2025 1;12(1):1-15. doi:10.1038/s41597-024-04232-w.

[48] Martins I. F, Teixeira A. L, Pinheiro L, Falcao A. O. A Bayesian approach to in Silico blood-brain barrier penetration modeling. J Chem Inf Model. 2012 6;52(6):1686-97. doi:10.1021/ci300124c.

[49] Subramanian G, Ramsundar B, Pande V, Denny R. A. Computational Modeling of β-Secretase 1 (BACE-1) Inhibitors Using Ligand Based Approaches. J Chem Inf Model. 2016 10;56(10):1936-49. doi:10.1021/acs.jcim.6b00290.

[50] Cassotti M, Ballabio D, Todeschini R, Consonni V. A similarity-based QSAR model for predicting acute toxicity towards the fathead minnow (Pimephales promelas). SAR QSAR Environ Res. 2015 3;26(3):217-43. doi:10.1080/1062936x.2015.1018938.

[51] Pedregosa F, Michel V, Grisel O, Blondel M, Prettenhofer P, Weiss R, et al.. Scikit-learn: Machine Learning in Python; 2011. `http://scikit-learn.sourceforge.net.`

[52] Svetnik V, Liaw A, Tong C, Christopher Culberson J, Sheridan R. P, Feuston B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. J Chem Inf Comput. 2003 11;43(6):1947-58. doi:10.1021/ci034160g.

[53] van Herwerden D, O'Brien J. W, Lege S, Pirok B. W. J, Thomas K. V, Samanipour S. Cumulative Neutral Loss Model for Fragment Deconvolution in Electrospray Ionization High-Resolution Mass Spectrometry Data. Anal Chem. 2023 8;95(33):12247-55. doi:10.1021/acs.analchem.3c00896.

[54] Yang F, van Herwerden D, Preud'homme H, Samanipour S. Collision Cross Section Prediction with Molecular Fingerprint Using Machine Learning. Molecules. 2022 10;27(19). doi:10.3390/molecules27196424/s1.

[55] Polishchuk P. G, Muratov E. N, Artemenko A. G, Kolumbin O. G, Muratov N. N, Kuz'min V. E. Application of random forest approach to QSAR prediction of aquatic toxicity. J Chem Inf Model. 2009 11;49(11):2481-8. doi:10.1021/ci900203n.

[56] Wen N, Liu G, Zhang J, Zhang R, Fu Y, Han X. A fingerprints based molecular property prediction method using the BERT model. J Cheminform. 2022 12;14(1):71. doi:10.1186/s13321-022-00650-3.

4

Prioritization of unknown LC-HRMS features based on predicted toxicity categories

Authors: Viktoriia Turkina, Jelle T. Gringhuis, Annemieke Petrignani, Garry Corthals, Antonia Praetorius, Jake W. O'Brien, Saer Samanipour

Published in: Environ. Sci. Technol. 2025, 59, 16, 8004–8015.

https://doi.org/10.1021/acs.est.4c13026

Abstract

Complex environmental samples contain a diverse array of known and unknown constituents. While Liquid Chromatography coupled with High-Resolution Mass Spectrometry (LC-HRMS) Non-Targeted Analysis (NTA) has emerged as an essential tool for the comprehensive study of such samples, the identification of individual constituents remains a significant challenge, primarily due to the vast number of detected features in each sample. To address this, prioritization strategies are frequently employed to narrow the focus to the most relevant features for further analysis.

In this study, we developed a novel prioritization strategy that directly links fragmentation and chromatographic data to aquatic toxicity categories, bypassing the need for individual compound identification. Given that features are not always well-characterized through fragmentation, we created two models: 1) a Random Forest Classification (RFC) model, which classifies fish toxicity categories based on MS^1, retention, and fragmentation data - expressed as cumulative neutral losses (CNLs) - when fragmentation information is available, and 2) a Kernel Density Estimation (KDE) model that relies solely on retention time and MS^1 data when fragmentation is absent. Both models demonstrated accuracy comparable to structure-based prediction methods. We further tested the models on a pesticide mixture in a tea extract measured by LC-HRMS, where the CNLs-based RFC model achieved 0.76 accuracy and the KDE model reached 0.61, showcasing their robust performance in real-world applications.

4.1 Introduction

The increasing number and variety of produced chemicals hinders the investigation and assessment of their exposure levels and impacts on human and environmental health. [1–3] Only specific groups of chemicals are routinely monitored as a regulatory measure, therefore, limiting the exposure assessment by excluding a diverse mixture of chemicals. [4, 5] In order to investigate chemical exposure realistically, accounting for the diversity of structures and their properties in samples necessitates the application of unbiased approaches. [5–7]

A major portion of contaminants of emerging concern (CECs) is covered by semi-polar and polar organic compounds. [8, 9] High-resolution tandem mass spectrometry coupled with reverse phase liquid chromatography (LC-HRMS) is a well-known technique to analyze such chemicals. [9, 10] To limit the bias towards a small set of standardized analytes, the LC-HRMS nontargeted (NTA) approach has been successfully developed and employed in environmental studies. [11–16] NTA does not require prior knowledge of the chemicals present in a sample and enables the analysis of all detectable compounds. However, the complex nature of environmental samples results in highly convoluted data, requiring careful and robust preprocessing. [17–20] Generally the number of detected features in NTA, data points constructed by retention time, chromatographic peak intensity, and m/z of precursor ions, often counts in the hundreds or thousands, though typically fewer than 5% of these can be identified. [19, 21–23]

Confident identification of detected features requires considerable computational and research resources. [17,23,24] Therefore, prioritization is used as strategy to focus the available resources on the most relevant features for a given study's objectives. [25] The prioritized species of interest undergo more detailed and thorough investigation. There are two ways of prioritization: online and offline. Online prioritization refers to a variety of parameters to acquire the data, e.g. to perform data-dependent acquisition, [25] consequently, limiting the acquired data according to predefined parameters. [21] On the other hand, offline prioritization strategies employ post hoc analysis of MS^1 and MS^2 (both data-dependent (DDA) and data-independent (DIA) acquisition) data, including intensity-based prioritization, statistical analysis, structural, and quantitative structure-activity relationship (QSAR) evaluation. [6] The online approach results in a cleaner dataset for further processing with

the risk of losing valuable information about highly relevant features. Thus, the offline approach is more advantageous for environmental studies and can also be used for retrospective studies. [26]

Environmental studies primarily focus on the properties and potential activities of detected chemicals, including their ecological toxicity. [27, 28] As a result, prioritization strategies often aim to highlight features that may pose significant (eco)toxicological risks. [29] For example, Peets et al. developed the MS2Tox R package, which uses a regression model to predict toxicity (-[LOG(mM)]) for organisms like fish, water fleas, and algae. [30, 31] Similarly, Rahu et al. developed an algorithm to predict endocrine-disrupting activity using binary or multi-label classification, [32] while Arturi et al. created MLinvitroTox, a tool to classify chemicals as toxic or non-toxic based on nearly 400 target-specific and over 100 cytotoxic endpoints. [33]

All these studies rely on predicting molecular formulas and fingerprints from MS^2 spectra using tools like CSI:FingerID/SIRIUS. [34] Therefore, the accuracy of these models is closely tied to the precision of the predicted molecular fingerprints. Incorrectly calculated fingerprints can significantly distort the predictions, and this accuracy can be compromised by various experimental factors, such as incomplete fragment detection, false positive fragments, or missing characteristic fragments due to instrumental noise. Bypassing the molecular fingerprint prediction and instead predicting chemical activity directly from chromatographic and MS data potentially reduces this uncertainty and enhances the interpretability of results.

Aquatic toxicity has been shown to be related to hydrophobicity of the chemicals, their molecular weights and structural alerts. [35–37] Each detected feature indirectly includes structural information: chromatographic retention times and mass spectral (MS) information, like accurate mass (MS^1) and fragmentation spectra (MS^2). Therefore, this information can be used to assess the toxicity of the chemicals without any knowledge of molecular structure, formula or fingerprint.

In this study, we developed a prioritization strategy based on acute aquatic toxicity using chromatographic and fragmentation information of unknown features. Instead of identifying each feature, we directly described their activity without any knowledge of molecular structure. We constructed two models: a Random Forest Classification (RFC) model to assign toxicity categories based on fragmentation data, and a Kernel Density

Estimation (KDE) model to determine the probability of features belonging to specific toxicity classes. These algorithms allow us to characterize unknown features at both the mass spectral and chromatographic levels. The performance of the constructed models was evaluated both on spectra from databases with measured toxicity values and on tea extracts spiked with 253 pesticide standards.

4.2 Methods

Since fragmentation information in NTA is often incomplete, we developed a probabilistic algorithm to predict toxicity categories that relies solely on chromatographic data, specifically retention time and MS^1 spectra. This algorithm uses KDE in both the retention and mass domains to identify regions of chromatograms that are more likely to correspond to higher or lower toxicity, effectively mapping toxicity. Additionally in the presence of reliable fragmentation data, the CNLs-based RFC model can be applied to categorize toxicity based on detected fragments.

4.2.1 Overall workflow

To develop the models we employed a semi-supervised approach. First, a structure-based classification model was built using an experimentally generated acute fish toxicity dataset and molecular fingerprints. This model was then used to expand toxicity knowledge by predicting the toxicity category for chemicals from large, existing spectral and environmentally relevant databases. The CompTox dataset [38] was used to develop a KDE model, while for the CNLs-based model, HRMS spectra were collected from sources such as MassBank, [39] MoNa, [40] GNPS, [41] and NIST20. [42] Finally, the results of the developed models were compared and validated using an experimentally acquired pesticide mixture in a tea extract (Figure 4.1).

4.2.2 Datasets

For the model development, validation, and testing we employed four different datasets namely: a set of experimental acute fish toxicity values - referred to as the fish toxicity data

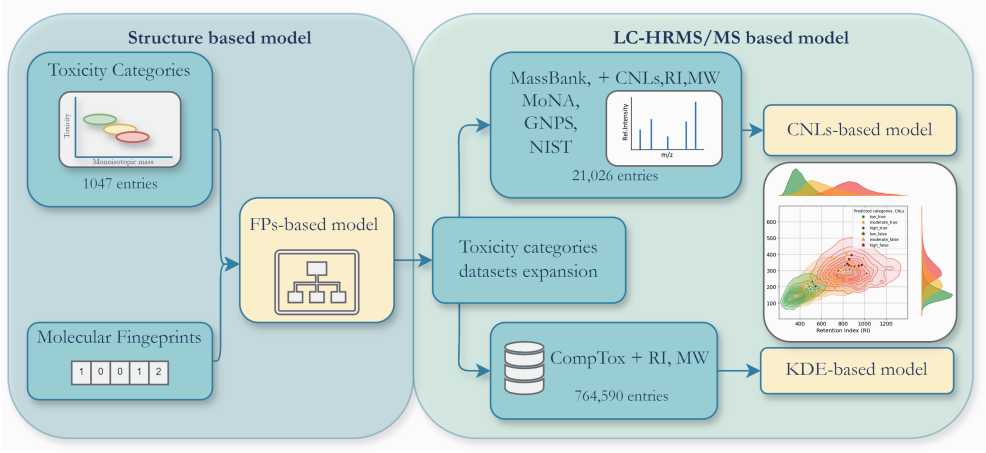


Figure 4.1: Overall workflow. A structure-based model was developed using a fish toxicity dataset and molecular fingerprints (FPs). This FPs-based model was then applied to the CompTox and Spectral databases, incorporating additional information such as molecular weight (MW) and retention indices (RI). A KDE-based model was constructed based on this data. Furthermore, fragmentation information (i.e., characteristic neutral losses, CNLs) was integrated to develop a CNLs-based model.

set, the CompTox dataset, a Cumulative Neutral Loss (CNL) dataset, and experimentally acquired LC-HRMS chromatogram of standard pesticide mixture in tea extracts - referred to as pesticide mixture dataset.

Fish toxicity dataset

The fish toxicity dataset used in this study was compiled from two sources: Cassotti et al. [43] and Schür et al. [44] These sources are themselves compilations of numerous empirical studies, relying on data collected from a wide range of experiments. The dataset includes concentration values, expressed in LOG(mg/L), representing the dose that causes 50% mortality in fathead minnows (Pimephales promelas) over a 96-hour exposure period (LC50). In total 1047 chemicals in the dataset come from various chemical families, including pharmaceuticals, pesticides, conventional persistent organic pollutants, and industrial chemicals. [45] Due to the inherent variability in *in vivo* experiments, where results can vary even under identical conditions, the dataset includes multiple repeated measurements for each chemical. The median value from these repetitions was used as the reference LC50 value for each compound.

The CompTox dataset

The CompTox Chemicals Dashboard, developed by the U.S. Environmental Protection Agency (EPA), is a vast database that has been curated over 15 years through both manual and automated efforts as part of the EPA's DSSTox project. It contains detailed information on 764,590 chemicals (as of March 2024), including identifiers such as CAS numbers and SMILES, synonyms, and key properties like solubility, logKow, and molecular weight. Additionally, it provides data on physicochemical characteristics, environmental fate, exposure, usage, and available toxicity results from both *in vivo* and *in vitro* studies. For our purposes, the CompTox dataset was filtered to exclude salts, metalloids, and compounds containing elements not typically found in fish toxicity datasets or not analyzable via reverse phased liquid chromatography (RPLC), leaving only chemicals composed of H, B, C, N, O, F, Si, P, S, Cl, Br, and I.

CNLs dataset

The CNLs were used in place of fragments to reduce the dimensionality of the dataset while retaining the structural information embedded in the spectra. The CNLs dataset was generated from entries in the MassBank EU, MoNA, GNPS, and NIST20 databases. These entries were obtained using electrospray ionization (positive mode) with a mass resolution of $\geq$ 5000, proton adducts, and included data from various mass analyzers (e.g., Q-TOF and Orbitrap) with different collision energies. To ensure data quality, only spectra with at least three recorded fragments were included. When multiple spectra were available for a single compound, they were merged using a ±0.01 Da mass window, yielding 21026 unique spectra. Rare fragments, present in only a single instance across multiple entries, were discarded to focus on more representative data.

For each compound, the CNLs were calculated by subtracting the fragment m/z values from the precursor ion mass. These CNLs were then converted into bit vectors representing masses from 0 to 1000 Da with a step size of 0.01 Da (± 5 mDa mass tolerance). In this representation, a value of 1 indicated the presence of a CNL in the spectrum, while a value of 0 indicated its absence. To further refine the dataset, CNLs that occurred in fewer than 100 spectra were removed, resulting in a final dataset of 2440 common CNLs, with mass values ranging from 1.03 to 382.19 Da.

In addition to the CNL bit vectors, the monoisotopic mass and predicted retention index were added as continuous features. To calculate monoisotopic mass, the mass of proton, 1.008 Da, was subtracted from the measured m/z value of the proton adduct. To predict retention indices the previously reported by van Herwerden et al. [46] model based on molecular fingerprints and retention indices calculated on the cocamide scale was employed. [47] This combination of features allowed us to incorporate key structural and chromatographic information into the model while minimizing the number of variables. To interpret the importance of individual CNLs, a search was conducted in the CompTox Chemicals Dashboard to identify relevant substructures. The prominent substructures frequently identified in the search results were used to inform model interpretation.

Pesticide mixture dataset

The prioritization models were further tested on experimentally acquired datasets of varying levels of tea extract matrix spiked with 253 pesticide standards. The samples were analysed by the previously reported RPLC-ESI-Q-TOF method. [48] The three standard mixtures of pesticides from Neochema were diluted with filtered water/ethanol solution (50%:50%, v/v, Blank) or filtered tea extract that was further diluted 1:10 or 1:100 with Blank to final analyte concentrations of 100 μg/L. Each sample was analysed twice. A detailed description of the samples, acquisition method, and data processing is provided in Supporting Information (Section 4.5) Section S1.

To obtain features and fragmentation information of detected pesticides, suspect screening was performed by applying the Universal Library Search Algorithm (ULSA). [49] A suspect list of the pesticides was compiled from the MS^2 spectra found in the MassBank EU, MoNA, and NIST20 databases, with a resolution higher than 5000 for each of these chemicals. To minimize detection of false positive fragments, a mass tolerance of 0.010 Da and a minimum MS^1 intensity of 2000 were applied for matching acquired fragments with suspects. Additionally, to increase the confidence of correctly detected fragments, a retention time tolerance of 0.5 min was employed for comparison of MS^1 and fragment profiles.

To convert retention times in retention indices based on the cocamide series a linear regression model was built between the retention time of 18 detected pesticides in all six chromatograms and their predicted retention indices (Section 4.5 Figure S3). The model

was applied to the rest of the detected pesticides (Section 4.5 Equation S2). Finally, the developed models were applied to the detected suspects to assign toxicity categories. Since 96h LC50 mortality of fathead minnows were unknown for detected pesticides, the predicted toxicity categories based on the structure presented via molecular fingerprints were used as benchmark toxicity categories.

4.2.3 Modeling

Assigning toxicity categories

A previous study demonstrated the advantage of using clustering algorithms as an alternative to expert-based toxicity categorization, which typically relies on fixed LC50 or EC50 thresholds [45]. These expert-defined thresholds often fail to account for the uncertainty in LC50 measurements or the structural similarities between chemicals. The k-means clustering algorithm, an unsupervised iterative method, clusters data based on the distances between measurements and a set of user-defined centroids. This approach takes advantage of structural similarities between chemicals when forming clusters.

In this study, we applied k-means clustering to categorize chemicals from a fish toxicity dataset, using scaled LC50 values [LOG(mg/L)], monoisotopic mass, and six elemental mass defects (EMDs): CO, CCl, CN, CS, CF, and CH as input features. [50] Since LC50 values primarily determine toxicity while structural information captures relationships between chemical structures, we assigned a weight of 1/7 to each of the monoisotopic mass and the six EMDs. This ensured that LC50 values and structural information contributed equally.

The algorithm converged after 12 iterations with an inertia of 323, forming three distinct clusters (Section 4.5 Figure S1). These clusters were then labeled as low, moderate, and high toxicity, based on their positions along the y-axis (LC50 values). The final clustering allowed clear differentiation between low toxicity (n = 238), high toxicity (n = 202), and moderate toxicity (n = 467) chemicals, providing a robust categorization informed by both toxicity data and structural characteristics.

Classification models

For modeling, a Random Forest Classifier (RFC) algorithm was implemented in Julia with scikit-learn. [51] Random Forest is a supervised algorithm that constructs several unique decision trees from bootstrap data. After model development, the major voting of predictions resulting from the individual trees is used to produce the final RFC model prediction. [52]

FINGERPRINT-BASED MODEL: To train the fingerprint-based (FPs-based) model, the fish toxicity dataset (n = 1047) was split into a training set (n = 854), test set (n = 95), and global test set (n = 98) via stratified sampling (Section 4.5 Figure S1). The input for the classification model was the optimizied fingerprint [53] and the output was toxicity categories. [45] The molecular fingerprints were calculated based on six different non-hashed molecular fingerprints, namely, Atom Pair 2D Count (AP2DC), [54] Electrotopological state (E-state), [55] Klekotha-Roth Count (KRC), [56] Molecular Access Systems (MACCS), [57] PubChem, [58] and Substructure Keys Count (SSC) FPs. [59] 340 bits out of combined 7073 were selected based on relevance to acute fish toxicity prediction. [53] The model was optimized on a training set, validated with a test set and finally tested with a global test set. Hyperparameter optimisation for the toxicity random forest classification models consisted of a number of estimators varying between 100 and 400 with a step of 100, minimum samples leaf between 2 and 10 with a step of 2, a maximum number of features of sqrt, log2, a maximum depth between 10 and 20 with a step of 2. Cross-validation of the hyperparameter optimisation was 3-fold. This FPs-based model was then utilized to predict the toxicity categories for CompTox and CNLs datasets.

CNLS-BASED MODEL: To train the CNLs-based model the CNLs dataset was randomly split into a training set (n = 18835), test set (n = 2093), and global test set (n = 98). The input for the classification model was the CNL values converted to a bit vector, monoisotopic mass, and predicted retention indices. The output of the CNL-based model was predicted probability of assignment to toxicity categories. The probability was further converted into predicted category based on the defined threshold. The threshold was calculated based on the difference in probabilities between all three categories using test set only. Hyperparameter optimisation was performed in the same manner as described in the Section 4.2.3. Fur-

thermore, to account for a class imbalance the weighted version of RFC was employed.

KDE based model

The KDE algorithm [60] was used for toxicity categories mapping on chromatographic dimensions using monoisotopic mass, predicted retention indices, and toxicity categories. Given fitted KDE, we calculated the probability for each chemical to belong to one of the predetermined toxicity categories using the formula:

$$\hat{f}(\mathbf{x}; \mathbf{H}) = \frac{1}{n|\mathbf{H}|^{1/2}} \sum_{i=1}^{n} K\left(\mathbf{H}^{-1/2}(\mathbf{x} - \mathbf{X}_i)\right) \tag{4.1}$$

where $\hat{f}(\mathbf{x})$ represents the estimated density at point x, n is the number of data points, d is the number of dimensions, H is the $d \times d$ bandwidth matrix, K is the multivariate kernel function, $|\mathbf{H}|$ is the determinant of the bandwidth matrix H, and $\mathbf{H}^{-1/2}$ is the matrix square root of the inverse of H. Because the classification data is unbalanced, the number of samples in each class was used as weights. This effectively cancels out the $1/n$ in the formula and allows for comparison between the kernel densities.

Bandwidth selection was carried out using Silverman's rule of thumb:

$$sigma_j = \left(\frac{4\hat{\sigma}_j^5}{3n_k}\right)^{\frac{1}{5}} \tag{4.2}$$

where σ_j denotes the standard deviation of the j-th feature in class C_k. This method ensures a smooth estimation of the density function, aiding in the effective prediction of toxicity classifications.

To evaluate the performance of the KDE based model, densities for all three toxicity classes were calculated for the fish toxicity dataset as only available experimentally measured values. The toxicity category was assigned based on the highest calculated probability (density).

Applicability domain calculations

The Applicability Domain (AD) defines the parameter space within which a predictive model is expected to make reliable predictions. In this study, we utilized the leverage approach to define and evaluate the AD of our predictive model [61]. The AD was assessed by calculating the leverage for each sample using the equation:

$$h_i = x_i^\top (X^\top X)^{-1} x_i \tag{4.3}$$

where x_i is the vector of predictor variables for the i-th sample, and X is the matrix of predictor variables for all training samples. The leverage values indicate the influence of each sample on the model's predictions.

The warning leverage threshold h^* is determined to identify samples that have high leverage and may lead to unreliable predictions. The threshold is typically set to three times the average leverage in the training set:

$$h^* = 3 \times h_{\text{avg}} \tag{4.4}$$

Thus, predictions for samples with leverage values greater than these thresholds are considered less reliable and are flagged as outside the AD. To ensure these do not affect the kernel density estimation, the weights for each sample were set at $1/h$.

Performance assessment

For the evaluation of the performance of the multiclass classification model, several metrics were computed, including precision, recall, F1-score, and accuracy. The classification metrics provide an overview of the model's performance for each class. The calculations for these metrics are as follows:

Precision: The ratio of correctly predicted positive observations to the total predicted positives.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4.5}$$

Recall (Sensitivity): The ratio of correctly predicted positive observations to all observa-

tions in the actual class.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4.6}$$

F1-score: The weighted average of Precision and Recall.

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.7}$$

Accuracy: The ratio of correctly predicted observations to the total observations

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4.8}$$

To further validate the classification models, the confusion matrix was computed, which provided a detailed breakdown of the model's predictions compared to the actual class labels. Because of the class imbalance, the confusion matrix was normalized to represent fractions, making it easier to interpret the proportion of correct and incorrect predictions for each class.

4.2.4 Calculations

All calculations were performed on a personal computer (PC) with an Intel Core i7-1260P central processing unit and 32 GB of RAM operating Windows 10 Education version 22H2. All of the data processing and statistical analyses were performed using Julia language version 1.6.7. The computational algorithms were computed using the ScikitLearn.jl version 0.7 library. The fish toxicity data set, the filtered CompTox data set, and the pesticide mixture dataset are publicaly available at `https://doi.org/10.5281/zenodo.14188167`. Additionally, a script to perform the calculations can be accessed at at `https://bitbucket.org/viktoriiaturkina/toxmap_prioritization.jl/src/main`.

4.3 Results and Discussion

In this study we developed two classification algorithms, machine learning and probabilistic, to predict fish toxicity categories based only on the chromatographic and fragmentation without identified structural information. Additionally, to assign the toxicity cate-

gories we trained the classification model based on the molecular FPs. For the probabilistic algorithm, the CompTox database was mapped on RPLC-HRMS space and KDE was fitted to each toxicity category assigned based on the known chemical structures (FPs). For the machine learning algorithm, the RFC model was trained on precursor ion mass (proton adduct), retention indices, and fragmentation information in the form of CNLs bit vector.

4.3.1 FPs-based model

The fingerprint-based model developed in this study was an RFC model trained on 900 structures described with optimized molecular fingerprint [53] and their experimental LC50 [LOG(mg/L)] values. [43–45] Optimized hyperparameters of the model were determined as 300 estimators, a minimum of 10 samples split, a maximum depth of 20, and sqrt as the number of features for split. To assess the performance of the model confusion matrix, accuracy, precision, recall, and F1-score for each toxicity category were calculated (Table 4.1).

The accuracy of the optimized model was 0.86 for the training set, 0.75 for the test set and 0.70 as a mean accuracy of 3-fold cross-validation, which was comparable to previously reported QSAR models for toxicity predictions. [62, 63] Performance assessment of the prediction results for chemicals from the test set showed that the model consistently performed well for low and moderate toxicity categories with an F1-score 0.80 and 0.73 correspondingly. However, the F1-score for the high toxicity category was lower compared to the other two categories, 0.54. Based on the confusion matrix and leverage analysis (Section 4.5 Figure S4), the lower recall for the high-toxicity class in the test set was primarily due to the misclassification of compounds with relatively low leverage values into the moderate-toxicity class. Compounds with lower leverage generally have lower monoisotopic mass and show greater similarity to compounds in the moderate and low-toxicity categories, that influences the accuracy of high-toxicity category predictions. Moreover, the chemicals assigned to the high-toxicity category exhibited greater variability in the structural representation with the smallest number of chemicals included, 145 for the training set and 13 for the test set.

To ensure a broader applicability domain for predictive purposes, the final fingerprint-based classification model was trained using a combination of the training and test sets. A separate global test set, containing measured toxicity values and available LC-HRMS data,

was reserved for the final model assessment. The validated model was then used to predict toxicity categories for the CompTox and CNLs datasets, expanding the dataset and supporting the development of a probabilistic approach and the training of a CNLs-based model.

Table 4.1: The performance assessment of the fingerprint-based classification model.

	Class	Precision	Recall	F1-score	Support	Accuracy
Training set	Low	0.90	0.83	0.87	263	0.86
	Moderate	0.83	0.94	0.88	446	
	High	0.92	0.66	0.77	145	
Test set	Low	0.72	0.75	0.73	24	0.75
	Moderate	0.81	0.79	0.80	58	
	High	0.54	0.54	0.54	13	
Combined dataset	Low	0.88	0.83	0.85	287	0.85
	Moderate	0.83	0.93	0.87	504	
	High	0.88	0.65	0.75	158	

4.3.2 KDE-based model

To create a KDE algorithm the toxicity categories and retention indices were predicted for each chemical from the CompTox dataset based on their structures. This information was then used to map toxicity regions within the RPLC-MS domain, linking toxicity categories to features in the chromatogram.

Out of the 764,590 chemicals of the filtered Comptox database, 83,969 chemicals (10.9%) were classified as low toxicity, 458,302 chemicals (59.9%) as moderate toxicity, and 222,319 chemicals (29.0%) as high toxicity class. However, only 20.8% of the CompTox dataset is well represented by the training set, meaning the calculated leverage for these chemicals is less than the leverage threshold, 3 times of the mean value of leverage in the training set, 1.02. To decrease the uncertainty of KDE we used calculated leverage values as weights to account for the less certain prediction of the fingerprint-based model. Therefore, chemicals that were well represented in the training set, gave more weights for the density estimation and more accurately represented the results. The weights used were $1/h$ with h being the leverage of a compound (Figure 4.2a).

The CompTox dataset contains a wide variety of structures, with the range of monoiso-

topic mass between 12.0 and 5,628.0 Da, and the predicted retention indices from 178.6 to 1,530.5 RI value. (Section 4.5 Figure S2) A partion of this space cannot be analysed with RPLC, either due to insufficient retention (e.g., highly hydrophilic chemicals, often with low masses, i.e., $<$ 70 Da) or excessive retention (e.g., highly hydrophobic chemicals with high mass and high retention indices). Several recent studies investigated algorithms for assigning chemicals to be analyzable with RPLC or other types of chromatography. [46, 64–66] We applied the machine learning algorithm developed by van Herwerden et al. [46] to the CompTox dataset to filter out chemicals that will not be analysed and therefore detected using RPLC. Of the 671,395 chemicals predicted to fall within the RPLC space, 78,087 (11.6%) were classified as low toxicity, 431,089 (64.2%) as moderate toxicity, and 162,219 (24.2%) as high toxicity category. Thus, mainly compounds predicted as highly toxic fell outside of RPLC space (Figure 4.2b).

Overall accuracy of the algorithm was 0.69. Similar results were observed for overall precision, recall and F1-score. The metrics of predictions for moderate and high toxicity categories showed comparable performance with precision 0.75-0.76, recall 0.6-0.64, and F1-score 0.67-0.70. Results for low-toxicity group with 0.58 for precision and 0.84 for recall, differed from the results for the rest of the categories. Overall, the algorithm showed consistently good performance among all categories.

Only 1.3% of chemicals from low-toxicity group were classified as highly toxic chemicals and 9.4% of high-toxicity chemicals were classified as low-toxicity chemicals. This indicated that low toxicity compounds generally do not occupy the region of high toxicity class compounds. The larger error occurs in misclassification between moderate and low toxicity groups (28%) and between high and moderate toxicity groups (30%) (Section 4.5 Figure S5). The developed algorithm based on KDE was as accurate as the FPs-based model, even though no actual structural information was provided to it.

4.3.3 CNLs-based model

For CNLs-based model construction, we predicted toxicity categories and retention indices for the CNLs dataset. Out of 20,928 chemicals in the dataset 3,152 chemicals were predicted as low-toxicity chemicals, 12,662 as moderate, and 5,114 as high toxicity. The number of chemicals assigned to different groups shows an imbalance in the dataset; therefore to train the model, the weighted version of RFC was employed.

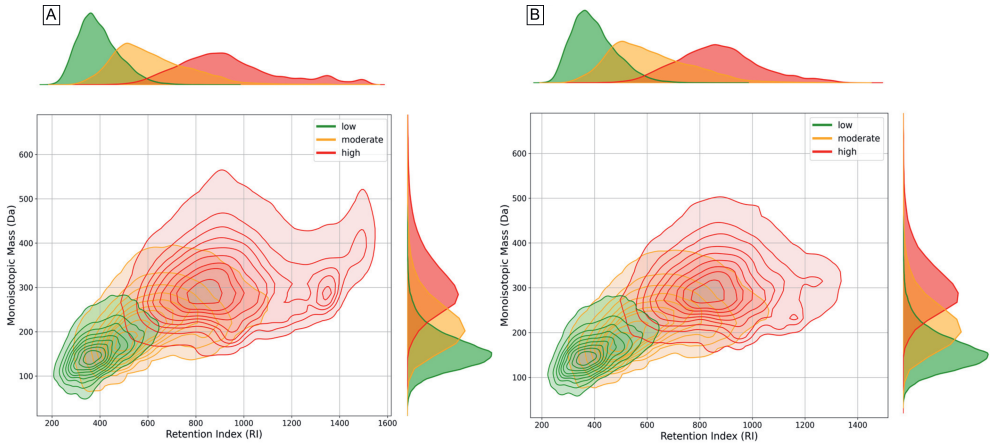


Figure 4.2: KDE joint plots of filtered on RPLC space Comptox dataset (a), and applied to KDE weights based on calculated leverage (b)

The optimized the CNLs-based model showed an overall accuracy of 0.76 for training set and 0.72 for test set. Therefore, the final model showed a high level of accuracy, comparable to fingerprint-based classification model. Moreover, precision, recall, and F1-score for training and test set show consistent result for all three toxicity groups (Section 4.5 Table S2), indicating excellent performance of the model regardless of the class imbalance.

This model consisted of 600 trees, a minimum of two samples in each leaf, a maximum depth of 20, and sqrt as a maximum number of features. A difference in probability threshold of 0.08 was applied to the classification results: if the difference in predicted probabilities across all three toxicity categories was less than 0.08 (or 8%), the chemical was assigned to the moderate toxicity category. This is because, in the KDE model, the moderate toxicity category is the most uncertain and tends to overlap with both the low and high toxicity categories.

The model explained 85% of the variance using 592 variables, including CNLs with masses ranging from 2.01 Da to 354.19 Da. Based on impurity decrease across trees, the largest contributions to the model came from retention index (RI) and accurate mass, cumulatively accounting for 36%. These were followed by contributions from CNLs with masses such as 17.03 Da (1.7%), likely corresponding to NH_3, 45.01 Da (CH_3NO), 27.01 Da (HCN), and 63.03 Da (HNO_2 or CH_2SO), each contributing around 1%. While these structures are plausible, alternative small molecules corresponding to these masses cannot be disregarded. Higher mass CNLs had an additional contribution, including 294.11

Da, potentially representing multiple aromatic rings and/or benzyloxy substructures, 216.13 Da, possibly linked to aromatic heterocycles like indole or quinoline, and 174.11 Da, which may correspond to a thiocarboxylic ester. This indicates that the model primarily relies on MS^1 data, supplemented by structural information that linked to the potentially toxicity-relevant substructures.

4.3.4 Comparison of the KDE and CNLs-based models

Based on the confusion matrices for training and test set the comparable performance to KDE algorithm was observed. The main misclassifications occur for a moderate toxicity group, while low toxicity barely ever classified as high toxicity group (1%) and vice versa (0%) (Section 4.5 Figure S6). However, the test and training set for the KDE model and CNLs based model were different therefore one-to-one comparison cannot be performed using these datasets. Thus, to further evaluate and compare prediction power between all three generated models we used a global test set. These 98 entries had both the CNL bit vector and fish toxicity experimentally determined values and were completely unknown to any of the models.

When comparing KDE and CNLs-based models with the FPs-model, the confusion matrices indicate that the FPs-based model produced comparable results in predicting low and high toxicity categories. This is a significant outcome, given that the KDE and CNLs-based models do not account for the structural representation of the compounds and are directly linked to chromatographic and mass spectrometry data. However, FPs-based model outperformed the other models in predicting the moderate toxicity group, achieving an accuracy of 0.88 compared to 0.47 for the KDE model and 0.43 for the CNL model (Section 4.5 Figure S6). This can be explained by the substantial overlap in the RPLC-MS space assigned to the moderate toxicity category, which has a mass range of 110.11 to 336.30 and retention indices between 285.28 and 1139.23, with the regions for low and high toxicity categories.

To further demonstrate the applicability of the developed approach, we applied it to an experimentally acquired dataset of a standard pesticide mixture at varying concentrations in a tea extract. Between 141 and 155 pesticides were detected out of 253 in the standard mixture across all three levels of the tea matrix (no tea, 100-fold dilution, and 10-fold dilution). The minimum number of 141 pesticides was detected in the 100-fold diluted tea,

while 155 were detected in the 10-fold diluted tea. No clear trend was observed between the number of detected pesticides and the matrix dilution. Additionally, the number of detected and matched fragments was compared across the three matrices. The median number of fragments per pesticide was 12 to 13, with the 25th and 75th percentiles ranging from 7 to 21-23. Overall, no significant influence of the matrix level was observed on the number of detected pesticides or the number of corresponding fragments. Since matrix effects and other parameters can influence not only the number but also the quality of detected fragments, the CNLs-based model was applied independently to each chromatogram (Section 4.5 Table S3 and Figure S7).

To keep prediction results comparable between the models based on KDE and CNLs, we retained only the features corresponding to the standard pesticide mixture detected in all samples. As part of the ULSA workflow, Suspect Screening was applied to identify corresponding features at both the chromatographic and fragmentation levels. In total 84 pesticides were detected in all three different matrices, 18 of which were used to establish a correlation between retention time and retention indices, yielding an R^2 = 0.91. The remaining 66 pesticides were used for further testing. Based on molecular structure, the FPs-based model classified the 66 pesticides as follows: 3 assigned to the low toxicity category, 48 to moderate toxicity, and 15 to high toxicity.

When applied to the detected pesticides, the CNLs-based model achieved an average accuracy of 0.76-0.80, while the KDE model had a lower accuracy of 0.61. Additionally, the accuracy of the CNLs model was not significantly influenced by different matrix levels, as variations in accuracy across repeated measurements were within 3%, similar to the variation between different matrices. This suggests that the CNLs model maintains stability and reliability across diverse experimental conditions, making it particularly useful for complex environmental matrices. We examined chemicals that were misclassified by either the KDE-based or CNLs-based models, meaning they were assigned to a different toxicity category compared to the benchmark (FPs-based) model. For instance, Metsulfuron-methyl (INCHIKEY: RSMUVYRMZCOLBH-UHFFFAOYSA-N) was predicted to belong to the high toxicity category by the KDE-based model, while both the CNLs-based and FPs-based models classified it as moderate toxicity. While there are no measured LC50 values available for fathead minnows, experimental LC50 values for Bluegill Sunfish and Rainbow Trout reported in the Pesticide Ecotoxicity Database from the EPA indicate LC50 values greater than 2 [LOG(mg/L)] for a 96-hour exposure. Based on classification

criteria, chemicals with LC50 values above 2 [LOG(mg/L)] and a monoisotopic mass of 381.07 Da typically fall into the moderate or low toxicity categories for fathead minnows. The CNLs and FPs models, therefore, offered a more accurate classification in this case, reflecting the expected behavior of Metsulfuron-methyl in aquatic species. On the other hand, Thiabendazole (INCHIKEY: WJCNZQLZVWNLKY-UHFFFAOYSA-N) was assigned to the moderate toxicity category by the CNLs-based model, while the KDE- and FPs-based models classified it as low toxicity. Again, no data were available for fathead minnows, but the LC50 values for Bluegill Sunfish and Rainbow Trout, as recorded by the EPA, range between -0.25 and 1.8 [LOG(mg/L)]. With an accurate mass of 201.04 Da, this chemical is typically assigned to either moderate or low toxicity categories. The discrepancy in classification highlights the challenges of classifying compounds that fall near the boundary between toxicity classes. In this case, the CNLs model's assignment of moderate toxicity may be more reflective of the compound's potential toxicity, given the reported LC50 values and structural characteristics.

Both models performed well on experimental and global test datasets. Overall, the CNLs-based model outperformed the KDE-based model in the pesticide dataset, demonstrating superior predictive power and more consistent classification.

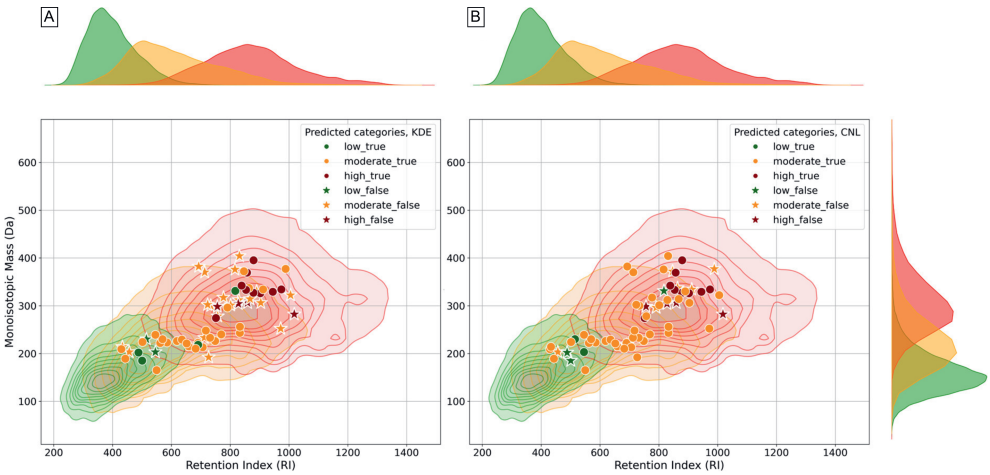


Figure 4.3: KDE map of RPLC space with the predicted toxicity categories for the detected pesticides based on KDE model (a), and CNLs-based model (b).

Several recent studies have focused on linking high-resolution accurate mass tandem mass spectrometry (MS^2) data to LC50 and EC50 values. These studies primarily aim to predict LC50 or EC50 values, which are then converted into toxicity categories. [30–33]

Typically, this process involves converting the spectra into estimated molecular formulas and predicted probabilistic molecular fingerprints using tools like SIRIUS. [34] The generated probabilistic fingerprints are then used for LC50 or EC50 prediction. As a result, the performance of these models depends on the accuracy of the molecular formula and fingerprint predictions. In contrast, our approach bypasses the need for molecular structure prediction by directly linking mass spectral signals to toxicity categories. Instead of using a regression model to predict continuous LC50 or EC50 values, we apply a classification algorithm that assigns compounds directly into toxicity categories. By relying solely on chromatographic and fragmentation data, our CNL- and KDE-based algorithms present a unique approach that differs from the structure-dependent models used in previous studies, making direct comparisons challenging.

4.4 Implications

In this study, we showed a novel feature prioritization approach based on predicting toxicity categories, by employing two algorithms: a probabilistic KDE-based model and a machine learning-based CNLs model. Due to the complex nature of environmental samples and the highly convoluted data, fragmentation information is not consistently available for all features. To mitigate the influence of the MS^2 signal availability we developed KDE-based algorithm, which relies solely on MS^1 and retention of the features. On the other hand the combination of retention and MS^1 and fragmentation patterns or CNLs enhances the prediction accuracy. Our approach connects chromatographic and fragmentation data directly to toxicity categories, bypassing the need for compound identification, a common requirement in the currently available prioritization strategies. Moreover, our models demonstrated accuracy comparable to those that depend on structural information (FPs-based model), enabling effective toxicity predictions even for compounds with unknown structures. This makes our models not only valuable for potential prioritization of unknown features in current NTA studies but also suitable for enhancing retrospective analyses of historical datasets.

One of the current limitations of the developed approach is the limited number of measured toxicity values, retention indices, and high-resolution mass spectra, which reduces the model's applicability domain. Furthermore, the model is currently developed based on data from a single trophic level and acute toxicity. More specific mode of action models

can give more insight into hazard scores and risk assessment. Another challenge arises from the fact that the model was trained on clean spectra obtained from publicly available spectral databases. In real LC-HRMS measurements, co-eluted compounds, instrumental artifacts, and background noise can introduce false positive fragments, potentially reducing prediction accuracy. As a result, robust and reliable data preprocessing is critical especially for DIA experiments to ensure the accuracy of the model.

4.5 Supporting Information

The Supporting Information is available at `https://doi.org/10.1021/acs.est.4c13026` and includes additional details on acquisition method for pesticide mixture dataset (Section S1), distribution on assigned toxicity categories and training, test, and global test set of fish toxicity dataset (Figure S1), plotted CompTox dataset with predicted toxicity categories (Figure S2), equation for inverse class frequency calculation (Equation S1), regression equation to convert *tr* into *RI* (Equation S2), additonal details on models performance (Figure S4-S6, Table S2), and number of fragments matched with suspect screening (Table S3, Figure S7).

Notes

The script to perform the calculations is available at `https://doi.org/10.5281/zenodo.14188167`. The datasets used for the development of the algorithms are available at `https://bitbucket.org/viktoriiaturkina/toxmap_prioritization.jl/src/main`

Acknowledgement

The authors are thankful to the members of Environmental Modeling & Computational Mass Spectrometry (www.emcms.info). S.S., A.P., and V.T. are thankful to the UvA Data Science Center and ChemistryNL TKI for the financial support (projects Edified and SCOPE). The authors thank Denice van Herwerden for her help in setting up the calculations of el-

emental mass defects. J.W.O. is the recipient of a National Health and Medical Research Council (NHMRC) Investigator Grant (EL12009209) funded by the Australian Government. K.V.T, and S.S. also acknowledge funding support from the Australian National Health and Medical Research Council (NHMRC; APP1185347). The Queensland Alliance for Environmental Health Sciences (QAEHS), The University of Queensland (UQ), gratefully acknowledges the financial support of Queensland Health. This research was funded in whole or in part by NHMRC Emerging Leadership Fellowship EL1 2009209. For the purposes of open access, the author has applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising from this submission.

4.6 References

[1] Landrigan P. J, Fuller R, Acosta N. J. R, Adeyi O, Arnold R, Basu N. N, et al. The Lancet Commission on pollution and health. Lancet. 2018 2;391(10119):462-512. doi:10.1016/s0140-6736(17)32345-0.

[2] Arp H. P. H, Aurich D, Schymanski E. L, Sims K, Hale S. E. Avoiding the Next Silent Spring: Our Chemical Past, Present, and Future. Environ Sci Technol. 2023 4;57(16):6355-9. doi:10.1021/acs.est.3c01735.

[3] Wang Z, Walker G. W, Muir D. C. G, Nagatani-Yoshida K. Toward a Global Understanding of Chemical Pollution: A First Comprehensive Analysis of National and Regional Chemical Inventories. Environ Sci Technol. 2020 3;54(5):2575-84. doi:10.1021/acs.est.9b06379.

[4] Rudén C, Hansson S. O. Registration, Evaluation, and Authorization of Chemicals (REACH) Is but the First Step–How Far Will It Take Us? Six Further Steps to Improve the European Chemicals Legislation. Environ Health Perspect. 2010 1;118(1):6. doi:10.1289/ehp.0901157.

[5] Hernández F, Bakker J, Bijlsma L, de Boer J, Botero-Coy A. M, Bruinen de Bruin Y, et al. The role of analytical chemistry in exposure science: Focus on the aquatic environment. Chemosphere. 2019 5;222:564-83. doi:10.1016/j.chemoshpere.2019.01.118.

[6] Ciccarelli D, Samanipour S, Rapp-Wright H, Bieber S, Letzel T, O'Brien J. W, et al. Bridging knowledge gaps in human chemical exposure via drinking water with non-target screening. Crit Rev Environ Sci Technol. 2024;55(3):190-214. doi:10.1080/10643389.2024.2396690.

[7] Samanipour S, Barron L. P, van Herwerden D, Praetorius A, Thomas K. V, O'Brien J. W. Exploring the Chemical Space of the Exposome: How Far Have We Gone? JACS Au. 2024 7;4(7):2412-25. doi:10.1021/jacsau.4c00220.

[8] Gosetti F, Mazzucco E, Gennaro M. C, Marengo E. Contaminants in water: non-target UHPLC/MS analysis. Environ Chem Lett. 2016 3;14:51-65. doi:10.1007/s10311-015-0527-1.

[9] Richardson S. D, Manasfi T. Water Analysis: Emerging Contaminants and Current Issues. Anal Chem. 2024 5;96(20):8184-219. doi:10.1021/acs.analchem.4c01423.

[10] Krauss M, Singer H, Hollender J. LC-high resolution MS in environmental analysis: From target screening to the identification of unknowns. Anal Bioanal Chem. 2010 6;397(3):943-51. doi:10.1007/s00216-010-3608-9.

[11] Sobus J. R, Wambaugh J. F, Isaacs K. K, Williams A. J, McEachran A. D, Richard A. M, et al. Integrating tools for non-targeted analysis research and chemical safety evaluations at the US EPA. J Expo Sci Environ Epidemiol. 2017 12;28(5):411-26. doi:10.1038/s41370-017-0012-y.

[12] Overdahl K. E, Sutton R, Sun J, Destefano N. J, Getzinger G. J, Ferguson P. L. Assessment of emerging polar organic pollutants linked to contaminant pathways within an urban estuary using non-targeted analysis. Environ Sci Process Impacts. 2021 4;23(3):429-45. doi:10.1039/d0em00463d.

[13] Creusot N, Huba K, Borel C, Ferrari B. J. D, Chèvre N, Hollender J. Identification of polar organic chemicals in the aquatic foodweb: Combining high-resolution mass spectrometry and trend analysis. Environ Int. 2024 1;183:108403. doi:10.1016/j.envint.2023.108403.

[14] Kutarna S, Tang S, Hu X, Peng H. Enhanced Nontarget Screening Algorithm Reveals Highly Abundant Chlorinated Azo Dye Compounds in House Dust. Environ Sci Technol. 2021 4;55(8):4729-39. doi:10.1021/acs.est.0c06382.

[15] Szabo D, Fischer S, Mathew A. P, Kruve A. Prioritization, Identification, and Quantification of Emerging Contaminants in Recycled Textiles Using Non-Targeted and Suspect Screening Workflows by LC-ESI-HRMS. Anal Chem. 2024 9;96(35):14150-9. doi:10.1021/acs.analchem.4c02041.

[16] Braun G, Herberth G, Krauss M, König M, Wojtysiak N, Zenclussen A. C, et al. Neurotoxic mixture effects of chemicals extracted from blood of pregnant women. Science. 2024 10;386(6719):301-9. doi:10.1126/science.adq0336.

[17] Petrie B, Barden R, Kasprzyk-Hordern B. A review on emerging contaminants in wastewaters and the environment: Current knowledge, understudied areas and recommendations for future monitoring. Water Res. 2015 4;72:3-27. doi:10.1016/j.watres.2014.08.053.

[18] Hohrenk L. L, Itzel F, Baetz N, Tuerk J, Vosough M, Schmidt T. C. Comparison of Software Tools for Liquid Chromatography-High-Resolution Mass Spectrometry Data Processing in Nontarget Screening of Environmental Samples. Anal Chem. 2020 1;92(2):1898-907. doi:10.1021/acs.analchem.9b04095.

[19] Hollender J, Schymanski E. L, Singer H. P, Ferguson P. L. Nontarget Screening with High Resolution Mass Spectrometry in the Environment: Ready to Go? Environ Sci Technol. 2017 10;51:11505-12. doi:10.1021/acs.est.7b02184.

[20] Lennon S, Chaker J, Price E. J, Hollender J, Huber C, Schulze T, et al. Harmonized quality assurance/quality control provisions to assess completeness and robustness of MS1 data preprocessing for LC-HRMS-based suspect screening and non-targeted analysis. Trends Anal Chem. 2024 5;174:117674. doi:10.1016/j.trac.2024.117674.

[21] Hulleman T, Turkina V, O'Brien J. W, Chojnacka A, Thomas K. V, Samanipour S. Critical Assessment of the Chemical Space Covered by LC-HRMS Non-Targeted Analysis. Environ Sci Technol. 2023 9;57(38):14101-12. doi:10.1021/acs.est.3c03606.

[22] González-Gaya B, Lopez-Herguedas N, Bilbao D, Mijangos L, Iker A. M, Etxebarria N, et al. Suspect and non-target screening: the last frontier in environmental analysis. Anal Methods. 2021 4;13(16):1876-904. doi:10.1039/d1ay00111f.

[23] Schymanski E. L, Jeon J, Gulde R, Fenner K, Ruff M, Singer H. P, et al. Identifying small molecules via high resolution mass spectrometry: Communicating confidence. Environ Sci Technol. 2014 2;48:2097-8. doi:10.1021/es5002105.

[24] Kind T, Fiehn O. Advances in structure elucidation of small molecules using mass spectrometry. Bioanal Rev. 2010 8;2(1):23-60. doi:10.1007/S12566-010-0015-9.

[25] Szabo D, Falconer T. M, Fisher C. M, Heise T, Phillips A. L, Vas G, et al. Online and Offline Prioritization of Chemicals of Interest in Suspect Screening and Non-targeted Screening with High-Resolution Mass Spectrometry. Anal Chem. 2024 3;96(9):3707-16. doi:10.1021/acs.analchem.3c05705.

[26] Rilievo G, Boscolo S, Pettenuzzo S, Matozzo V, Fabrello J, Roverso M, et al. From a validated targeted method to a retrospective UHPLC-HRMS non-targeted analysis unveiling COVID-19-related contaminants in clams. Have we bias in marine model organisms for ecotoxicological studies? Chemosphere. 2024 9;364:142994. doi:10.1016/j.chemosphere.2024.142994.

[27] Silva M, Capps S, London J. Community-Engaged Research and the Use of Open Access ToxVal/ToxRef In Vivo Databases and New Approach Methodologies (NAM) to Address Human Health Risks From Environmental Contaminants. Birth Defects Res. 2024 9;116(9):e2395. doi:10.1002/bdr2.2395.

[28] Vermeulen R, Schymanski E. L, Barabási A. L, Miller G. W. The exposome and health: where chemistry meets biology. Science. 2020 1;367(6476):392. doi:10.1126/science.aay3164.

[29] Meekel N, Vughs D, Béen F, Brunner A. M. Online Prioritization of Toxic Compounds in Water Samples through Intelligent HRMS Data Acquisition. Anal Chem. 2021 3;93(12):5071-80. doi:10.1021/acs.analchem.0c04473.

[30] Peets P, Wang W. C, Macleod M, Breitholtz M, Martin J. W, Kruve A. MS2Tox Machine Learning Tool for Predicting the Ecotoxicity of Unidentified Chemicals in Water by Nontarget LC-HRMS. Environ Sci Technol. 2022 11;56(22):15508-17. doi:10.1021/acs.est.2c02536.

[31] Peets P, Rian M. B, Martin J. W, Kruve A. Evaluation of Nontargeted Mass Spectral Data Acquisition Strategies for Water Analysis and Toxicity-Based Feature Prioritization by MS2Tox. Environ Sci Technol. 2024;58:17406-18. doi:10.1021/acs.est.4c02833.

[32] Rahu I, Kull M, Kruve A. Predicting the Activity of Unidentified Chemicals in Complementary Bioassays from the HRMS Data to Pinpoint Potential Endocrine Disruptors. J Chem Inf Model. 2024 4;64(8):3093-104. doi:10.1021/acs.jcim.3c02050.

[33] Arturi K, Hollender J. Machine Learning-Based Hazard-Driven Prioritization of Features in Nontarget Screening of Environmental High-Resolution Mass Spectrometry Data. Environ Sci Technol. 2023 11;57(46):18067-79. doi:10.1021/acs.est.3c00304.

[34] Dührkop K, Fleischauer M, Ludwig M, Aksenov A. A, Melnik A. V, Meusel M, et al. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. Nat Methods. 2019 3;16(4):299-302. doi:10.1038/s41592-019-0344-8.

[35] Zushi Y, Hanari N, Nabi D, Lin B. L. Mixture Touch: A Web Platform for the Evaluation of Complex Chemical Mixtures. ACS Omega. 2020 4;5(14):8121-6. doi:10.1021/acsomega.0c00340.

[36] Zhang R, Guo H, Hua Y, Cui X, Shi Y, Li X. Modeling and insights into the structural basis of chemical acute aquatic toxicity. Ecotoxicol Environ Saf. 2022 9;242:113940. doi:10.1016/j.ecoenv.2022.113940.

[37] Yang H, Lou C, Li W, Liu G, Tang Y. Computational Approaches to Identify Structural Alerts and Their Applications in Environmental Toxicology and Drug Discovery. Chem Res Toxicol. 2020 6;33(6):1312-22. doi:10.1021/acs.chemrestox.0c00006.

[38] Williams A. J, Grulke C. M, Edwards J, McEachran A. D, Mansouri K, Baker N. C, et al. The CompTox Chemistry Dashboard: A community data resource for environmental chemistry. J Cheminform. 2017 11;9(1):1-27. doi:10.1186/s13321-017-0247-6.

[39] MassBank Europe Mass Spectral DataBase (2023, October, 15);. https://massbank.eu/MassBank/.

[40] MassBank of North America (2023, October, 15);. https://mona.fiehnlab.ucdavis.edu/downloads.

[41] Reference Spectral Libraries - GNPS Documentation (2023, October, 15);. https://ccms-ucsd.github.io/GNPSDocumentation/downloadlibraries/.

[42] NIST20: Updates to the NIST Tandem and Electron Ionization Spectral Libraries (2023, October, 15);. https://www.nist.gov/programs-projects/nist20-updates-nist-tandem-and-electron-ionization-spectral-libraries.

[43] Cassotti M, Ballabio D, Todeschini R, Consonni V. A similarity-based QSAR model for predicting acute toxicity towards the fathead minnow (Pimephales promelas). SAR QSAR Environ Res. 2015 3;26(3):217-43. doi:10.1080/1062936x.2015.1018938.

[44] Schür C, Gasser L, Perez-Cruz F, Schirmer K, Baity-Jesi M. A benchmark dataset for machine learning in ecotoxicology. Sci Data. 2023 10;10(1):1-20. doi:10.1038/s41597-023-02612-2.

[45] Samanipour S, O'Brien J. W, Reid M. J, Thomas K. V, Praetorius A. From Molecular Descriptors to Intrinsic Fish Toxicity of Chemicals: An Alternative Approach to Chemical Prioritization. Environ Sci Technol. 2023 11;57(46):17950-8. doi:10.1021/acs.est.2c07353.

[46] van Herwerden D, Nikolopoulos A, Barron L. P, O'Brien J. W, Pirok B. W. J, Thomas K. V, et al. Exploring the chemical subspace of RPLC: A data driven approach. Anal Chim Acta. 2024 8;1317:342869. doi:10.1016/j.aca.2024.342869.

[47] Aalizadeh R, Nikolopoulou V, Thomaidis N. S. Development of Liquid Chromatographic Retention Index Based on Cocamide Diethanolamine Homologous Series (C(n)-DEA). Anal Chem. 2022 11;94(46):15987-96. doi:10.1021/acs.analchem.2c02893.

[48] van Herwerden D, O'Brien J. W, Lege S, Pirok B. W. J, Thomas K. V, Samanipour S. Cumulative Neutral Loss Model for Fragment Deconvolution in Electrospray Ionization High-Resolution Mass Spectrometry Data. Anal Chem. 2023 8;95(33):12247-55. doi:10.1021/acs.analchem.3c00896.

[49] Samanipour S, Reid M. J, Bæk K, Thomas K. V. Combining a Deconvolution and a Universal Library Search Algorithm for the Nontarget Analysis of Data-Independent Acquisition Mode Liquid Chromatography-High-Resolution Mass Spectrometry Results. Environ Sci Technol. 2018 4;52(8):4694-701. doi:10.1021/acs.est.8b00259.

[50] van Herwerden D, O'Brien J. W, Choi P. M, Thomas K. V, Schoenmakers P. J, Samanipour S. Naive Bayes classification model for isotopologue detection in LC-HRMS data. Chemom Intell Lab Syst. 2022 4;223:104515. doi:10.1016/j.chemolab.2022.104515.

[51] Pedregosa F, Michel V, Grisel O, Blondel M, Prettenhofer P, Weiss R, et al.. Scikit-learn: Machine Learning in Python; 2011. `http://scikit-learn.sourceforge.net.`

[52] Svetnik V, Liaw A, Tong C, Christopher Culberson J, Sheridan R. P, Feuston B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. J Chem Inf Comput. 2003 11;43(6):1947-58. doi:10.1021/ci034160g.

[53] Turkina V, Messih M. R. W, Kant E, Gringhuis J, Petrignani A, Corthals G, et al. Molecular Fingerprints Optimization for Enhanced Predictive Modeling. ChemRxiv. 2024 2;preprint(doi:10.26434/chemrxiv-2024-zr2vr). doi:10.26434/chemrxiv-2024-zr2vr.

[54] Smith D. H, Carhart R. E, Venkataraghavan R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. J Chem Inf Comput. 1985 5;25(2):64-73. doi:10.1021/ci00046a002.

[55] Hall L. H, Kier L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. J Chem Inf Comput. 1995;35(6):1039-45. doi:10.1021/ci00028a014.

[56] Klekota J, Roth F. P. Chemical substructures that enrich for biological activity. Bioinformatics. 2008;24(21):2518-25. doi:10.1093/bioinformatics/btn479.

[57] Durant J. L, Leland B. A, Henry D. R, Nourse J. G. Reoptimization of MDL keys for use in drug discovery. J Chem Inf Comput. 2002 11;42(6):1273-80. doi:10.1021/ci010132r.

[58] PubChem Substructure Fingerprint V1.3 (2024, January, 23);. https://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.pdf.

[59] Yap C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. J Comput Chem. 2011 5;32(7):1466-74. doi:10.1002/jcc.21707.

[60] Węglarczyk S. Kernel density estimation and its application. ITM Web Conf. 2018;23:00037. doi:10.1051/itmconf/20182300037.

[61] Gramatica P. Principles of QSAR models validation: internal and external. QSAR Comb Sci. 2007 5;26(5):694-701. doi:10.1002/QSAR.200610151.

[62] Yu X. Global classification models for predicting acute toxicity of chemicals towards Daphnia magna. Environ Res. 2023 12;238:117239. doi:10.1016/j.envres.2023.117239.

[63] Sheffield T. Y, Judson R. S. Ensemble QSAR Modeling to Predict Multispecies Fish Toxicity Lethal Concentrations and Points of Departure. Environ Sci Technol. 2019 11;53(21):12793-802. doi:10.1021/acs.est.9b03957.

[64] Black G, Lowe C, Anumol T, Bade J, Favela K, Feng Y. L, et al. Exploring chemical space in non-targeted analysis: a proposed ChemSpace tool. Anal Bioanal Chem. 2022 1;415(1):35. doi:10.1007/S00216-022-04434-4.

[65] Lowe C. N, Isaacs K. K, McEachran A, Grulke C. M, Sobus J. R, Ulrich E. M, et al. Predicting compound amenability with liquid chromatography-mass spectrometry to improve non-targeted analysis. Anal Bioanal Chem. 2021 12;413(30):7495. doi:10.1007/s00216-021-03713-w.

[66] Alygizakis N, Konstantakos V, Bouziotopoulos G, Kormentzas E, Slobodnik J, Thomaidis N. S. A Multi-Label Classifier for Predicting the Most Appropriate Instrumental Method for the Analysis of Contaminants of Emerging Concern. Metabolites. 2022 3;12(3):199. doi:10.3390/metabo2030199/s1.

5

From Spectra to Local Networks: Evaluating MS^2 Similarity Metrics

Authors: Viktoriia Turkina, Quinten van Erp, Annemieke Petrignani, Garry Corthals, Jake W. O'Brien, Saer Samanipour

Published in: Analytical Chemistry, 2026, 98, 24, 17937-17947.

https://doi.org/10.1021/acs.analchem.6c01202

Abstract

MS^2 spectral similarity is fundamental to interpreting LC-HRMS-NTA data. Beyond the commonly used cosine similarity, a wide range of alternative metrics, including distance-, probability-, and machine-learning-based approaches, provide different perspectives on spectral matching. In this study, we extracted 8,290 subdatasets from publicly available LC-HRMS/MS libraries (MassBank, MoNa, GNPS, and NIST), each containing spectra sharing a precursor m/z within a 5 mDa tolerance. We evaluated 20 similarity metrics by constructing single-generation local molecular networks. Most metrics failed to produce pure networks, achieving complete resolution in only ≈ 8% of cases at the recommended 0.7 threshold and ≈ 20% even when individually optimized. The number of fragment ions showed little influence on network resolution. Instead, performance was driven primarily by the similarity metric and thresholding behavior, indicating that library matching alone rarely supports identification confidence above level 3. These findings highlight the need for cumulative strategies that integrate multiple similarity perspectives and orthogonal information, such as retention time or index.

5.1 Introduction

A vast number of environmentally relevant chemicals can be found in our surroundings and potentially affect both human and environmental health. [1–3] Most of these chemicals remain poorly characterized due to the complexity of environmental samples and the limitations of current analytical methods. [3–5] This leads to a number of unknowns or emerging contaminants that may pose serious, yet undetected, threats. [6,7]

One of the most advanced techniques for investigating both well-defined (known) and poorly defined (unknown) compounds is non-targeted analysis by liquid chromatography coupled with high-resolution mass spectrometry (LC-HRMS-NTA). This technique aims to detect as broad range of chemical compounds as possible. [8,9] Ideally, all detected compounds would be confidently identified and structurally characterized. However, achieving confident identification remains a major challenge due to the vast chemical diversity present in complex samples and the large number of possible candidate structures associated with each detected feature. [10–12] Consequently, the use of public spectral and chemical databases has become common practice in exposomics, as it supports standardized confidence reporting for NTA identification results. [10,12]

Nevertheless, confident identification is essential for interpretation and decision-making. [13] Misidentification among structurally similar compounds can lead to incorrect interpretation of results, as closely related structures may exhibit markedly different effects on human and environmental health. [14]

A common strategy for characterizing unknowns is to investigate their MS/MS (MS^2) spectra, with the assumption that MS^2 spectral similarity serves as a proxy for structural similarity. [15,16] However, the relationship between spectral similarity scores and true chemical identity is imperfect. [16] Structurally distinct compounds can produce highly similar fragmentation patterns, whereas structurally related compounds may fragment differently depending on experimental conditions. [17–19] Consequently, spectral similarity metrics must balance detecting true matches against avoiding false ones, but this trade-off is rarely evaluated systematically across different methods.

When confident compound identification is not feasible, molecular networking (MN) has emerged as a powerful framework for organizing MS^2 spectra into clusters of related molecules, enabling the discovery of chemical families and transformation

products across large datasets. [20–23] In this context, MN can also be applied to isobaric compounds to provide additional insight into their spectral and structural relatedness. However, MN fundamentally relies on pairwise MS^2 similarity assessment and user-defined thresholds. Therefore, the structure and interpretability of resulting networks are directly dependent on the behavior of the underlying similarity metrics.

Over time, various similarity metrics have been developed and can broadly be categorized as cosine-based, distance-based, probability-based, and machine learning-based methods, each with its advantages and limitations. Among these, cosine similarity remains one of the most widely used metrics due to its simplicity and efficiency. [24] By treating spectra as vectors, cosine similarity calculates the angle between them, where a perfect overlap (angle of 0°) yields a similarity score of 1.0.

Cosine scoring is simple and interpretable but highly sensitive to noise, missing peaks, and intensity scaling, which can reduce its robustness across different instruments and acquisition settings. Entropy-based metrics address some of these issues by considering the distribution of fragment intensities across spectra, potentially improving tolerance to peak intensity variations. [24,25] Binary similarity coefficients, such as the Jaccard or Dice index, treat spectra as sets of fragment peaks and compare them based on shared presence or absence. That makes them robust to intensity differences, yet they ignore quantitative information, which can limit their discriminatory power. More recently, machine learning approaches, including deep learning models like MS2DeepScore [26] or Spec2Vec [27], have been developed for spectral matching. These methods can capture complex, non-linear relationships between spectra, but they depend on large, well-curated training datasets and may lack interpretability compared with conventional similarity metrics.

In general, alternatives to cosine-based similarities have demonstrated enhanced performance in structural similarity assessment and showed improved accuracy in compound identification. [24, 25] However, comprehensive comparative evaluations that incorporate both classical and emerging similarity metrics remain limited. This knowledge gap may hinder the standardization and reproducibility of LC-HRMS-NTA data-processing workflows, potentially impacting the reliability of downstream interpretations and conclusions. To address this, we systematically benchmarked 20 MS^2 spectral similarity metrics, including classical, entropy-based, and machine learning approaches, using controlled local molecular networks constructed from spectra sharing nearly identical precursor masses. Rather than evaluating complete molecular networking

workflows or library-search pipelines, our goal is to isolate and investigate the ability of MS^2 similarity metrics alone to discriminate multiple candidate compounds which remain plausible after precursor mass filtering. By focusing on local, isobaric candidate sets, we provide an assessment of compound-level resolution achievable using MS^2 information alone. This study provides an extensive, threshold-resolved comparison of similarity metrics and examines how they affect cluster resolution within local molecular networks.

This study focuses on large-scale statistical behavior of similarity metrics rather than detailed spectrum-level interpretation. While specific spectral features such as peak density, noise, and shared fragment patterns are known to influence similarity scores, a systematic investigation of these effects would require a dedicated, feature-level analysis. Here, we instead evaluate how these combined effects manifest across diverse datasets in terms of network connectivity and clustering behavior.

5.2 Experimental Section

5.2.1 Overall Workflow

The dataset was compiled from MS^2 spectra extracted from publicly available databases and divided into 8,290 subdatasets that met predefined selection criteria (Section 5.2.2 Dataset). For each subdataset, similarity matrices were computed using 20 different similarities (Section 5.5 Section S1). An iterative thresholding approach, ranging from 0.0 to 1.0 with a step size of 0.1, was then applied to construct local molecular networks, and confusion matrices based on the resulting connections were recorded. Performance metrics derived from these confusion matrices were subsequently calculated and analyzed to evaluate the behavior of each similarity measure across thresholds. Finally, the complete database containing all subnetworks and the optimal threshold determined for each similarity measure has been made publicly available at `https://raw.githubusercontent.com/ViktoriiaTurkina/From-Spectra-to-Local-Networks-Evaluating-MS-Similarity/refs/heads/main/optimlal_thr_database.csv`. (Figure 5.1).

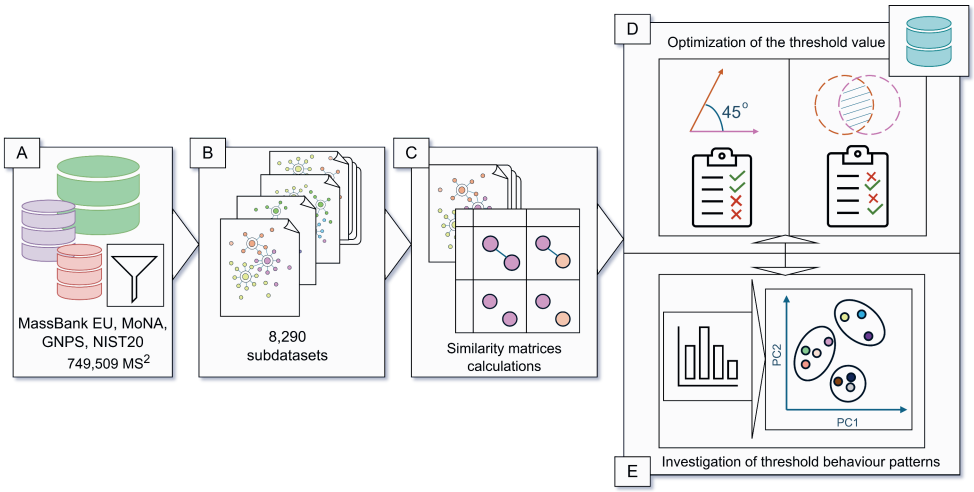


Figure 5.1: Overview of the study workflow, including dataset compilation and filtering (A), subdivision into smaller datasets according to predefined criteria (B), construction of local molecular networks including similarity and confusion matrices calculation (C), determination of the optimal threshold (D), and evaluation of similarity measure performance across thresholds(E).

5.2.2 Dataset

The MS^2 spectra dataset was combined from entries in the MassBank EU, [28, 29] MoNA, [30] GNPS, [31] and NIST20 [32] databases (in total 1,095,389 spectra). Only experimentally acquired spectra were retained from MoNA. The dataset includes data from both Q-TOF and Orbitrap mass analyzers acquired under diverse experimental conditions, including varying collision energies. This variability is inherent to public spectral libraries and is intentionally retained to evaluate similarity metric behavior under realistic, heterogeneous conditions. Spectra with no recorded resolution and with a resolution of ≤ 5000 were filtered out, keeping only high resolution spectra. To ensure data quality, only spectra with at least three recorded fragments were included, resulting in a total of 988,534 spectra. We acknowledge that duplicate spectra cannot be fully excluded when using aggregated public databases. However, their impact is limited, as our analysis focuses on relative differences between similarity metrics across many subnetworks rather than on individual spectral matches.

Firstly, the dataset was divided into two datasets: measured in positive and negative mode. Only data acquired in positive mode was investigated in this study. The dataset were divided into subdatasets that met three criteria: each subdataset contained at least five

spectra, two unique InChIKeys identifiers, and all the spectra within a subdataset shared the same precursor ion mass within a $\pm$5 mDa tolerance. Since all precursor ions within a dataset are already identical within m/z tolerance, precursor ions were removed from the recorded MS^2 spectra, as retaining them would artificially inflate similarity values. Intensities were normalized by dividing each intensity by the most intense fragment ion. In total for m/z tolerance of $\pm$5 mDa, 749,509 spectra (27,717 unique InChIKeys) were grouped into 8,290 of such subdatasets.

5.2.3 Similarity Metrics

To calculate similarity metrics, which requires alignment (Section S1 in Section 5.5), each MS^2 spectra in a subdataset were aligned pairwise prior to the calculations. All m/z values of the fragment ions were kept from both spectra, resulting in two vectors of intensities in case of continuous similarity or binary vectors in case of a binary calculations. The alignment was performed within tolerance of $\pm$5 mDa. If fragments were detected in one spectra out of two, their intensity were recorded as 0 and added to a corresponded vector. The constructed vectors were further used for similarity calculations.

In this study we investigated distance metrics and similarity indices as Jaccard similarity (both binary and taking intensities into account), entropy and weighted entropy, [25] and machine learning based, MS2DeepScore. [26] It should be noted that MS2DeepScore is trained on large, publicly available spectral repositories, and therefore some degree of overlap between training and evaluation data cannot be fully excluded. The full list of the 20 investigated similarity metrics is reported in SI (Section S1 in Section 5.5). Pairwise comparison of each spectrum resulted in a similarity matrix for each subdataset and each similarity measure.

5.2.4 Local Molecular Networks

Rather than using a generic molecular network where all nodes are connected through several generations, only a subset of the closest edges were utilized in this study (star-like network). The use of a single central node and first-generation network architecture allows each evaluated edge to correspond to a direct pairwise MS^2 similarity comparison between a query spectrum (the central node) and its candidate spectra within the same precursor-

mass–constrained subset. By restricting the analysis to direct connections, we avoid the influence of indirect connections through intermediate nodes, causing nodes to cluster together despite weak or absent direct similarity. Therefore, star-like networks ensure that observed differences in network connectivity and purity can be directly attributed to the behavior of the underlying similarity metrics rather than network propagation effects.

A similarity matrix was first computed for all spectra in a subdataset, and a threshold was defined to estimate significant molecular similarity for network construction. For each InChIKey in a subdataset, all nodes with a similarity higher than the threshold were considered connected via edges, and the node with the highest number of edges was selected as the central node; in cases of ties, a node was randomly chosen. Each local network contains multiple central nodes, with the number of central nodes corresponding to the number of unique InChIKeys present in the subdataset used to generate the local network (Section 5.5 Figure S1).

For each subset corresponding to a single molecular network, an empty graph was initialized with a number of nodes equal to the number of entries in the subset. Edges for each central node were then added iteratively: non-central nodes with similarity greater than or equal to the threshold were connected to a central node, and the resulting connections were classified as true positives (TP) if the nodes shared the same InChIKey or false positives (FP) if they differed. Non-central nodes with similarity below the threshold were evaluated as false negatives (FN) if they shared the same InChIKey or true negatives (TN) if they differed. Connections between central nodes were also evaluated; edges were added if similarity exceeded the threshold and classified as FP, whereas non-connections were recorded as TN. This procedure was applied across all subsets and the thresholds from 0.0 to 1.0 with the step of 0.1 to construct local networks and to quantify network performance (Section 5.5 Figures S1 - S3). In this study, TP, FP, FN, and TN are used as statistical descriptors of identity-consistency under a given similarity threshold and should not be interpreted as strict indicators of identification correctness.

For each local molecular network, the optimal threshold is defined as the value that yields the maximum Jaccard score (Equation 1), since this metric accounts for both correctly connected nodes and those that are incorrectly disconnected. Thus, the Jaccard score was used in this study to evaluate cluster purity and is equivalent to the purity metric.

$$\text{Jaccard score} = \frac{TP}{TP + FP + FN} \tag{1}$$

The star-like topology used here should be viewed as an evaluation tool rather than a representation of commonly applied molecular networks. [15] While global and multi-generation networks provide valuable context for chemical exploration, our architecture allows direct attribution of observed performance differences to the similarity metrics themselves.

5.2.5 Performance assessment

Based on the confusion matrix (TP, TN, FP, FN), defined in Section 5.2.4 Local Molecular Networks, performance metrics were calculated for each subdataset: Jaccard score, sensitivity, specificity, precision, false discovery rate (FDR), accuracy, and F1-score (Equation 1 and Equations S2 - S7 in Section 5.5). For each similarity measure, network performance was assessed at varying thresholds using calculated metrics. The Jaccard score was chosen instead of accuracy as a measure of purity because as it accounts for the overlap between predicted and actual positive connections, while accuracy (Section 5.5 Equation S6) also accounts for true negatives. Thereby, focus on Jaccard score allows to reduce bias from the large number of true negatives, which is common in molecular networks. To enable comparison across similarities, the median value of each performance metric was calculated at each threshold.

The performance metrics reported in this study are intended as diagnostic indicators of MS^2 similarity behavior rather than absolute measures of identification correctness. Because MS^2 similarity alone cannot always uniquely distinguish chemical identity, especially for isomeric or closely related compounds, these metrics reflect the discriminative limits of similarity-based matching under the evaluated conditions.

To assess the level of structural similarity within unresolved subnetworks, we calculated Tanimoto similarity [33] using two molecular fingerprint representations: PubChem [34] and Atom Pair Count (APC). [35] Unresolved subnetworks were defined as those in which no similarity metric achieved a purity of 1.0, regardless of the applied threshold. For each unresolved local network, pairwise Tanimoto similarities were calculated separately for connected and disconnected node pairs (i.e., compound structures) at the optimal sim-

ilarity threshold, and the values were averaged within each group for the subnetwork. For this analysis, we used local networks constructed using cosine similarity as the representative MS^2 similarity metric, as it is the most widely used and commonly accepted approach in the field.

5.2.6 Multivariate analysis

To explore relationships among the evaluated spectral similarity metrics, a multivariate analysis was performed. Principal component analysis (PCA) was applied to a matrix comprising of the median values of four performance metrics, such as specificity, sensitivity, precision, and Jaccard score. The metrics were calculated across all sub-networks and similarity thresholds. The PCA was conducted using scikit-learn, with data mean-centered and scaled to unit variance before the analysis.

Additionally, hierarchical clustering was performed to group similarities with similar performance profiles. Euclidean distance was used as the similarity metric, and the ward linkage method was applied to construct the dendrogram. The resulting hierarchy was cut to determine five major clusters. The number of clusters was defined based on PCA results.

5.2.7 Calculations

All calculations were performed on a personal computer (PC) with an Intel Core i7-1260P central processing unit and 32 GB of RAM operating Windows 10 Education version 22H2. All data processing and statistical analysis were performed using Julia language version 1.10.6. The script is publicly available at `https://github.com/ViktoriiaTurkina/From-Spectra-to-Local-Networks-Evaluating-MS-Similarity`.

5.3 Results and Discussions

Initially, 749,509 spectra acquired in positive ion mode extracted from the combined spectral library were grouped into 13,638 subdatasets. Of these, 5,348 subdatasets (i.e. 5,348 unique chemicals) were excluded from the analysis because they contained only a single

InChIKey and therefore could not be used to assess cluster resolution based on multiple unique InChIKeys.

The dataset used in this study comprised of a total of 8,290 subdatasets of MS^2 spectra that met the predefined filtering criteria (Section 5.2.2 Dataset), using an m/z tolerance of $\pm$5 mDa. The majority of these subnetworks were small, typically containing only a few unique molecular identifiers. Approximately 26% (2,140 subdatasets) consisted of two unique InChIKeys, each represented on average by 21 spectra (within range between 5 and 214) and covering a precursor ion m/z range of 59.05–1753.73 Da. Another 14.8% (1,223 subdatasets) contained three unique InChIKeys, averaging 32 spectra (within range between 6 and 209) and spanning over m/z 73.07–1590.85 Da. Subdatasets with four or more unique InChIKeys accounted for fewer than 10% of the total, while the largest subdataset contained 119 unique InChIKeys and 1,097 spectra for a precursor ion at m/z 135.0441 (Section 5.5 Figure S4). These distributions indicate that the vast majority of subnetworks are small, with relatively few unique molecular entities per network, and only a minor fraction of sub-networks consist of a larger number of distinct compounds.

It is important to note that these statistics strongly depend on the database used and applied filtering criteria. For example, increasing the m/z tolerance reduces the number of subdatasets to 4,254, while increasing the minimum number of InChIKeys and spectra per subdataset to three and seven, respectively. The used filtering criteria was designed to be as close as possible to real-world applications of spectral matching against experimental databases.

5.3.1 Performance trends across thresholds

We evaluated classification metrics, including sensitivity, specificity, precision, FDR, accuracy, F1-score, and Jaccard score (purity) for 20 spectral similarity metrics across increasing similarity thresholds, from 0 to 1.0 with a step of 0.1. The values of the performance metrics varied widely across the full range of thresholds. Therefore, to capture general trends, the median value of each metric was calculated across 8,290 subnetworks for every similarity measure and threshold (Figure 5.2).

False links are removed more rapidly than true connections, when looking at the trends in the median numbers of true positives (TPs) and false positives (FPs) across thresholds. At

low thresholds, many spectra are connected, including both true and false links, resulting in dense molecular networks. By a threshold of approximately 0.1, many false connections have already disappeared. In contrast, some true positives, particularly under distance-based similarities such as Chebyshev and Manhattan, are lost early as well. These metrics are the most conservative, as they are highly restrictive and retain only spectra that are extremely similar, discarding all others. Even small differences between spectra result in very low similarity scores, which causes true matches to be discarded at relatively low thresholds (Section 5.5 Figure S5). At low thresholds, sensitivity is high, since most true connections are included, but precision is low because false positives are also linked. As thresholds increase, weaker connections (i.e. low similarity spectra) are progressively removed. At moderate thresholds, true but lower-scoring connections are lost earlier than some high-scoring false positives, leading to a rapid decline in sensitivity and F1-score. At higher thresholds, nearly all remaining links correspond to high-confidence matches, producing sharply decreasing FDR and near-perfect specificity. Precision increased steadily with threshold, reflecting the gradual removal of false-positive connections as similarity cutoffs became tighter. The complementary FDR metric decreased accordingly, confirming the expected trade-off between inclusiveness and selectivity. Together, these patterns show that while higher thresholds improve precision, they do so at the expense of overall match completeness, as reflected in declining sensitivity and F1-score. Accuracy and F1-score exhibited nearly identical threshold-dependent behavior, both peaking between thresholds of 0.4 and 0.7, indicating an optimal balance between correctly identified and missed spectral matches. Beyond this range, Accuracy and F1-score declined as stricter thresholds caused a loss of true connections within molecular networks. This similarity is related to the dominance of true negatives and the relatively low number of false assignments in the evaluated local networks. Since both metrics depend primarily on the proportion of remaining true links (TP) relative to false (FP) or missing ones (FN), they respond similarly to threshold changes. Thus, the loss of true connections dominates the observed performance.

Importantly, false connections, positive or negative, do not necessarily indicate incorrect or misleading matches in a practical annotation context. Instead, they also reflect inherent spectral overlap between chemically distinct compounds or variability introduced by experimental conditions. In particular, unresolved cases involve positional or regioisomers, for which near-identical fragmentation patterns are expected.

Jaccard score values exhibited a distinct peak at intermediate thresholds (0.4–0.6), with maximal median values ranging from 0.45 to 0.52 for w-Entropy, Entropy, and Jaccard. Beyond this range, performance declined, indicating that moderate similarity thresholds provide the best balance between intra-cluster homogeneity and inter-cluster separation. The position of this peak varied among similarities: low-threshold maxima (0.0–0.4) were observed, for example, for Bhattacharyya, Manhattan, and Jaccard similarities, intermediate peaks (0.4–0.6) for bi-Cosine, and high-threshold maxima (>0.6) for MS2DeepScore. This variation reflects differences in how strictly each measure retains spectral links within clusters. For example, Jaccard and bi-Jaccard perform well at lower thresholds but rapidly lose both true and false connections as thresholds increase. These similarities are prone to network fragmentation and may underestimate relationships between moderately similar spectra. In contrast, MS2DeepScore preserves more connections even at high thresholds, demonstrating greater tolerance to spectral variability, although this behavior also increases the likelihood of retaining false-positive links. The maximum achievable Jaccard score (cluster purity) also differed among similarities. For instance, cosine peaked at a similar threshold as w-entropy but reached a lower median value (0.45 versus 0.53), indicating differences in clustering selectivity (Figure 5.2 and Figures S6-S7 in Section 5.5).

To illustrate these general trends, figure 5.3 shows how increasing the similarity threshold (0.1, 0.7, and 0.9) affects local molecular network structure for five representative similarity metrics: bi-Cosine, Inner Product, Jaccard, MS2DeepScore, and w-Entropy (Section S7 in Section 5.5). At low thresholds (0.1), networks appear highly connected, incorporating both true and false links. As the threshold increases, weaker connections are progressively removed, resulting in fragmentation of the network into smaller, more specific clusters. Similarities such as Jaccard and bi-Cosine rapidly lose connectivity, while MS2DeepScore maintains more links even at high thresholds, consistent with its broader tolerance for spectral variation. These examples visually demonstrate the threshold-dependent trade-off between network completeness and purity described above.

Overall, our results showed that thresholding first breaks true connections before eliminating residual false ones. This pattern is consistent with the conservative behavior of most similarity metrics. Therefore, these similarities are strongly biased toward high-confidence connections: as thresholds increase, they rapidly eliminate uncertain

or moderately similar spectra, resulting in steadily increasing precision and specificity but a pronounced loss of sensitivity. This selectivity favors network purity and minimizes false positives at the expense of completeness. Consequently, the resulting molecular networks are relatively sparse but are likely to consist of connections characterized by high similarity spectra.

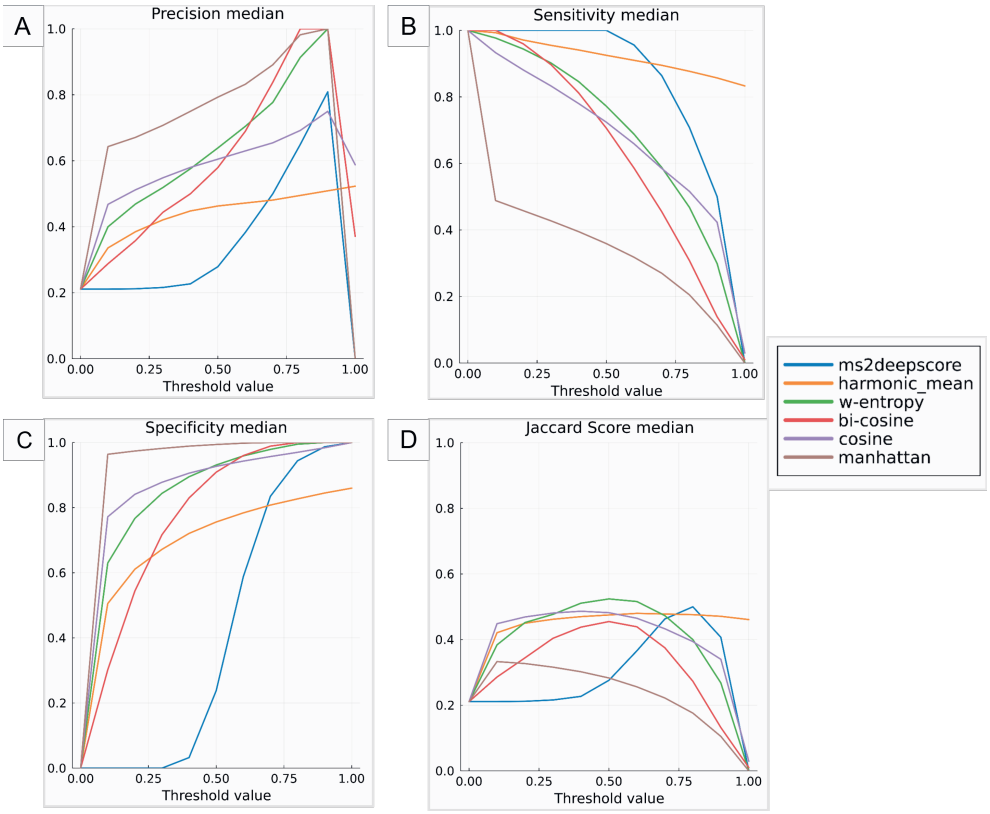


Figure 5.2: Performance for selected similarity metrics across similarity thresholds (0–1.0, step 0.1), evaluated using precision (A), sensitivity (B), specificity (C), and Jaccard score (purity) (D). The representative similarity metrics were selected from each category discussed in Section 5.3.2, including distance-based (e.g., Manhattan, Harmonic Mean), continuous (e.g., Cosine, W-Entropy), binary (e.g., Bi-Cosine), and machine-learning–based approaches (e.g., MS2DeepScore).

5.3.2 Clustering and Multivariate Analysis

To investigate patterns in performance trends across all similarity metrics, principal component analysis (PCA) was performed on a matrix combining selectivity, sensitivity, precision, and Jaccard score values (Section 5.5 Figure S9). The first two principal components

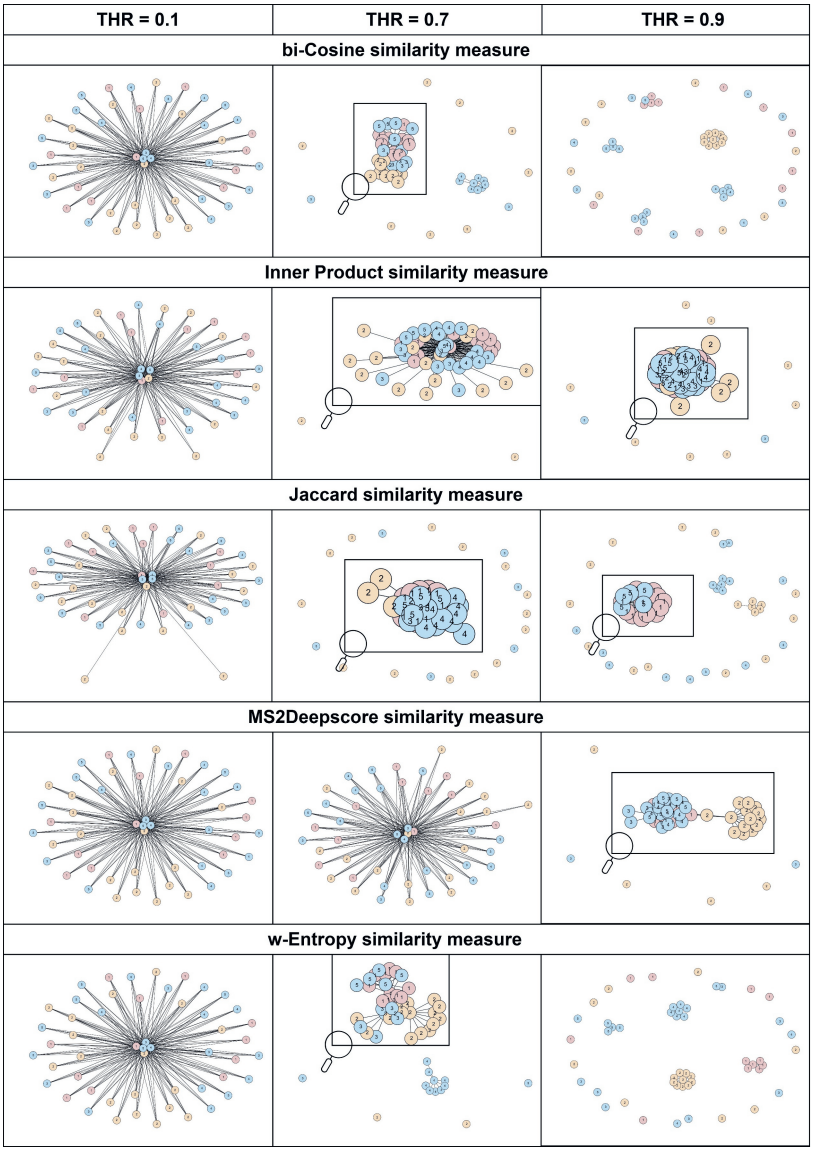


Figure 5.3: Threshold-dependent effects of similarities for the subdataset with the precursor ion m/z 116.143. Shown are the effects of bi-Cosine, Inner Product, Jaccard, MS2DeepScore, and w-Entropy across thresholds (0.1, 0.7, 0.9), Section S7 in Section 5.5.

captured the most variance (i.e. 80%) among metrics, highlighting differences in their overall behavior across thresholds. Additionally, hierarchical clustering (HCA) was then applied to the same data as PCA, and five clusters were defined by cutting the dendrogram at k = 5. This combined approach allowed us to group similarity metrics according to their shared response patterns and underlying computational characteristics.

HCA revealed five groups, each reflecting fundamental differences in computational formulation and information content (Section 5.5 Figure S10). MS2DeepScore formed a clearly isolated cluster, distinct from all other metrics. Its deep learning–based architecture captures nonlinear spectral relationships in a learned embedding space, explaining its different performance profile compared with formula-based similarities. In contrast, Bhattacharyya2, Inner Product, and Harmonic Mean formed a second cluster, reflecting their shared emphasis on proportional fragment intensities and continuous intensity correlations. These metrics displayed similar threshold-dependent behaviors, balancing sensitivity and precision across the evaluated sub-networks. The remaining similarity metrics were less distinctly separated in the PCA space, yet hierarchical clustering identified three additional groups. The first included Cosine, Intersection, Spectral Contrast Angle, Dice, Squared Euclidean, w-Entropy, Entropy, and Dot Product. These metrics rely on geometric or information-theoretic formulations that quantify spectral overlap and intensity coherence, leading to comparable performance across thresholds. Another group comprised Probabilistic Symmetric Chi-squared, Jaccard, Manhattan, Chebyshev, and Penrose shape, which are distance-based metrics emphasizing absolute rather than proportional differences in fragment intensities. Their conservative behavior results in higher specificity but lower sensitivity, often producing more fragmented molecular networks. Finally, Baroni–urbani–buser, bi-Jaccard, and bi-Cosine formed a distinct cluster representing binary similarity formulations. These metrics operate on the presence or absence of fragments rather than their intensities, producing discrete, stepwise threshold responses that distinguish them from continuous-intensity metrics.

Overall, the clustering reflects the conceptual divisions among the similarity metrics. Continuous intensity-based metrics, such as cosine, entropy variants, and inner product, form cohesive groups, while distance based metrics segregate due to their sensitivity to dissimilarity. Binary metrics occupy a separate space defined by their discrete representation of spectral data. MS2DeepScore clustered separately from all other metrics, showing distinct behavior characterized by retaining both true and false links at high thresholds,

suggesting a less restrictive connection profile.

5.3.3 Set threshold

Among all datasets, 674 subnetworks (8.1%) reached the highest possible purity (1.0) at the commonly recommended similarity threshold of 0.7. [36] Of these, 255 subnetworks (3.1%) were resolved exclusively by a single similarity metrics, and 131 subnetworks (1.6%) were uniquely resolved using MS2DeepScore, meaning that no other single similarity metric was able to achieve the purity of 1.0 for these cases. Despite these cases of complete separation, approximately 92% of subnetworks were not resolved by any similarity metric. This indicates that, in most instances, the combination of MS^2 spectral information content and the applied similarity metrics is insufficient to achieve fully pure and independent clusters of unique compounds within local molecular networks. Consequently, a substantial fraction of nodes could not be confidently assigned to a single molecular structure, even after extensive database matching. Across all datasets and similarity metrics, the median purity was 0.36. The highest median purity values (0.48–0.46) were observed for Inner Product, w-Entropy, MS2DeepScore, and Entropy. However, the distribution of purity values varied widely between similarity metrics, and in 82% of all cases, at least one measure failed to achieve any resolution showing a minimum purity of 0.0 (Section 5.5 Figure S11). MS2DeepScore and Inner Product exhibited broader distributions than Entropy, Cosine, or Dice, suggesting that their performance may be more sensitive to dataset composition. Overall, while some methods achieved perfect clustering in specific datasets, purity remains highly dataset-dependent and rarely exceeded moderate resolution.

Finally, we compared performance metrics, such as sensitivity, specificity, and precision (Section 5.5 Figures S12 - S14), where MS2DeepScore, Harmonic Mean, and Bhattacharyya 2 displayed the highest sensitivity, characterized by high medians and narrow interquartile ranges, followed by Inner Product. Thus, the deeplearning–based MS2DeepScore performed competitively with traditional similarities, suggesting that learned vector-space embeddings and classical spectral similarities achieve comparable accuracy in identifying correct matches at this threshold. In contrast, Manhattan, bi-Jaccard, and Jaccard distance showed the lowest sensitivities, indicating limited ability to recover true positive connections. Across all metrics, specificity remained consistently high, whereas sensitivity showed greater dispersion. This pattern suggests that most

metrics are conservative at the 0.7 threshold, effectively minimizing false positives but at the cost of reduced detection of true connections. However, the number of unresolved cases remains significant. Even when using large and comprehensive databases, overall accuracy typically does not exceed 50%. This limitation is evident from the global statistics, which reveal a saturation effect: increasing database size improves coverage but not proportional accuracy.

5.3.4 Optimal threshold

Different similarity metrics have a pronounced effect on both the resulting molecular network structure and accuracy of connected nodes. Each metric can be characterized by an optimal similarity threshold. In this study, the optimal threshold for each local molecular network was determined as the value at which the Jaccard score reached its maximum. Such a threshold represents the best balance between correctly connected (i.e. TP) and incorrectly disconnected (i.e. FN) nodes within a local network. Across all networks, the median optimal thresholds ranged from as low as 0.1 to as high as 0.8, indicating high variability in the sensitivity of different similarity metrics.

Metrics such as Manhattan and Chebyshev distances exhibited narrow distributions skewed toward lower threshold values, suggesting that they connect spectra only at low similarity levels and behave conservatively by minimizing false-positive connections. In contrast, metrics like bi-Cosine displayed broader distributions, indicating greater variability in performance across datasets. While MS2DeepScore was skewed toward higher thresholds, reflecting its tendency to retain connections even at stricter cutoffs, this behavior may also increase the number of false positives depending on the reference library. This variability underscores that each similarity metric captures different aspects of spectral match, leading to distinct network connectivity and optimal filtering behavior. Additionally, each similarity produces networks of different densities and reaches its optimal balance between true and false connections at distinct threshold levels for each network. Given the optimal thresholds, local molecular networks reached different maximum values for the Jaccard score, with median values across all datasets ranging between 0.41 and 0.60. Interestingly, Inner Product and Harmonic Mean displayed a bimodal distribution of optimal thresholds, with a primary peak around 0.5 and a secondary peak near 1.0. This pattern suggests two characteristic behaviors across datasets: one around

0.5, representing partially resolved networks, and another near 1.0, corresponding to cases where networks were almost perfectly resolved into pure clusters (Figure 5.4 and Section 5.5 Figure S15). A total of 1,891 subnetworks (22.8% of all evaluated) reached a Jaccard score of 1.0 at the evaluated thresholds, indicating fully resolved subnetworks. This proportion is almost twice as high as that obtained using the similarity threshold of 0.7. Among these, 256 subnetworks were resolved exclusively by a single similarity measure. This number is nearly identical to that observed at the fixed threshold, differing by only ten cases. With the majority resolved by w-entropy (55 cases, 21.5%) and MS2DeepScore (53 cases, 20,7%), cumulatively covering 42,2% of the total number of resolved networks with a single similarity metric. However, despite optimized thresholds, the share of these cases of complete separation increase only by around 10% resulting in 80.2% in contrast to previous 92% of unresolved by any similarity metric sub-networks. Therefore, even with optimized thresholds, a major fraction of networks remains unresolved.

These observations are illustrated by practical examples (Section S12 in Section 5.5). For a simple candidate set (m/z 268.1993, two InChIKeys), optimized thresholds yielded clearly separated clusters (Jaccard score $\approx$ 0.90 at threshold of 0.5 using entropy similarity) (Section 5.5 Figure S18). In contrast, for a more complex set (m/z 229.0932, 38 InChIKeys), the maximum Jaccard score remained low (Jaccard score $\approx$ 0.14 at 0.6 using Jaccard similarity) (Section 5.5 Figure S19), and the network was highly interconnected. This highlights that, even under optimized conditions, resolving candidate structures depends strongly on dataset complexity.

Finally, we examined structural similarity within unresolved subnetworks at the optimal threshold using Tanimoto similarity calculated from molecular fingerprints. Both connected and disconnected compound pairs exhibited broad and strongly overlapping similarity distributions. This indicates that unresolved subnetworks contain both structurally similar (e.g., isomeric) and structurally distinct compounds (Section 5.5 Figures S16 - S17). In other words, compounds retained in the network and those excluded at the optimized threshold do not differ systematically in their structural similarity. Consequently, unresolved local networks at individually optimized thresholds do not consist exclusively of structurally similar compounds.

These findings indicate that optimizing the threshold for each similarity metric and dataset allows some subnetworks to be fully resolved, with all clusters containing a single compound. Overall, this demonstrates that tailoring the threshold to the behavior of each

similarity metric can improve network purity and, potentially, increase confidence of identification through information-driven selection rather than ranking candidates alone.

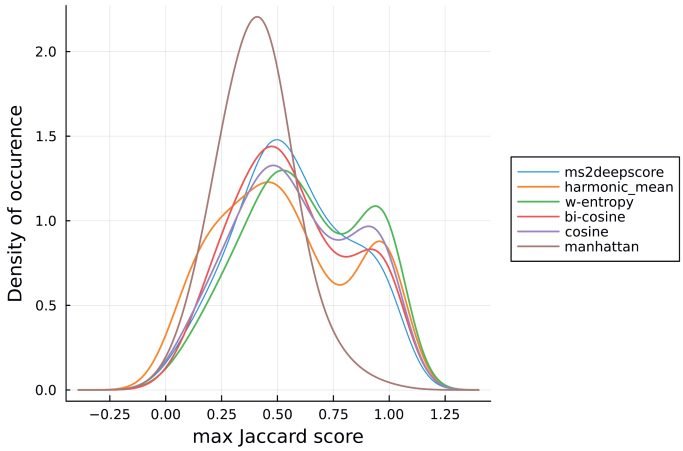


Figure 5.4: Density distributions of the maximum Jaccard scores achieved across all subnetworks for representative similarity metrics from each category discussed in Section 5.3.2 is shown, including distance-based (e.g., Manhattan, Harmonic Mean), continuous (e.g., Cosine, W-Entropy), binary (e.g., Bi-Cosine), and machine-learning–based approaches (e.g., MS2DeepScore).

5.3.5 Number of fragments versus purity of networks

We compared the distribution of fragment counts between resolved and unresolved subnetworks to evaluate whether the number of fragments per spectrum influences network resolution. Initially, 1125 subnetworks were classified as resolved and 6655 as unresolved. The maximum number of unique compounds in the unresolved dataset was 119, whereas in the resolved dataset it was only eight. This pattern indicates that subnetworks containing a larger number of unique compounds tend to be more difficult to resolve, suggesting that chemical heterogeneity within a network reduces the likelihood of achieving complete separation. To minimize the influence of this effect on the result of the analysis, only subnetworks with no more than eight unique compounds were retained from the unresolved dataset, resulting in 4534 subnetworks used for comparison. A similar filtering step was applied to precursor ion masses: since the maximum precursor m/z in resolved subnetworks was 1125 and 1753 in unresolved ones, only subnetworks with precursor masses below 1125 m/z were retained, yielding 4503 unresolved networks for analysis.

The overall shapes of the distributions (along with their maximum, minimum, and me-

dian values) did not differ significantly between resolved and unresolved networks, based on the results of density fitting of fragment count distributions across all subnetworks (Figure 5.5). This suggests that the number of fragments per spectrum has only a weak effect on the ability to resolve clusters in a network, and that factors other than fragmentation complexity, such as spectral similarity behavior or chemical diversity within a network, likely play a more dominant role in determining network resolution.

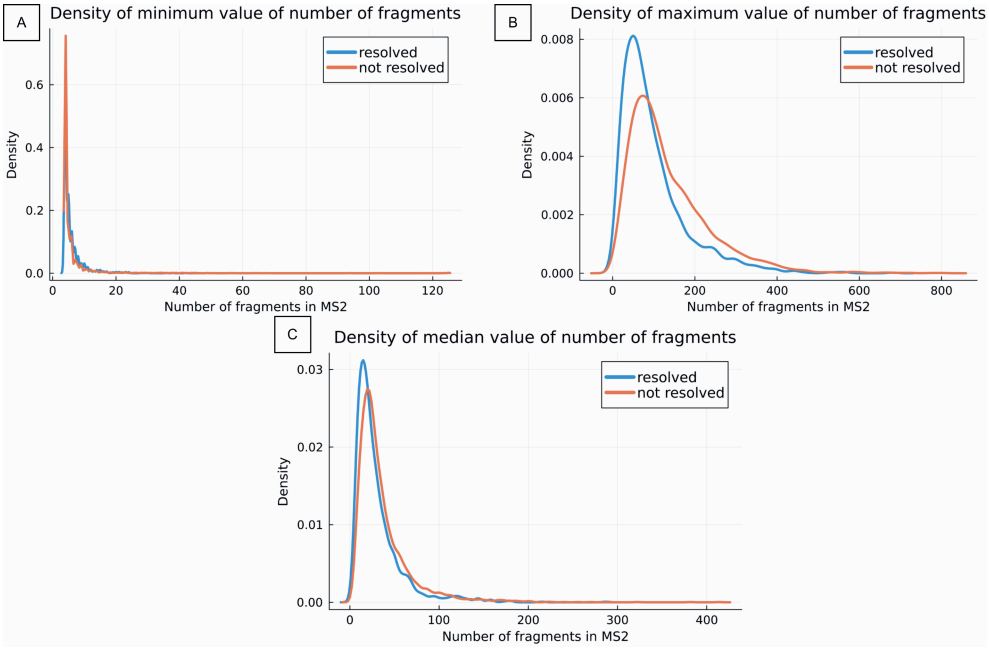


Figure 5.5: Density distributions of the minimum (A), maximum (B), and median (C) fragment counts across all subnetworks. Fragment count refers to the number of MS^2 fragments per spectrum. "resolved" denotes local networks in which at least one similarity metric achieved a purity of 1.0, while "not resolved" denotes local networks in which a purity of 1.0 was not achieved by any similarity metric.

5.4 Conclusions

Our results suggest that adjusting the number of fragments included in spectral similarity calculations does not significantly improve the ability to separate chemically distinct compounds within molecular networks. The resolution of subnetworks appears to be largely independent of fragmentation complexity, suggesting that similarity-based connectivity is driven more by the chosen metrics and thresholding behavior than by the number of fragment ions. This finding highlights an important limitation of pairwise spectral

comparisons: improvements in resolution are unlikely to be achieved simply by tuning fragment inclusion criteria or applying generic similarity metrics with the commonly used threshold. Instead, a more effective strategy involves cumulative approaches that combine information across multiple candidate matches and incorporate network-level relationships. By integrating network-based information, such as shared connectivity patterns, spectral neighborhoods, or co-occurrence within chemical families. These methods can provide a more robust and informative representation of molecular similarity. This network-informed perspective shifts the focus from isolated spectrum-to-spectrum comparisons towards a collective evaluation of spectral and structural relationships, improving both annotation reliability and interpretability.

Although applying the optimal thresholds improved network resolution, raising the proportion of completely resolved subnetworks (Jaccard score = 1.0) from 8.1% to 22.8%, complete separation of candidate compounds remained rare. Perfect spectral discrimination is generally unachievable for complex datasets. Even with optimized thresholds and advanced similarity metrics, many subnetworks remain partially mixed due to intrinsic spectral overlap and limited reference coverage in existing libraries. Consequently, most library-based spectral searches can rarely achieve annotation confidence beyond level 3. These results further highlights the need for orthogonal sources of information, such as retention behavior, method chemical space coverage, ion mobility, and analyst's prior knowledge, for being able to increase the confidence in identifications.

From a practical perspective, the results of this study suggest several ways in which local molecular networking approach can be applied in NTA workflows. One example (Section S12 in Section 5.5), is a two-step strategy in which similarity metrics and their corresponding thresholds are first calibrated using reference data, and subsequently applied to candidate spectra associated with a given precursor ion m/z value. The resulting local molecular network provides an indication of whether candidate structures can be separated based on MS^2 similarity.

More broadly, the observed behavior of similarity metrics can lead their selection depending on the analytical objective. Metrics differ in their sensitivity to spectral variability, resulting in varying degrees of connectivity and cluster fragmentation. In addition, evaluating network structure across metrics or threshold settings can provide insight into the stability of similarity-based relationships, distinguishing between cases where candidate separation is feasible and those where structural ambiguity remains.

In this sense, local molecular networks are not intended as standalone identification tools, but rather as a complementary framework for interpreting similarity-based evidence and assessing the resolving power of MS^2 similarity in a given context.

Overall, the design of the study simulates a critical decision point common to both molecular networking and library searching, namely, how confidently MS^2 similarity alone can separate structurally distinct compounds once mass-based filtering has been satisfied. Consequently, the results should not be interpreted as a measure of performance for analog discovery across wide chemical space or as a direct estimate of identification accuracy in specific databases. Instead, they provide an assessment of the resolving power and limitations of commonly used similarity metrics in challenging yet realistic scenarios involving isobars, isomers, and closely related compounds.

Moreover, in addition to structural similarity, experimental heterogeneity across the spectral libraries, including differences in instrument type, collision energy, and acquisition protocols, contributes to variability in MS^2 spectra. Such heterogeneity can reduce similarity between spectra of the same compound measured under different conditions, while increasing similarity between different compounds acquired under similar conditions. These effects also contribute to the observed trend patterns across similarity metrics. However, our results highlight the fundamental limitations of spectral similarity metrics, especially when mass spectra are used as the sole source of information for compound identification.

5.5 Supporting Information

The Supporting Information is available at `https://pubs.acs.org/doi/10.1021/acs.analchem.6c01202` with definitions of similarity metrics used in this study (Section S1), performance metrics calculations (Section S3), explanation of confusion-matrix generation (Figures S1–S3), distribution of InChIKeys across all sub-data sets (Figure S4), heatmaps of true-positive and false-positive counts across thresholds for all similarities (Figure S5), median performance trends for all metrics and similarities (Figures S7–S6), structural information and threshold-dependent network behavior for the example sub-data set (m/z 116.143) (Figure S8), PCA score plot and hierarchical clustering of similarity metrics (Figures S9 and S10), boxplots of Jaccard score, sensitivity, specificity, and preci-

sion at threshold 0.7 (Figures S11–S14), density distributions of maximum Jaccard score (Figure S15), schematics illustrating local network construction, central-node selection, density plots of average Tanimoto similarity (Figures S16-S17), and an application example (Section S12).

Notes

The script to perform the calculations is available at `https://github.com/ViktoriiaTurkina/From-Spectra-to-Local-Networks-Evaluating-MS-Similarity`.

Acknowledgements

The authors express their gratitude to all the members of Environmental Modeling & Computational Mass Spectrometry (www.emcms.info). They are also thankful to Jeremy Quail for their kind help and contribution to this study. S.S., A.P., and V.T. thank the ChemistryNL TKI and the UvA Data Science Centre for their funding support (projects EDIFIED and SCOPE). J.W.O. is the recipient of a National Health and Medical Research Council (NHMRC) Investigator Grant (EL12009209) funded by the Australian Government. J.W.O. also acknowledges financial support from the Australian National Health and Medical Research Council (NHMRC; APP1185347). The Queensland Alliance for Environmental Health Sciences (QAEHS), The University of Queensland (UQ), gratefully acknowledges the financial support of Queensland Health. For the purposes of open access, the author has applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising from this submission.

5.6 References

[1] Vermeulen R, Schymanski E. L, Barabási A. L, Miller G. W. The exposome and health: where chemistry meets biology. Science. 2020 1;367(6476):392. doi:10.1126/science.aay3164.

[2] Arp H. P. H, Aurich D, Schymanski E. L, Sims K, Hale S. E. Avoiding the Next Silent Spring: Our Chemical Past, Present, and Future. Environ Sci Technol. 2023 4;57(16):6355-9. doi:10.1021/acs.est.3c01735.

[3] Samanipour S, Barron L. P, van Herwerden D, Praetorius A, Thomas K. V, O'Brien J. W. Exploring the Chemical Space of the Exposome: How Far Have We Gone? JACS Au. 2024 7;4(7):2412-25. doi:10.1021/jacsau.4c00220.

[4] Escher B. I, Stapleton H. M, Schymanski E. L. Tracking complex mixtures of chemicals in our changing environment. Science. 2020;367(6476):388-92. doi:10.1126/science.aay6636.

[5] Hulleman T, Turkina V, O'Brien J. W, Chojnacka A, Thomas K. V, Samanipour S. Critical Assessment of the Chemical Space Covered by LC-HRMS Non-Targeted Analysis. Environ Sci Technol. 2023 9;57(38):14101-12. doi:10.1021/acs.est.3c03606.

[6] McEachran A. D, Sobus J. R, Williams A. J. Identifying known unknowns using the US EPA's CompTox Chemistry Dashboard. Anal Bioanal Chem. 2017 3;409:1729-35. doi:10.1007/s00216-016-0139-z.

[7] Sauvé S, Desrosiers M. A review of what is an emerging contaminant. Chem Cent J. 2014;8:15. doi:10.1186/1752-153x-8-15.

[8] Hollender J, Schymanski E. L, Singer H. P, Ferguson P. L. Nontarget Screening with High Resolution Mass Spectrometry in the Environment: Ready to Go? Environ Sci Technol. 2017 10;51:11505-12. doi:10.1021/acs.est.7b02184.

[9] Manz K. E, Feerick A, Braun J. M, Feng Y. L, Hall A, Koelmel J, et al. Non-targeted analysis (NTA) and suspect screening analysis (SSA): a review of examining the chemical exposome. J Expo Sci Environ Epidemiol. 2023 6;33(4):524-36. doi:10.1038/s41370-023-00574-6.

[10] Schymanski E. L, Jeon J, Gulde R, Fenner K, Ruff M, Singer H. P, et al. Identifying small molecules via high resolution mass spectrometry: Communicating confidence. Environ Sci Technol. 2014 2;48:2097-8. doi:10.1021/es5002105.

[11] Alygizakis N, Lestremau F, Gago-Ferrero P, Gil-Solsona R, Arturi K, Hollender J, et al. Towards a harmonized identification scoring system in LC-HRMS/MS based non-target screening (NTS) of emerging contaminants. Trends Anal Chem. 2023 2;159:116944. doi:10.1016/j.trac.2023.116944.

[12] Hollender J, Schymanski E. L, Ahrens L, Alygizakis N, Béen F, Bijlsma L, et al. NORMAN guidance on suspect and non-target screening in environmental monitoring. Environ Sci Eur. 2023 9;35(1):75. doi:10.1186/s12302-023-00779-4.

[13] Dulio V, Koschorreck J, van Bavel B, van den Brink P, Hollender J, Munthe J, et al. The NORMAN Association and the European Partnership for Chemicals Risk Assessment (PARC): let's cooperate! Environ Sci Eur. 2020 7;32(1):100. doi:10.1186/S12302-020-00375-W.

[14] Pérez-Pereira A, Carrola J. S, Tiritan M. E, Ribeiro C. Enantioselectivity in ecotoxicity of pharmaceuticals, illicit drugs, and industrial persistent pollutants in aquatic and terrestrial environments: A review. Sci Total Environ. 2024 2;912:169573. doi:10.1016/j.scitotenv.2023.169573.

[15] Aron A. T, Gentry E. C, McPhail K. L, Nothias L. F, Nothias-Esposito M, Bouslimani A, et al. Reproducible molecular networking of untargeted mass spectrometry data using GNPS. Nat Protoc. 2020 5;15(6):1954-91. doi:10.1038/s41596-020-0317-5.

[16] Schollée J. E, Schymanski E. L, Stravs M. A, Gulde R, Thomaidis N. S, Hollender J. Similarity of High-Resolution Tandem Mass Spectrometry Spectra of Structurally Related Micropollutants and Transformation Products. J Am Soc Mass Spectrom. 2017 12;28(12):2692-704. doi:10.1007/s13361-017-1797-6.

[17] Bazsó F. L, Ozohanics O, Schlosser G, Ludányi K, Vékey K, Drahos L. Quantitative Comparison of Tandem Mass Spectra Obtained on Various Instruments. J Am Soc Mass Spectrom. 2016 8;27(8):1357-65. doi:10.1007/s13361-016-1408-y.

[18] Ji H, Xu Y, Lu H, Zhang Z. Deep MS/MS-Aided Structural-Similarity Scoring for Unknown Metabolite Identification. Anal Chem. 2019 5;91(9):5629-37. doi:10.1021/acs.analchem.8B05405.

[19] Kind T, Tsugawa H, Cajka T, Ma Y, Lai Z, Mehta S. S, et al. Identification of small molecules using accurate mass MS/MS search. Mass Spectrom Rev. 2018 7;37(4):513-32. doi:10.1002/mas.21535.

[20] Yu J. S, Kwak Y. B, Kee K. H, Wang M, Kim D. H, Dorrestein P. C, et al. A versatile toolkit for drug metabolism studies with GNPS2: from drug development to clinical monitoring. Nat Protoc. 2025 9;15:1-35. doi:10.1038/s41596-025-01237-6.

[21] Wang Y, Wang Y, Zhang Z, Xu K, Fang Q, Wu X, et al. Molecular networking: An efficient tool for discovering and identifying natural products. J Pharm Biomed Anal. 2025 7;259:116741. doi:10.1016/j.jpba.2025.116741.

[22] Li W, Wu Z, Xu Y, Long H, Deng Y, Li S, et al. Emerging LC-MS/MS-based molecular networking strategy facilitates foodomics to assess the function, safety, and quality of foods: recent trends and future perspectives. Trends Food Sci Technol. 2023 9;139:104114. doi:10.1016/j.tifs.2023.07.011.

[23] Dalla Valle N, Garcia-Aloy M, Robatscher P, Franceschi P, Oberhuber M. Improving Spectral Similarity and Molecular Network Reliability through Noise Signal Filtering in MS/MS Spectra. Anal Chem. 2025 7;97(29):15873-82. doi:0.1021/acs.analchem.5c02109.

[24] Dlugas H, Zhang X, Kim S. Comparative analysis of continuous similarity measures for compound identification in mass spectrometry-based metabolomics. Chemom Intell Lab Syst. 2025 8;263:105417. doi:10.1016/j.chemolab.2025.105417.

[25] Li Y, Kind T, Folz J, Vaniya A, Mehta S. S, Fiehn O. Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification. Nat Methods. 2021 12;18(12):1524-31. doi:10.1038/s41592-021-01331-z.

[26] Huber F, van der Burg S, van der Hooft J. J. J, Ridder L. MS2DeepScore: a novel deep learning similarity measure to compare tandem mass spectra. J Cheminform. 2021 10;13(1):1-14. doi:10.1186/s13321-021-00558-4.

[27] Huber F, Ridder L, Verhoeven S, Spaaks J. H, Diblen F, Rogers S, et al. Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships. PLoS Comput Biol. 2021 2;17(2):e1008724. doi:10.1371/journal.pcbi.1008724.

[28] Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, et al. MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom. 2010 7;45(7):703-14. doi:10.1002/jms.1777.

[29] MassBank Europe Mass Spectral DataBase (2023, October, 15);. https://massbank.eu/MassBank/.

[30] MassBank of North America (2023, October, 15);. https://mona.fiehnlab.ucdavis.edu/downloads.

[31] Reference Spectral Libraries - GNPS Documentation (2023, October, 15);. https://ccms-ucsd.github.io/GNPSDocumentation/downloadlibraries/.

[32] NIST20: Updates to the NIST Tandem and Electron Ionization Spectral Libraries (2023, October, 15);. https://www.nist.gov/programs-projects/nist20-updates-nist-tandem-and-electron-ionization-spectral-libraries.

[33] Bajusz D, Rácz A, Héberger K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? J Cheminform. 2015 5;7(1):20. doi:10.1186/s13321-015-0069-3.

[34] Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res. 2021 1;49(D1):D1388-95. doi:10.1093/nae/gkaa971.

[35] Smith D. H, Carhart R. E, Venkataraghavan R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. J Chem Inf Comput. 1985 5;25(2):64-73. doi:10.1021/ci00046a002.

[36] FBMN Workflow - GNPS Documentation (2025, October, 27);. https://ccms-ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking/.

6

Quantifying Fragment Ambiguity in LC-HRMS/MS Spectral Databases

Authors: Viktoriia Turkina, Lapo Renai, Annemieke Petrignani, Garry Corthals, Jake W. O'Brien, Saer Samanipour

Manuscript is in preparation.

https://doi.org/10.26434/chemrxiv.15005624/v1

Abstract

LC–HRMS/MS-based non-target analysis relies heavily on MS^2 spectral matching for compound annotation. However, not all fragments contribute equally to identification confidence, and quantitative measures of fragment ambiguity remain lacking. In this study, we developed a data-driven framework to assess fragment-level ambiguity through the Fragment Ambiguity Score (FAS), which combines fragment stability and specificity within precursor-mass-defined candidate spaces.

FAS was evaluated using five experimental spectral databases (MassBank, MoNA, GNPS, NIST20, and MSnLib), comprising together 1,026,855 spectra and an *in silico* spectral library containing 256,865 spectra. The results showed that MS^2 spectra contain both diagnostic and ambiguous fragments, with diagnostic fragments representing only a subset of the total fragmentation information available for a compound. Only 14.1% of fragments in the experimental databases and 10.7% in the *in silico* library exhibited low ambiguity (FAS $\leq$ 0.5). Fragment ambiguity showed a strong correlation with database composition, with fragments appearing highly diagnostic in smaller databases often becoming substantially more ambiguous when evaluated against larger and more diverse spectral collections. Differences between experimental and *in silico* databases further demonstrated the influence of database origin and composition on fragment ambiguity estimates.

These findings demonstrate that the number of matched fragments alone is insufficient for confidence assessment and highlight the importance of complementary sources of evidence. The proposed framework provides a quantitative basis for assessing fragment ambiguity in spectral databases.

6.1 Introduction

LC-HRMS/MS-based non-target analysis (NTA) has become a central analytical strategy for the detection and identification/annotation of unknown chemicals in complex environmental, biological, and food-related samples. [1–3] Despite its widespread application and rapid methodological developments, confident identification of detected features remains one of the major challenges in NTA workflows. [4] Since MS^2 spectra similarity is considered a proxy for chemical structure similarity, tandem mass spectrometry has been extensively applied for compound annotation and structural elucidation. [5–7] Although several computational approaches for structural elucidation have recently emerged, [8–10] MS^2 spectral matching against experimentally acquired reference spectra remains one of the most reliable annotation approaches when reference spectra are available. [11] In this workflow, experimental query spectra are compared against spectral libraries, and compound identification is inferred from the similarity between fragmentation patterns. Consequently, the performance of spectral matching strongly depends on both the quality and coverage of spectral databases, as well as on the structural information content encoded in MS^2 spectra. When carefully performed, spectral library matching supports tentative structure annotation corresponding to confidence level 2a according to the Schymanski identification scale. [12]

Spectral library matching algorithms integrate several similarity features, including the number of matched fragment ions, fragment ion intensity agreement, modified cosine or alternative similarity scores, and m/z tolerances. [13–15] The number of matched fragment ions is commonly used as a key indicator of spectral-match confidence and is typically evaluated together with similarity scores when assessing reliability of library-based compound annotations. [16] However, the use of matched-fragment counts as a confidence metric for library search assumes that all matched fragments contribute equally to compound discrimination.

In practice, it has been rarely discussed that not all fragments contribute equally to confident annotation via spectral matching. While some fragments are highly characteristic and provide strong evidence for a specific structure, others may represent common neutral losses or low-mass ions that are frequently observed across many unrelated compounds. Such non-specific fragments can increase ambiguity during spectral interpreta-

tion and may contribute to false-positive annotations. [17] Current spectral matching approaches often evaluate spectra using similarity metrics evaluated on the total number of fragment ions, while largely treating all fragments as equally informative. However, the actual discriminative power of individual fragments has not been systematically assessed and quantified.

Quantifying fragment-level ambiguity requires a large collection of reference spectra that captures fragment occurrence patterns across diverse chemical space and experimental conditions. Spectral databases have therefore became an essential resource for both compound annotation and computational modeling in mass spectrometry. Beyond their primary role in spectral library matching and non-target identification workflows, large-scale spectral repositories are increasingly used to develop and train computational methods for predicting compound properties and activities, [18–22] fragmentation behavior, [23] molecular representations, [24, 25] and structure-spectrum relationships. [7, 26, 27] As the size and diversity of publicly available spectral databases continue to expand, they support data-driven algorithm development. However, despite their growing use for both annotation and modeling applications, the uniqueness and specificity of fragments across compounds within these databases have not yet been systematically investigated. Evaluating fragment-level uniqueness and incorporating this information into spectral matching workflows could provide an additional layer of confidence assessment without requiring additional experimental measurements.

In this study, we present a data-driven statistical framework for quantifying fragment-level ambiguity in MS^2 spectra. Using large curated spectral databases, both experimental and *in silico*, compounds were grouped into precursor-mass-defined candidate sets, and fragment occurrence statistics were evaluated within and across isobaric compounds. To capture both fragment detection consistency across spectra of the same compound and fragment occurrence across different compounds, we introduce the Fragment Ambiguity Score (FAS), a metric designed to estimate how diagnostic or ambiguous individual fragments are for compound identification. The proposed framework enables systematic assessment of fragment uniqueness and provides additional source of information for confidence estimation in spectral matching workflows for non-target analysis.

6.2 Experimental Section

6.2.1 Overall Workflow

Spectral databases were first curated through a series of filtering and preprocessing steps. Next, the spectra were grouped into subdatasets with the same precursor ion mass using a mass tolerance of ± 0.005 Da. That represents a set of compounds that could not be distinguished based on precursor m/z values alone, which often leads to false identifications. Fragment occurrence statistics were then calculated within each subdataset to estimate fragment reproducibility and specificity. These statistics were combined into the Fragment Ambiguity Score (FAS), which was subsequently used to assess fragment-level ambiguity across multiple spectral databases (Figure 6.1).

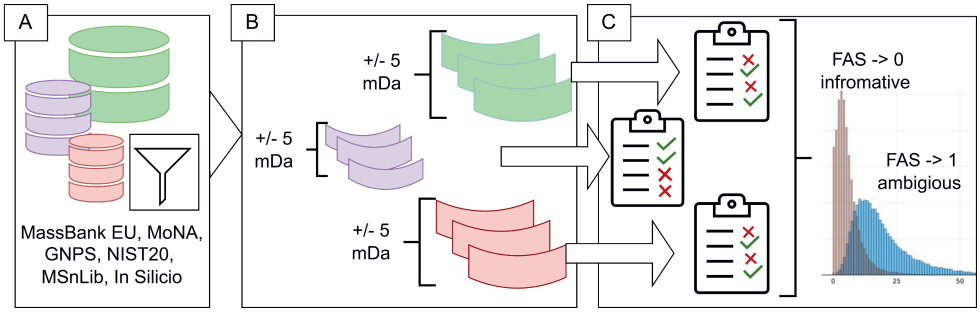


Figure 6.1: Overview of the overall workflow. Spectral databases were curated (A) and grouped into precursor-mass-defined candidate spaces (±0.005 Da) (B). Fragment occurrence statistics and FAS were calculated within each isobaric group. These metrics were used to evaluate fragment-level ambiguity (C).

6.2.2 Databases

The MS^2 spectral dataset was assembled from entries in the MSnLib, [28] MassBank EU 2025.10, [29, 30] MoNA (LC-MS/MS Positive Mode), [31] GNPS curated spectral library (GNPS-LIBRARY.msp), [32] and NIST20 [33] databases. Only experimentally acquired spectra were retained from MoNA, and only MS^2 spectra were retained from MSnLib (MS2Lib).

In addition to the experimental databases, an *in silico* spectral library was included to investigate how computationally generated fragmentation patterns influence fragment am-

biguity estimates. The *in silico* spectral library used in this study was published by Egede Frøkjær and Hansen [34] and consists of spectra generated using CFM-ID 4.4.7. [35] The *in silico* library was subjected to the same preprocessing and filtering workflow as the experimental datasets to ensure methodological consistency and comparability.

To assess the influence of database composition on fragment ambiguity, all calculations were first performed individually for each spectral database and subsequently repeated using a combined dataset containing all retained MS^2 spectra. As the final dataset was assembled from multiple public spectral repositories, duplicate spectra could not be completely excluded.

6.2.3 Databases preprocessing - clean up

The collected spectra originated from both Q-TOF and Orbitrap mass analyzers and were acquired under diverse instruments and experimental conditions, including varying collision energies. Characteristic fragments themselves are often condition-dependent, and therefore, acquisition variability was intentionally retained to evaluate fragment ambiguity under realistic conditions.

Only spectra acquired in positive electrospray ionization (ESI+) mode were included in this study. Spectra lacking structural identifiers or adduct type information were excluded. To ensure high spectral quality, only spectra acquired at a reported resolution greater than 5000 were retained. Furthermore, spectra were required to contain at least three recorded fragment ions and a precursor ion m/z value < 650 Da. Compounds above this mass are relatively rare in the investigated datasets and, therefore, often exhibit distinct fragmentation behavior. Restricting the analysis to m/z < 650 improves the representativeness of the dataset while avoiding bias from sparsely represented high-mass compounds.

Fragments with the relative intensity less than 5% were removed from the fragmentation spectra. A cumulative neutral loss (CNL) filter was subsequently applied using the combined database as a reference. CNLs were grouped using a mass tolerance of ± 0.01 Da, and the ones observed fewer than 100 times across the combined dataset were considered insufficiently frequent. Fragments associated with these low-frequency CNLs were excluded from further analysis. Finally, spectra containing an exceptionally large number of fragments ($>$95th percentile of the fragment-count distribution) were excluded to limit

the influence of outlier spectra.

To reduce variability associated with uncommon ionization products, an additional adduct filtering step was applied specifically to the NIST20 library. Adduct types representing 95% of the cumulative adduct type population were selected. Thus, only more frequently observed adduct types within the NIST20 library were retained.

For each retained spectrum, the following data were extracted and used for subsequent analysis: accession number, InChI, SMILES, InChIKey, exact mass, precursor ion m/z, adduct type, instrument type, ionization, fragmentation mode, collision energy, MS^2 m/z values, and MS^2 peak intensities. A combined experimental database was then compiled by merging all entries from the preprocessed and filtered spectral datasets. A detailed statistical analysis of the datasets is discussed in Section 6.3.1 Fragment Ambiguity Across Databases.

6.2.4 Fragment Ambiguity Score calculation

To calculate fragment ambiguity, we first generated consensus spectra by merging all spectra associated with the same InChIKey. Fragment ions within $\pm$0.01 Da were considered identical when constructing consensus spectra and recording fragment occurrences across all available spectra. The resulting fragment frequencies were then used to estimate the probability of observing a given fragment for the target compound and across competing compounds.

Compounds were grouped according to precursor ion m/z using a tolerance of $\pm$0.005 Da. Each resulting group represented candidate compounds that could not be distinguished on the basis of precursor m/z alone, reflecting a realistic identification scenario in which candidate structures must be differentiated using MS^2 fragmentation information. Different tolerances were used for precursor and fragment ions to reflect their distinct mass error characteristics. [36, 37] The stricter precursor mass tolerance ensured the narrowest possible candidate set, whereas the wider $\pm$0.01 Da fragment tolerance the lower mass accuracy of MS^2 fragment measurements, which commonly arise from the reduced resolving power used during MS^2 acquisition. This broader tolerance reduces the risk of excluding genuine fragment matches.

The detection consistency was quantified as the probability of observing fragment f for

compound c, calculated as:

$$P(TP) = \frac{N_{f,c}}{N_c} \tag{6.1}$$

where $N_{f,c}$ denotes the number of spectra of compound c in which fragment f was detected, and N_c denotes the total number of spectra available for compound c. Thus, $P(TP)$ represents the probability of observing fragment f for compound c and serves as a measure of fragment detection consistency.

Fragment specificity was evaluated using:

$$FDR = \frac{FP}{TP + FP} \tag{6.2}$$

where TP denotes occurrences of the fragment within the target compound; FP denotes occurrences of the fragment in other compounds belonging to the same precursor m/z candidate group. Fragments exhibiting high FDR values are frequently observed across multiple candidate compounds and therefore possess limited discriminative power.

To jointly account for fragment reproducibility and specificity, a Fragment Ambiguity Score (FAS) was defined as:

$$FAS = 1 - P(TP)(1 - FDR) \tag{6.3}$$

The score ranges from 0.0 to 1.0:

FAS = 0.0 indicates a highly diagnostic fragment; FAS = 1.0 indicates a highly ambiguous fragment.

Low FAS values correspond to fragments that are both reproducible for the target compound and uncommon among candidate compounds that fall within the same precursor ion m/z window.

6.2.5 Statistical analysis

To estimate the proportion of diagnostic fragments within each spectrum, fragments with a FAS of ≤ 0.5 were classified as diagnostic. This threshold was selected to ensure that defined diagnostic fragments exhibited both sufficient detection consistency and specificity. According to Equation 6.3, a fragment detected in at least 50% of the spectra associated with a given compound and exhibiting no false-positive occurrences (FDR = 0) yields a FAS value of 0.5. Consequently, fragments classified as diagnostic were required to demonstrate a minimum level of detection consistency while remaining specific to the target compound. Fragments observed less frequently were considered more ambiguous, reducing the likelihood of assigning unstable or rarely observed fragments as diagnostic. The FAS ≤ 0.5 threshold was selected for statistical comparison and does not represent a strict definition of a diagnostic fragment.

To quantify the similarity between the distributions of total fragments and diagnostic fragments within each database, a distribution overlap coefficient (OVL) was calculated. [38] The OVL was estimated from normalized histogram densities according to:

$$OVL = \sum_{i} \min(p_i, q_i) \tag{6.4}$$

where (p_i) and (q_i) denote the probabilities associated with bin (i) for the distributions of total and diagnostic fragments, respectively. The coefficient ranges from 0 to 1, where 0 indicates no overlap between the distributions and 1 indicates identical distributions. Higher OVL values, therefore, indicate that the number of diagnostic fragments closely follows the total number of detected fragments within a database.

In addition, the fraction of diagnostic fragments was calculated as the number of fragments with FAS ≤ 0.5 divided by the total number of recorded fragments in each dataset. This metric was used to compare the overall proportion of low-ambiguity fragments across databases.

6.2.6 Experimental dataset

To demonstrate the practical application of FAS, an experimental LC–HRMS/MS dataset previously reported by Renai et al. [39] was utilized. The dataset consisted of mixtures of

54 internal standards prepared from individual stock solutions (1000 mg/L) and diluted to a final concentration of 0.5 mg/L prior to analysis.

Chromatographic separation was performed using an ACQUITY UPLC system (Waters, Milford, MA, USA) coupled to a Synapt G2 high-resolution mass spectrometer (Waters) equipped with an electrospray ionization (ESI) source. Data were acquired in data-independent acquisition (DIA, MSE) mode over an m/z range of 70–1200. Analyses were performed in both positive and negative ionization modes using stepped collision energies to promote the generation of structurally informative fragments. Each mixture was analyzed in triplicate under identical instrumental conditions.

Because the present study focused on positive-ion fragmentation patterns, only spectra acquired in positive ESI mode were included in the subsequent FAS analysis. Detailed instrumental parameters and acquisition settings are provided in the original publication and summarized in the Supplementary Information.

6.2.7 Calculations

All calculations were performed on a personal computer (PC) with an Intel Core i7-1260P central processing unit and 32 GB of RAM operating Windows 10 Education version 22H2. All data processing and statistical analysis were performed using Julia language version 1.10.6. The script is publicly available at `https://doi.org/10.5281/zenodo.21035799`.

6.3 Results

6.3.1 Fragment Ambiguity Across Databases

NIST20 and MS2Lib contained considerably larger numbers of spectra and unique compounds than MassBank, MoNA, and GNPS (Table 6.1). Differences were also observed in spectral richness, measured as the number of fragments per spectrum. NIST20 exhibited the highest median fragment count (28 fragments per spectrum), compared to 17, 16, 15, and 11 fragments for MoNA, MS2Lib, MassBank, and GNPS, respectively. Furthermore, NIST20 showed substantially greater diversity in adduct types, comprising 887 distinct precursor ion annotations, whereas the remaining databases contained no more than 48 adduct types. These differences in database size, spectral richness, and adduct diversity

are expected to influence fragment occurrence statistics and, consequently, fragment ambiguity scores.

Table 6.1: Summary of the spectral databases after preprocessing and filtering.

Database	Spectra	Compounds	Median fragments (IQR, max)	Median precursor m/z	Number of adduct ion types
MassBank	61722	5536	15 (7–33, 248)	293.1 (212.1–377.2)	48
MoNA	57924	6787	17 (7–38, 308)	315.1 (231.6–409.0)	25
GNPS	6560	2784	11 (6–21, 150)	348.2 (272.2–459.1)	35
NIST20	580657	25473	28 (16–55, 374)	269.2 (209.0–349.2)	813
MS2Lib	319992	41663	16 (11–24, 112)	333.2 (307.1–385.1)	7
Combined	1026855	70129	19 (11–32, 374)	322.1 (262.1–375.2)	887
in silico	256865	83398	21 (12–32, 71)	265.1 (190.0–368.2)	1

The distribution of the number of total and diagnostic fragments per compound varied across the investigated spectral databases (Figure 6.2). The less populated databases, MassBank, MoNA, and GNPS, exhibited similar fragment-count distributions characterized by relatively low numbers of total fragments per compound. In contrast, NIST20 and MS2Lib showed considerably broader distributions and substantially larger numbers of fragments per compound.

NIST20 and MS2Lib contain substantially larger numbers of spectra and compounds, resulting in broader coverage of chemical space and fragmentation behavior. Moreover, compounds in these databases are frequently represented by multiple spectra acquired under different experimental conditions, including varying collision energies. Aggregation of these spectra increases the likelihood of observing additional fragmentation pathways and, consequently, increases the number of unique fragments associated with a compound. Finally, NIST20 and MS2Lib are highly curated reference libraries, whereas MassBank, MoNA, and GNPS primarily function as spectral repositories with different levels of

curation. Therefore, differences in acquisition protocols, curation practices, and spectral optimization across these resources are likely to contribute to the observed differences in the number of recorded fragments per compound.

MassBank, MoNA, and GNPS exhibited moderate to high overlap between the distributions of total and diagnostic fragments with OVL coefficient equal to 0.44, 0.48, and 0.90, respectively, suggesting that a considerable proportion of detected fragments were indicated as diagnostic. In contrast, NIST20 and MS2Lib showed substantially lower overlap (0.14 and 0.33, respectively). Despite containing substantially richer spectra, the distributions of diagnostic fragments in these databases were shifted toward lower fragment counts.

NIST20 and MS2Lib exhibited distributions strongly skewed toward ambiguous fragment fractions, indicating that only a minority of recorded fragments contributed substantial discriminatory information (Section 6.5 Figure S1). In contrast, GNPS, MassBank, MoNA showed a large proportion of compounds with diagnostic fragment fractions approaching unity. Overall, 5.9%, 17.6%, 24.8%, 28.3%, and 84.1% of the fragments in NIST20, MassBank, MoNA, MS2Lib, and GNPS, respectively, were classified as diagnostic (FAS ≤ 0.5) (Figure S1).

This behavior suggests that increasing database size and chemical diversity provide a more comprehensive assessment of fragment specificity. As the number of candidate compounds available for comparison increases, fragments that appear diagnostic in smaller databases are more frequently observed in unrelated compounds and therefore receive higher ambiguity scores. Consequently, larger databases reveal a greater proportion of shared and non-specific fragments, resulting in a reduced fraction of fragments classified as diagnostic.

6.3.2 Comparison with an *in silico* Spectral Database

To investigate whether fragment ambiguity differs between experimentally acquired and computationally predicted spectra, the proposed workflow was applied to both experimental spectral libraries and an *in silico* spectral database. The resulting distributions of diagnostic fragments were subsequently compared between the two data sources (Figure 6.2).

For the experimental databases, the overlap between total and diagnostic fragment

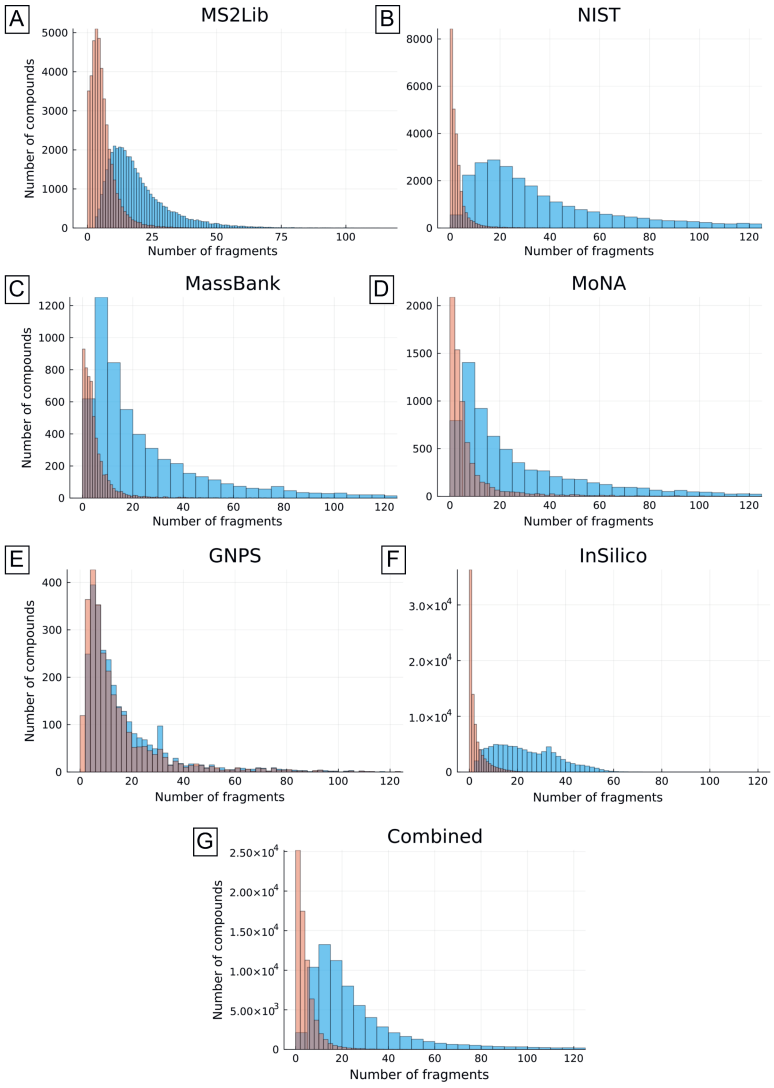


Figure 6.2: Distribution of the total number of fragments per compound (blue) and the number of diagnostic fragments per compound (FAS $\leq$0.5, orange) across the investigated spectral databases. Histograms are shown for MS2Lib (A), NIST20 (B), MassBank (C), MoNA (D), GNPS (E), the *in silico* spectral library (F), and the combined experimental database (G).

count distributions generally decreased with increasing database size and spectral coverage. In the *in silico* database, 10.7% of fragments were classified as diagnostic (FAS $\leq$ 0.5), which is intermediate between NIST (5.9%) and MS2Lib (28.3%). The overlap coefficient was 0.22, comparable to that observed for NIST20 (0.14) and MS2Lib (0.33), indicating a substantial difference between the distributions of total and diagnostic fragment counts. The distribution of diagnostic fragment fractions (Section 6.5 Figure S1) was likewise strongly skewed toward low values, meaning that for most compounds only a small proportion of the predicted fragments were classified as diagnostic, similar to the larger experimental databases.

In experimental spectral libraries, compounds are frequently represented by spectra acquired under diverse conditions, including different collision energies, instruments, and laboratories. Consequently, experimentally acquired fragments are not consistently observed across all spectra of a compound, resulting in lower fragment detection consistency. In contrast, *in silico* spectra are generated under a uniform computational framework, where fragments are predicted directly from molecular structure and retained based on chemically plausible fragmentation pathways. As a result, predicted fragments are expected to exhibit higher reproducibility across spectra. Nevertheless, despite this increased reproducibility, the *in silico* library exhibited diagnostic-fragment fractions similar to those of the larger experimental databases. This suggests that the large number and diversity of compounds represented in the database have a stronger influence on fragment ambiguity than the increased reproducibility of predicted spectra.

6.3.3 Combined Database

To evaluate the effect of database size and chemical diversity on fragment ambiguity, the proposed workflow was applied to a combined spectral database containing all curated spectra from the individual repositories. By integrating multiple experimental databases, the combined dataset provides the broadest coverage of compounds and fragmentation patterns considered in this study.

The combined dataset similarly to NIST20 and MS2Lib exhibited a reduced proportion of diagnostic fragments (FAS $\leq$ 0.5) with overlap coefficient 0.24 and distributions strongly skewed toward a high ambiguity-fragment fraction. Overall, 14.1% of fragments in the combined database exhibited low ambiguity (FAS $\leq$ 0.5). Again, this trend reflects the

increased likelihood of observing the same fragment across a larger number of candidate compounds, resulting in higher fragment ambiguity scores. Consequently, fragments that appeared highly specific within smaller databases were frequently reclassified as less diagnostic when evaluated against the broader chemical space represented in the combined database. Therefore, the combined database provides a more comprehensive assessment of fragment ambiguity because fragments are evaluated against a larger and more chemically diverse set of candidate compounds.

The influence of database composition on fragment ambiguity is illustrated by the examples of Hippuric acid (InChiKey: QIAFMBKCNZACKA-UHFFFAOYSA-N) (Figure 6.3) and Lumichrome (InChiKey: ZJTJUVIJVLLGSP-UHFFFAOYSA-N) (Section 6.5 Figure S2). First, the consensus spectra generated from different databases did not necessarily contain the same set of fragments, reflecting differences in spectral coverage and database content. Consequently, both the number of diagnostic fragments (FAS $\leq$ 0.5) and the minimum FAS values observed for a compound varied between databases (Section 6.5 Figure S2).

For Hippuric acid, only the GNPS database contained two fragments, m/z 105.04 and 106.00, with FAS values equal to 0.0 (i.e. fully diagnostic). This indicates that these fragments were repeatedly observed for Hippuric acid but were not detected in any competing isobaric compounds within the GNPS dataset. However, when evaluated using the combined database, the fragment at m/z 105.04 remained the most diagnostic fragment but exhibited a higher ambiguity score (FAS = 0.40) than within GNPS database, while the neighboring fragment at m/z 106.00 was assigned a substantially higher ambiguity score (FAS = 0.96). The same m/z 105.04 fragment received a considerably higher ambiguity score in the *in silico* database (FAS = 0.89). This increase in the ambiguity is linked to the relative number of spectra of Hippuric acid (TP) compared with the number of competing candidate structures within the corresponding precursor-mass window in each database (FP). The combined spectral database contain 32 TPs against 21 of FPs, whereas in *in silico* database contains a single spectrum of TP and 8 FPs. A higher number of competing candidates per Hippuric acid spectrum leads to a higher FAS.

A similar trend was observed for Lumichrome. Fragments classified as diagnostic in GNPS and the *in silico* database frequently exhibited substantially higher ambiguity scores in NIST and the combined database. As a result, the ranking of fragments according to their diagnostic value changed considerably between databases, and different fragments were selected as the most diagnostic for the same compound.

Together, these examples demonstrate that fragment ambiguity is strongly dependent on database composition and chemical coverage. As the number and diversity of candidate compounds increase, fragments are evaluated against a broader chemical space, revealing fragment sharing that may remain undetected in smaller databases. Consequently, fragments that appear highly diagnostic in individual repositories may become substantially less diagnostic when assessed using a larger and more diverse database. The combined database, therefore provides a more comprehensive assessment of fragment ambiguity and highlights information that may not be apparent when individual repositories are considered in isolation.

6.3.4 Application

Suspect screening is one of the most widely applied approaches for candidate annotation in LC-HRMS/MS NTA. To evaluate the practical utility of FAS, the proposed framework was applied to the dataset reported by Renai et al. [39] Fragment-level ambiguity was assessed using the combined spectral database generated in this study (Figure 6.4).

A total of 30 internal standards were detected in positive ionization mode. However, 12 of these compounds were not represented in the curated database and were therefore added to enable assessment of their fragment ambiguity. Subsequently, FAS values were calculated for all matched fragments associated with the detected compounds.

The analysis revealed that most candidate annotations were supported primarily by commonly observed fragments exhibiting relatively high ambiguity scores. For the majority of compounds (19 out of 30) zero or only one matched fragment was classified as diagnostic. Only 11 compounds were associated with more than one diagnostic fragments. These results suggest that successful suspect screening annotations frequently rely on fragments that are widely shared among structurally unrelated compounds and therefore provide limited discriminatory power when considered individually.

The findings further demonstrate that the number of matched fragments alone may not adequately reflect identification confidence. As the application study was performed using standards dissolved in solvent, it represents a near-ideal scenario for suspect screening with minimal background interference. Fragment ambiguity is expected to increase in complex samples containing co-eluting compounds and matrix-derived ions, further

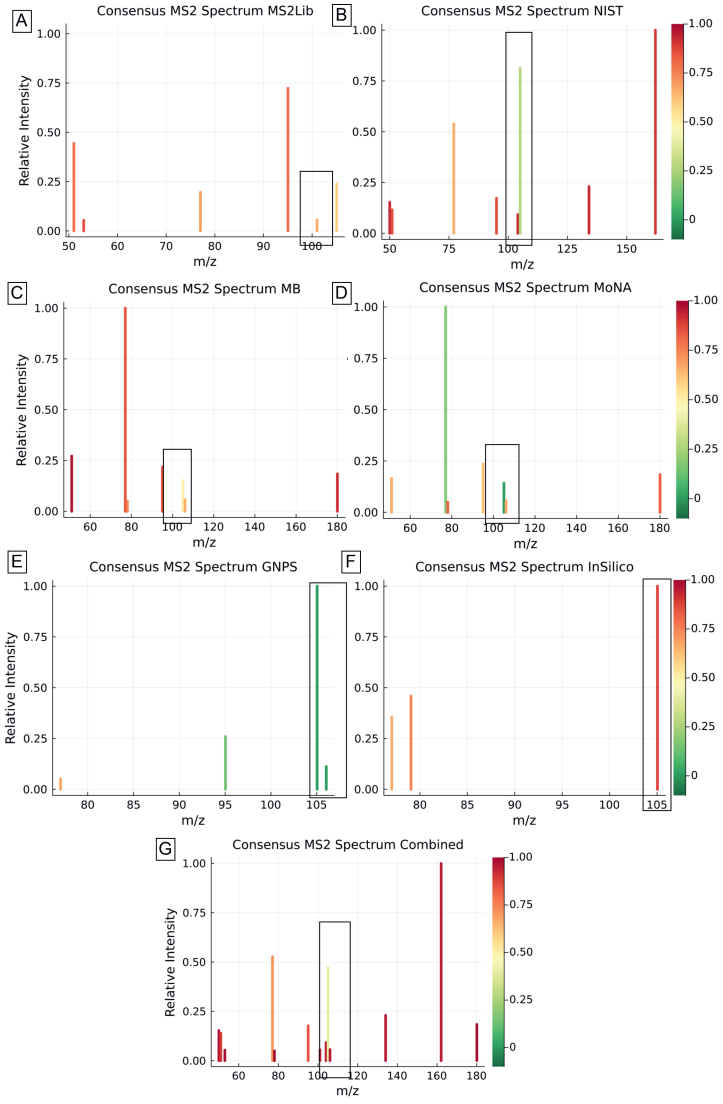


Figure 6.3: Consensus MS^2 spectra of Hippuric acid (InChiKey: QIAFMBKCNZACKA-UHFFFAOYSA-N) generated form MS2Lib (A), NIST20 (B), MassBank (C), MoNA, (D) GNPS (E), *in silico* spectral library (F), and the combined experimental database (G). Fragment colors represent FAS values, ranging from diagnostic (green) to high ambiguous (red). The grey boxes highlight the fragment at m/z 105.04, which was consistently assigned a lower FAS value and therefore classified as less ambiguous across experimental databases.

highlighting the value of incorporating FAS to assess the diagnostic value of individual spectral matches and support more informed confidence assessment in suspect screening workflows.

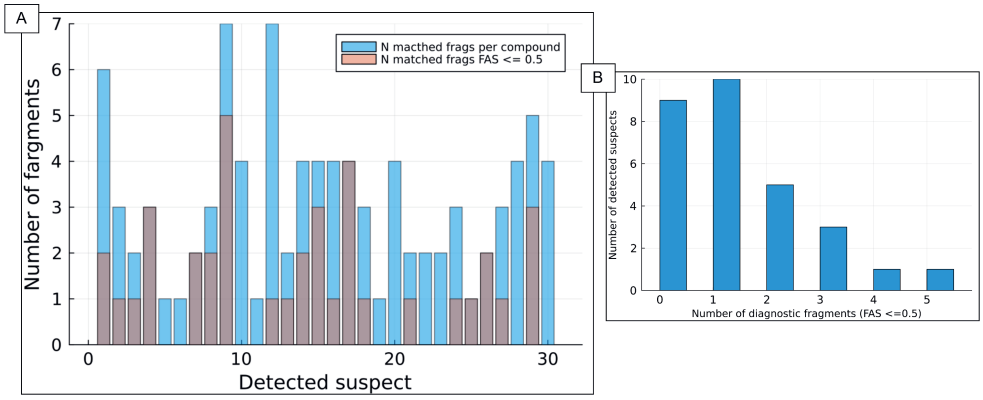


Figure 6.4: Application of FAS to suspect screening results. (A) Number of matched fragments for each detected suspect (blue) and the subset of matched fragments classified as diagnostic (FAS $\leq$0.5, orange). (B) Distribution of detected suspects according to the number of diagnostic matched fragments.

6.3.5 Potential Applications of Fragment Ambiguity Scoring

The proposed Fragment Ambiguity Score provides a quantitative measure of fragment-level ambiguity and therefore has several potential applications in LC–HRMS/MS annotation workflows.

First, FAS could be incorporated into spectral matching algorithms by weighting matched fragments according to their diagnostic value. Under such an approach, matches supported by highly diagnostic fragments would contribute more strongly to identification confidence than matches based primarily on commonly observed fragments.

Figure 6.5 illustrates this concept using a simplified example. A generated spectra of Trimethoprim (InChiKey: IEDVJHCEMCRBQM-UHFFFAOYSA-N) was matched against reference spectra of two candidates with identical precursor ion masses (m/z 291.146): Trimethoprim and 1-[1-(Cyclohex-3-ene-1-carbonyl)pyrrolidin-3-yl]triazole-4-carboxylic acid (InChiKey: HGRZJPNBQQWSAA-UHFFFAOYSA-N). Based solely on the number of matched fragments, the false candidate would appear more plausible because four fragments are matched compared with only three for trimethoprim. However, all four

matched fragments supporting the false candidate exhibit high ambiguity scores (FAS $>$ 0.5), indicating that they are commonly observed and provide limited discriminatory power. In contrast, the trimethoprim match contains a fragment with a substantially lower FAS value, indicating a higher degree of specificity. Consequently, consideration of fragment ambiguity favors the correct candidate despite the lower number of matched fragments.

More generally, identification confidence could be assessed not only by the number of matched fragments but also by the number and proportion of diagnostic fragments supporting a candidate structure. Multiple low-FAS fragments observed simultaneously may provide stronger evidence than a larger number of highly ambiguous fragments. Such information could support confidence assessment in suspect screening and library-matching workflows. To facilitate this, a fragment-level FAS database is generated in this study and is made publicly available at `https://doi.org/10.5281/zenodo.21035799`. Such a database provides a resource for evaluating fragment ambiguity within the proposed annotation framework (Section 6.5 Figure S3).

Second, FAS may assist method development and optimization of acquisition conditions. By identifying which fragments provide the greatest discriminatory power, fragmentation conditions such as collision energy as well as the adducts may be optimized to maximize the production of diagnostic fragments and improve compound characterization.

Finally, FAS may support spectrum interpretation and database curation by identifying fragments that contribute little discriminatory power across large spectral collections. Currently, fragment selection is guided primarily by fragment intensity, as the most abundant product ions are typically assumed to provide the most useful information. FAS provides a complementary criterion by quantifying the diagnostic value of individual fragments. Combining fragment intensity with FAS could therefore guide the selection of fragments that are not only abundant but also highly specific to the target compound. However, it should be noted that the present framework evaluates fragments individually. Meaning that fragments with high ambiguity scores may still contribute valuable structural information when interpreted in combination with other fragments. Future work should therefore investigate the ambiguity of fragment combinations and fragmentation motifs in addition to individual product ions.

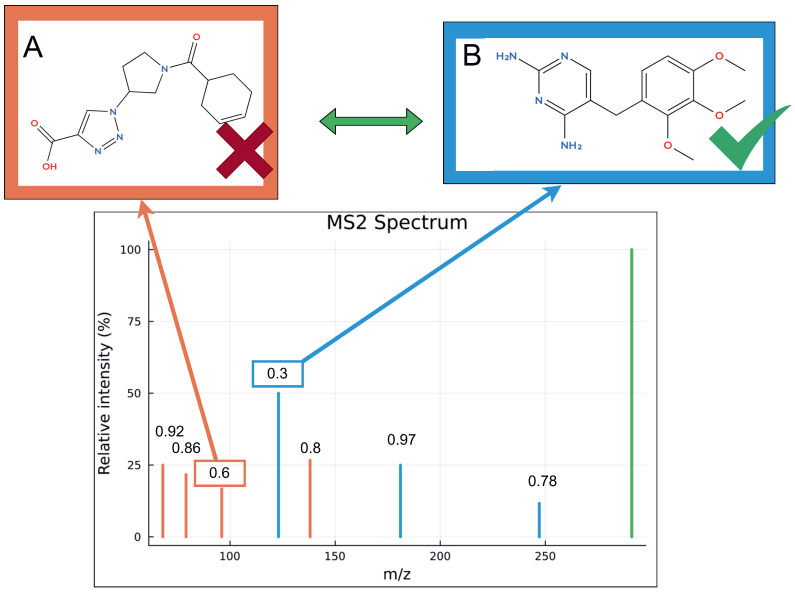


Figure 6.5: Example application of FAS in spectral matching. Candidate B is favored despite fewer matched fragments due to the presence of a low-FAS (more diagnostic) fragment.

6.4 Conclusions

The proposed framework provides a quantitative approach for assessing fragment ambiguity and has potential applications in spectral matching and confidence assessment workflows. By distinguishing diagnostic fragments from commonly occurring ones, FAS may support more transparent interpretation of spectral matches and help prioritize spectral evidence during compound annotation.

Analysis of multiple experimental spectral databases demonstrated that MS^2 spectra contain a mixture of highly diagnostic and non-specific fragments. While diagnostic fragments were observed for the majority of compounds, they represented only a subset of the total fragmentation information available. Consequently, the number of matched fragments alone is insufficient to assess identification confidence, as fragments differ substantially in their ability to discriminate between candidate structures.

Several limitations should be acknowledged. The proposed approach remains dependent on the quality, completeness, and representativeness of available spectral databases. Furthermore, the present study focused exclusively on positive electrospray ionization

mode. In addition, the current framework evaluates fragments individually. While highly ambiguous fragments generally contribute limited discriminatory power on their own, combinations of fragments may collectively provide strong structural evidence. Therefore, fragment-level ambiguity should not be interpreted independently from the overall fragmentation pattern.

Future work should extend the framework to negative ionization mode and explore approaches for incorporating fragment combinations into ambiguity scoring. Integration of fragment-level ambiguity into spectral matching algorithms to further improve the interpretation of MS^2 evidence in non-target analysis.

The proposed framework has direct implications for spectral matching workflows. Fragment-level weighting based on their diagnostic value may improve confidence estimation, reduce ambiguity among competing candidates, and support more transparent interpretation of spectral matching results. More broadly, the findings suggest that confident compound annotation should not rely solely on MS^2 spectral similarity but should incorporate complementary sources of evidence, including retention behavior, structural metadata, and orthogonal analytical information.

Overall, this study demonstrates that fragment ambiguity is a measurable property of MS^2 data and that not all matched fragments contribute equally to compound identification. These findings emphasize that confident annotation should not rely solely on the number of matched fragments, but should also consider the diagnostic value of individual fragments together with complementary sources of evidence.

6.5 Supporting Information

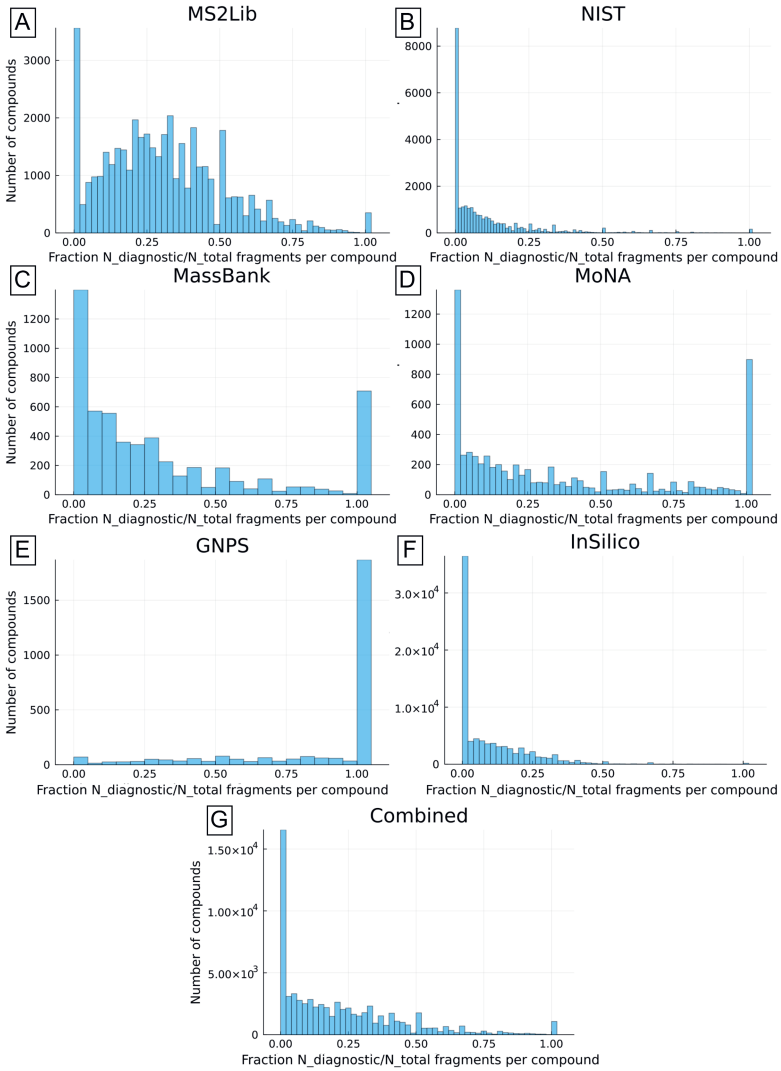


Figure S1: Distribution of the fraction of diagnostic fragments per compound across the investigated spectral databases: MS2Lib (A), NIST20 (B), MassBank (C), MoNA (D), GNPS (E), *in silico* spectral library (F), and the combined experimental database (G). Diagnostic fragments were defined as fragments with FAS $\leq$ 0.5. The fraction was calculated as the ratio of diagnostic fragments to the total number of fragments recorded per compound.

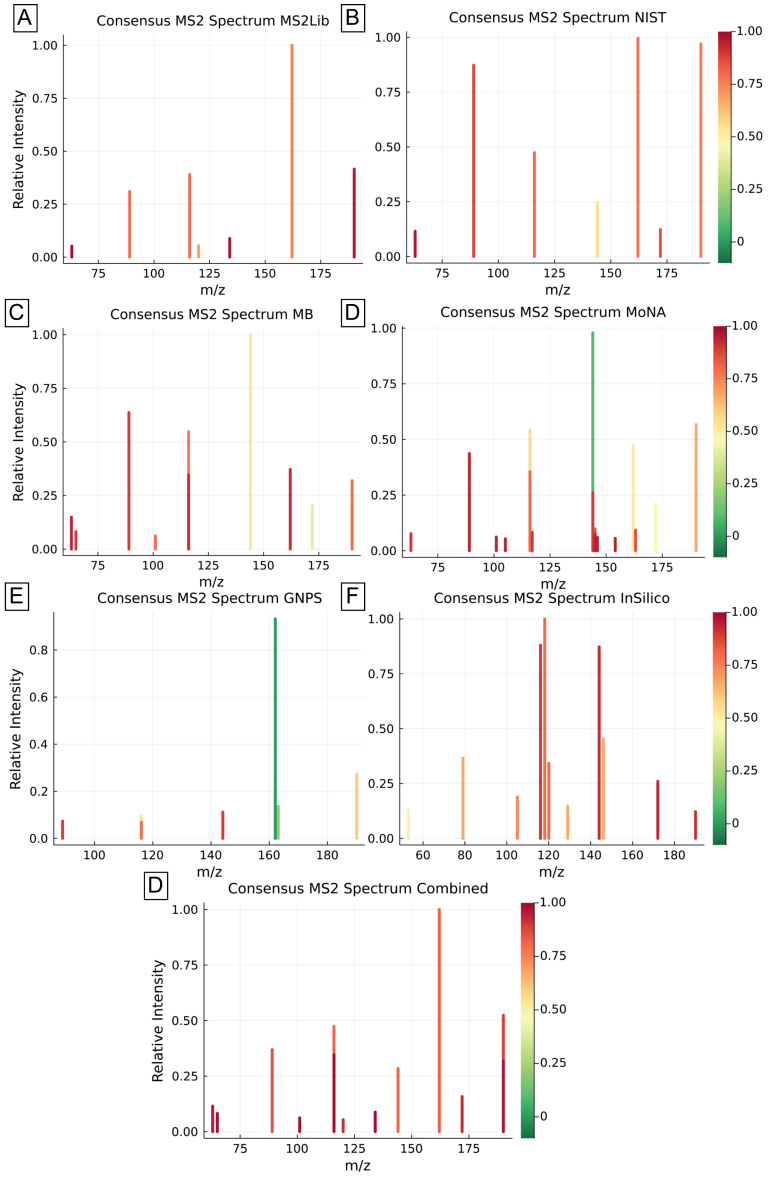


Figure S2: Consensus MS^2 spectra of Lumichrome (InChiKey: ZJTJUVIJVLLGSP-UHFFFAOYSA-N) generated form MS2Lib (A), NIST20 (B), MassBank (C), MoNA (D), GNPS (E), *in silico* spectral library (F), and the combined experimental database (G). Fragment colors represent FAS values, ranging from more diagnostic (green) to high ambiguity (red).

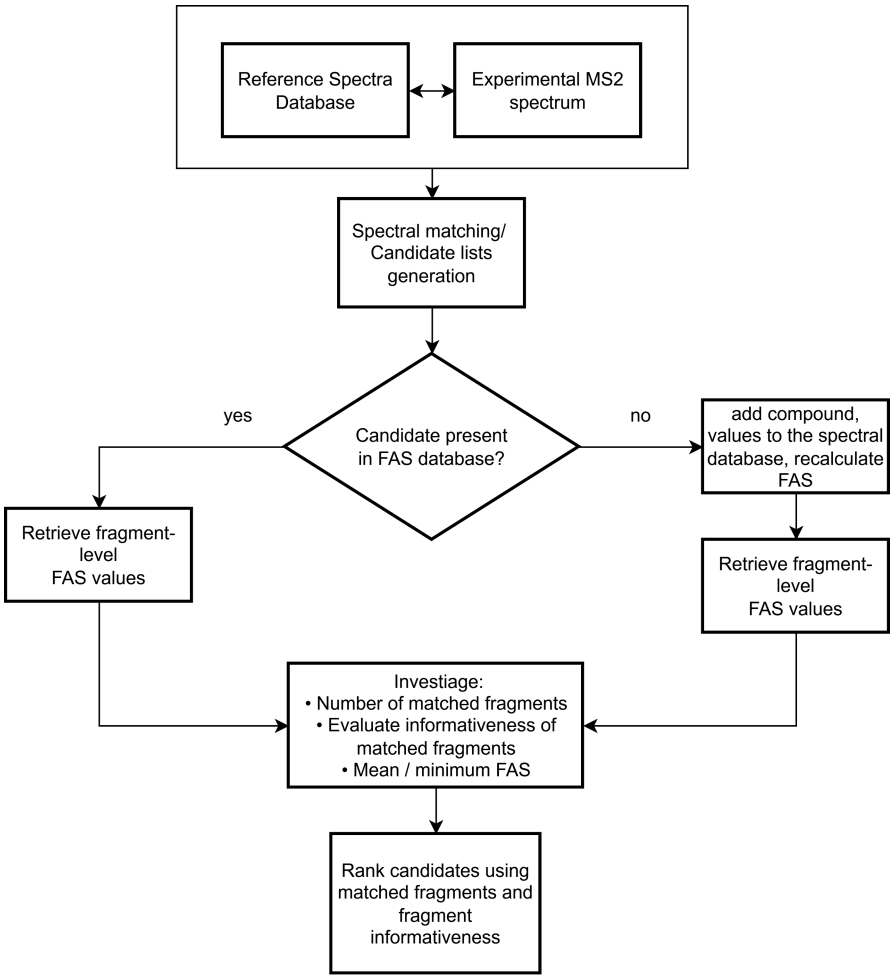


Figure S3: Proposed workflow for integrating the Fragment Ambiguity Score (FAS) database into spectral matching workflows. Following candidate generation, fragment-level FAS values are retrieved and used to evaluate the number and proportion of diagnostic fragments supporting each candidate. This information can be incorporated into confidence assessment and candidate prioritization. New compounds can be added to the database, enabling recalculation of FAS values as spectral coverage expands.

Notes

The code used to perform all calculations and the processed spectral databases generated in this study are publicly available at `https://doi.org/10.5281/zenodo.21035799`. Due to licensing restrictions, the NIST20 database is not included in the repository.

Acknowledgement

The authors express their gratitude to all the members of Environmental Modeling & Computational Mass Spectrometry (www.emcms.info). S.S., A.P., and V.T. thank the ChemistryNL TKI and the UvA Data Science Centre for their funding support (projects EDIFIED and SCOPE). J.W.O. is the recipient of a National Health and Medical Research Council (NHMRC) Investigator Grant (EL12009209) funded by the Australian Government. J.W.O. also acknowledges financial support from the Australian National Health and Medical Research Council (NHMRC; APP1185347). The Queensland Alliance for Environmental Health Sciences (QAEHS), The University of Queensland (UQ), gratefully acknowledges the financial support of Queensland Health.

6.6 References

[1] Manz K. E, Feerick A, Braun J. M, Feng Y. L, Hall A, Koelmel J, et al. Non-targeted analysis (NTA) and suspect screening analysis (SSA): a review of examining the chemical exposome. J Expo Sci Environ Epidemiol. 2023 6;33(4):524-36. doi:10.1038/s41370-023-00574-6.

[2] Vermeulen R, Schymanski E. L, Barabási A. L, Miller G. W. The exposome and health: where chemistry meets biology. Science. 2020 1;367(6476):392. doi:10.1126/science.aay3164.

[3] Lai Y, Koelmel J. P, Walker D. I, Price E. J, Papazian S, Manz K. E, et al. High-Resolution Mass Spectrometry for Human Exposomics: Expanding Chemical Space Coverage. Environ Sci Technol. 2024 7;58(29):12784-822. doi:10.1021/acs.est.4C01156.

[4] Hollender J, Schymanski E. L, Ahrens L, Alygizakis N, Béen F, Bijlsma L, et al. NORMAN guidance on suspect and non-target screening in environmental monitoring. Environ Sci Eur. 2023 9;35(1):75. doi:10.1186/s12302-023-00779-4.

[5] Aron A. T, Gentry E. C, McPhail K. L, Nothias L. F, Nothias-Esposito M, Bouslimani A, et al. Reproducible molecular networking of untargeted mass spectrometry data using GNPS. Nat Protoc. 2020 5;15(6):1954-91. doi:10.1038/s41596-020-0317-5.

[6] Schollée J. E, Schymanski E. L, Stravs M. A, Gulde R, Thomaidis N. S, Hollender J. Similarity of High-Resolution Tandem Mass Spectrometry Spectra of Structurally Related Micropollutants and Transformation Products. J Am Soc Mass Spectrom. 2017 12;28(12):2692-704. doi:10.1007/s13361-017-1797-6.

[7] Huber F, van der Burg S, van der Hooft J. J. J, Ridder L. MS2DeepScore: a novel deep learning similarity measure to compare tandem mass spectra. J Cheminform. 2021 10;13(1):1-14. doi:10.1186/s13321-021-00558-4.

[8] Vitale G. A, Xia S. N, Dührkop K, Zare Shahneh M. R, Brötz-Oesterhelt H, Mast Y, et al. Enhancing tandem mass spectrometry-based metabolite annotation with online chemical labeling. Nat Commun. 2025 7;16(1):6911. doi:10.1038/s41467-025-61240-z.

[9] Hupatz H, Rahu I, Wang W. C, Peets P, Palm E. H, Kruve A. Critical review on in silico methods for structural annotation of chemicals detected with LC/HRMS non-targeted screening. Anal Bioanal Chem. 2024 8;417(3):473-93. doi:10.1007/s00216-024-05471-x.

[10] Rosina L, Ortega T, Dietrich J, Wandy J, Mol H, Van Der Hooft J. J. J. Large-scale discovery and annotation of hidden substructure patterns in mass spectrometry profiles. bioRxiv. 2025 7. doi:10.1101/2025.06.19.659491.

[11] Bittremieux W, Schmid R, Huber F, Van Der Hooft J. J. J, Wang M, Dorrestein P. C. Comparison of Cosine, Modified Cosine, and Neutral Loss Based Spectrum Alignment For Dis-

covery of Structurally Related Molecules. J Am Soc Mass Spectrom. 2022 9;33(9):1733-44. doi:10.1021/jasms.2C00153.

[12] Schymanski E. L, Jeon J, Gulde R, Fenner K, Ruff M, Singer H. P, et al. Identifying small molecules via high resolution mass spectrometry: Communicating confidence. Environ Sci Technol. 2014 2;48:2097-8. doi:10.1021/es5002105.

[13] Samanipour S, Reid M. J, Bæk K, Thomas K. V. Combining a Deconvolution and a Universal Library Search Algorithm for the Nontarget Analysis of Data-Independent Acquisition Mode Liquid Chromatography-High-Resolution Mass Spectrometry Results. Environ Sci Technol. 2018 4;52(8):4694-701. doi:10.1021/acs.est.8b00259.

[14] Rutz A, Bittremieux W, Schmid R, Cailloux O, van der Hooft J. J. J, Beniddir M. A. Scalability of mass spectrometry-based metabolomics for natural extracts libraries exploration: current status, challenges, and opportunities. Nat Prod Rep. 2025. doi:10.1039/d5np00034c.

[15] Spectral library search - mzmine documentation;. `https://mzmine.github.io/mzmine_documentation/module_docs/id_spectral_library_search/spectral_library_search.html`.

[16] Scheubert K, Hufsky F, Petras D, Wang M, Nothias L. F, Dührkop K, et al. Significance estimation for large scale metabolomics annotations by spectral matching. Nat Commun. 2017 11;8(1):1494. doi:10.1038/s41467-017-01318-5.

[17] Stein S, Gates R. Mass Spectral Reference Libraries: An Ever-Expanding Resource for Chemical Identification. Anal Chem. 2012 9;84(17):7274-82. doi:10.1021/ac301205z.

[18] Peets P, Wang W. C, Macleod M, Breitholtz M, Martin J. W, Kruve A. MS2Tox Machine Learning Tool for Predicting the Ecotoxicity of Unidentified Chemicals in Water by Nontarget LC-HRMS. Environ Sci Technol. 2022 11;56(22):15508-17. doi:10.1021/acs.est.2c02536.

[19] Boelrijk J, van Herwerden D, Ensing B, Forré P, Samanipour S. Predicting RP-LC retention indices of structurally unknown chemicals from mass spectrometry data. J Cheminform. 2023 2;15(1):28. doi:10.1186/s13321-023-00699-8.

[20] Ngan H. L, Turkina V, van Herwerden D, Yan H, Cai Z, Samanipour S. Machine Learning for Enhanced Identification Probability in RPLC/HRMS Nontargeted Workflows. Anal Chem. 2025 8;97(33):18028-35. doi:10.1021/acs.analchem.5c01873.

[21] Nikolopoulos A, van Herwerden D, Turkina V, Kruve A, Baerenfaenger M, Samanipour S. Ionization efficiency prediction of electrospray ionization mass spectrometry analytes based on molecular fingerprints and cumulative neutral losses. J Cheminform. 2025 12;17(1):183. doi:10.1186/s13321-025-01129-7.

[22] Kreutzer Y, Rahu I, Norinder U, Kruve A. Molecular networking, conformal predictions and revised fingerprint-based models for discovering endocrine disruptors in mixtures. Anal Bioanal Chem. 2026 1;418(5):1445-57. doi:10.1007/s00216-025-06303-2.

[23] Stravs M. A, Dührkop K, Böcker S, Zamboni N. MSNovelist: de novo structure generation from mass spectra. Nat Methods. 2022 7;19(7):865. doi:10.1038/s41592-022-01486-3.

[24] Dührkop K, Fleischauer M, Ludwig M, Aksenov A. A, Melnik A. V, Meusel M, et al. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. Nat Methods. 2019 3;16(4):299-302. doi:10.1038/s41592-019-0344-8.

[25] Bushuiev R, Bushuiev A, Samusevich R, Brungs C, Sivic J, Pluskal T. Self-supervised learning of molecular representations from millions of tandem mass spectra using DreaMS. Nat Biotechnol. 2025 5;44(4):630-40. doi:10.1038/s41587-025-02663-3.

[26] Huber F, Ridder L, Verhoeven S, Spaaks J. H, Diblen F, Rogers S, et al. Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships. PLoS Comput Biol. 2021 2;17(2):e1008724. doi:10.1371/journal.pcbi.1008724.

[27] de Jonge N. F, Chekmeneva E, Schmid R, Joas D, Truong L. J, van der Hooft J. J. J, et al. Cross ionization mode chemical similarity prediction between tandem mass spectra in metabolomics. Nat Commun. 2026 2;17(1):2483. doi:10.1038/s41467-026-69083-y.

[28] Brungs C, Schmid R, Heuckeroth S, Mazumdar A, Drexler M, Šácha P, et al. MSnLib: efficient generation of open multi-stage fragmentation mass spectral libraries. Nat Methods. 2025 9;22(10):2028-31. doi:10.1038/s41592-025-02813-0.

[29] Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, et al. MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom. 2010 7;45(7):703-14. doi:10.1002/jms.1777.

[30] MassBank Europe 2025.10 (2026, February, 11);. `https://github.com/MassBank/MassBank-data/tree/2025.10`.

[31] MassBank of North America (2026, February, 19);. `https://mona.fiehnlab.ucdavis.edu/downloads`.

[32] Downloading Reference Spectral Libraries - GNPS Documentation (2026, February, 11);. `https://external.gnps2.org/gnpslibrary`.

[33] NIST 20 MS/MS Library (2023, October, 15);. `https://www.sisweb.com/software/nist-msms.htm#2`.

[34] Egede Frøkjær E, Hansen M. Large-scale generation of in silico based spectral libraries to annotate dark chemical space features in non-target analysis. Anal Bioanal Chem. 2025 9;417(27):6107-19. doi:10.1007/s00216-025-06034-4.

[35] Wang F, Allen D, Tian S, Oler E, Gautam V, Greiner R, et al. CFM-ID 4.0 – a web server for accurate MS-based metabolite identification. Nucleic Acids Res. 2022 7;50(W1):W165-74. doi:10.1093/nar/gkac383.

[36] Kind T, Tsugawa H, Cajka T, Ma Y, Lai Z, Mehta S. S, et al. Identification of small molecules using accurate mass MS/MS search. Mass Spectrom Rev. 2018 7;37(4):513-32. doi:10.1002/mas.21535.

[37] Wilding-McBride D, Dagley L. F, Spall S. K, Infusini G, Webb A. I. Simplifying MS1 and MS2 spectra to achieve lower mass error, more dynamic range, and higher peptide identification confidence on the Bruker timsTOF Pro. PLoS ONE. 2022 7;17(7):e0271025. doi:10.1371/journal.pone.0271025.

[38] Clemons T. E, Bradley E. L. A nonparametric measure of the overlapping coefficient. Comput Stat Data Anal. 2000 7;34(1):51-61. doi:10.1016/s0167-9473(99)00074-2.

[39] Renai L, Turkina V, Chojnacka A, Gargano A. F. G, Samanipour S. Measurable Feature Prediction for Estimating Chemical Space Coverage in LC–ESI–HRMS Nontargeted Analysis. Anal Chem. 2026 3;98(10):7637-43. doi:10.1021/acs.analchem.5C07705.

7

Conclusion and outlook

The central objective of this thesis was to improve the understanding of how chemical information can be extracted and utilized within LC-HRMS-based non-target analysis (NTA) workflows. This was approached from multiple perspectives, ranging from the extent to which current workflows capture environmentally relevant chemical space, to strategies for extracting information from unknown features, and finally to the assumptions and limitations underlying MS^2-based compound annotation. Although these topics are often investigated independently, the results presented throughout this thesis demonstrate that they are closely interconnected and collectively determine the extent to which unknown chemicals can be characterized in environmental samples.

A first major finding of this work is that the identified chemical space remains only a small fraction of the environmentally relevant chemical universe. The meta-analysis presented in ***Chapter 2*** showed that published LC-HRMS-based NTA studies collectively covered only a limited proportion of the estimated chemical space. Furthermore, no clear relationship was observed between individual experimental configurations and achieved chemical space coverage, largely because most NTA studies employ broad and generic workflows designed to maximize overall applicability rather than targeting specific regions of chemical space. These findings highlight that current NTA approaches are still

far from providing comprehensive coverage of the exposome and that both analytical and computational developments are required to broaden the range of measurable and identifiable chemicals.

The lack of harmonized reporting across published NTA studies remains a major obstacle for systematic comparison of analytical workflows. Experimental conditions, data-processing workflows, compound identifications, and associated metadata are often reported inconsistently, hindering systematic evaluation of the relationship between methodological choices and the resulting identified chemical space coverage. Standardized reporting practices, together with broader sharing of experimental data, are therefore essential for improving reproducibility, facilitating cross-study comparisons, and advancing our understanding of how analytical strategies influence the observable chemical space.

However, increasing the number of detectable chemicals alone does not guarantee a better understanding of the exposome. Once compounds are detected, meaningful information must still be extracted from the available data. The results of this thesis demonstrate that the success of such information extraction, particularly through predictive modeling approaches, depends strongly on how chemical information is represented. ***Chapter 3*** showed that molecular representation is not a trivial data preprocessing step but a determining factor in predictive model performance. Combining complementary molecular fingerprints through feature-level fusion consistently improved predictive accuracy across multiple regression and classification tasks. This finding suggests that no single representation captures all relevant aspects of molecular structure and that complementary structural descriptions can provide a more complete representation of chemical information. More broadly, these results emphasize that the success of downstream prediction and annotation workflows depends strongly on the information retained during the representation stage.

While the investigated fingerprint fusion strategies consistently improved predictive performance, the employed molecular representations capture only selected aspects of chemical structure and do not explicitly encode features such as stereochemistry or three-dimensional molecular characteristics. The focus on interpretable non-hashed fingerprints enabled systematic evaluation of complementary structural information, but future work should also explore whether similar gains can be achieved using learned representations derived from molecular graphs, spectra, or multimodal data sources.

While molecular fingerprints describe chemicals from a structural perspective, LC–HRMS measurements provide an experimental representation through chromatographic and fragmentation behavior. The results of ***Chapter 4*** demonstrated that useful information can be extracted directly from these experimental representations without requiring explicit compound identification. By linking chromatographic and fragmentation information to toxicity categories, unknown features could be prioritized based on their potential environmental relevance. The comparable performance of the developed models to structure-based approaches suggests that LC-HRMS data contain substantial information about chemical properties that remains underutilized in conventional NTA workflows. This finding is particularly important because the majority of detected features remain unidentified, highlighting the possibility of obtaining meaningful risk-relevant information before structural elucidation is achieved.

The developed models were constrained by the availability of toxicity data, retention information, and high-quality MS^2 spectra, and were therefore developed for RPLC-ESI(+)-HRMS measurements and acute toxicity at a single trophic level. Extending these approaches to additional analytical conditions, such as negative ionization mode, will require substantially larger and more representative datasets, as the currently available data are considerably more limited and may introduce biases in model development and evaluation. Future work should therefore focus on expanding the applicability domain of these approaches through the incorporation of broader chemical diversity, additional analytical conditions, and a wider range of environmental and toxicological endpoints, including persistence, bioaccumulation, transformation behavior, and exposure relevance. Such developments would not only improve model generalizability but also further advance identification-independent characterization of unknown features, enabling more effective prioritization of chemicals before structural elucidation is achieved.

The success of these approaches relies on the assumption that chromatographic and fragmentation measurements contain sufficient information to characterize chemical properties. This naturally raises the broader question of how much structural information is encoded in MS^2 spectra and how reliably that information can be interpreted. The investigations presented in ***Chapters 5*** and ***6*** further explored the information content of MS^2 spectra, which constitute one of the most important data sources for compound annotation in NTA. The results demonstrate that the relationship between molecular structure and fragmentation spectra is considerably more complex than often assumed. Although

spectral similarity is widely used as a proxy for structural similarity, the evaluation of twenty similarity metrics revealed limited ability to consistently resolve structurally distinct compounds using MS^2 information alone. Even under optimized conditions, most similarity metrics failed to generate highly specific local molecular networks. These findings indicate that spectral similarity should not be interpreted as direct evidence of structural identity and support the growing consensus that reliable annotation requires the integration of multiple orthogonal sources of information.

The evaluated spectra originated from multiple public libraries and encompassed substantial experimental heterogeneity, including differences in instrumentation and acquisition conditions. The inclusion of spectra from multiple public databases and diverse acquisition conditions enabled the evaluation of similarity metrics and threshold behavior under realistic NTA conditions, where spectral heterogeneity is unavoidable. Future research in this area will require more comprehensive and harmonized spectral databases, including standardized reporting of acquisition parameters and associated metadata. Such resources would enable systematic investigation of how experimental conditions influence spectral similarity and facilitate the development of more robust annotation frameworks. In parallel, future research should focus on integrating spectral similarity with complementary sources of evidence to improve compound discrimination and confidence estimation.

This conclusion is further reinforced by the fragment-level analysis presented in ***Chapter 6***. Not all fragments contribute equally to compound discrimination. Fragments frequently considered characteristic in small spectral collections became substantially more ambiguous when evaluated against larger and more chemically diverse databases. Consequently, the number of matched fragments alone cannot be considered a reliable measure of annotation confidence. Instead, confidence assessment should account for both fragment specificity and the underlying search space in which spectral matching is performed.

Future research should extend the concept of fragment ambiguity beyond individual fragments to combinations of fragments and entire fragmentation patterns. Integrating fragment-level ambiguity into spectral matching and confidence estimation frameworks will enable more transparent interpretation of MS^2 evidence and improve annotation reliability. However, achieving a deeper understanding of the diagnostic value of MS^2 fragments will require more comprehensive spectral databases together with harmonized re-

porting of experimental metadata. Such resources would facilitate systematic evaluation of fragment specificity and may ultimately guide experimental design by identifying acquisition conditions and fragmentation patterns that provide the greatest structural discrimination

Collectively, the results of this thesis suggest that the future development of NTA lies not only in expanding detectable exposome chemical space, but also in extracting more information from the data already available. While modern mass spectrometers can generate enormous amounts of data, the ability to convert these measurements into reliable chemical knowledge remains limited. Expanding chemical space coverage, developing more informative molecular and spectral representations, prioritizing unknown features before identification, and evaluating the reliability of MS^2-based evidence are complementary strategies that contribute to a more comprehensive characterization of environmental chemical exposure. Ultimately, advancing NTA will depend on improving our ability to extract, interpret, and evaluate chemical information throughout the analytical workflow, rather than focusing solely on compound identification.

Summary

The continuous growth in the production and use of chemicals has resulted in increasingly complex chemical mixtures in environmental and biological samples. In addition to intentionally manufactured substances, numerous transformation products, metabolites, and impurities contribute to the chemical exposome. Characterizing these chemicals and assessing their potential impacts on environmental and human health remain major challenges due to the vast size and diversity of chemical space.

Liquid chromatography coupled with high-resolution tandem mass spectrometry (LC–HRMS/MS) has become one of the most important analytical techniques for investigating complex samples. In particular, non-target analysis (NTA) enables the detection of a broad range of compounds without requiring prior knowledge of their identity. However, despite the ability of NTA to detect hundreds or thousands of features in a single sample, only a small fraction can typically be identified with confidence. As a result, a substantial gap remains between chemical detection and chemical understanding.

The aim of this thesis was to improve the understanding of chemical space coverage and chemical information extraction within LC–HRMS-based NTA workflows. This was addressed through investigations of chemical space coverage, molecular and spectral representations, feature prioritization, and compound annotation.

First, a meta-analysis of LC–HRMS-based NTA studies published between 2017 and 2023 was conducted to evaluate the extent of chemical space currently covered by envi-

ronmental NTA. The results showed that only a small fraction of the estimated chemical space is represented by reported compounds, highlighting the limited coverage achieved by existing workflows and the need for more comprehensive analytical and computational strategies.

The challenge of prioritizing unknown features was subsequently investigated from both molecular and experimental perspectives. At the molecular level, different strategies for combining non-hashed molecular fingerprints were evaluated across multiple regression and classification tasks. The results demonstrated that integrating complementary fingerprints improves predictive performance, emphasizing the importance of molecular representation in cheminformatics applications. At the experimental level, a prioritization strategy was developed that directly links chromatographic and fragmentation information to aquatic toxicity categories without requiring compound identification. The developed models showed performance comparable to structure-based approaches, demonstrating that meaningful toxicological information can be extracted directly from LC–HRMS measurements.

The final part of the thesis focused on the interpretation of tandem mass spectrometry data for compound annotation. First, twenty MS^2 similarity metrics were systematically evaluated using large collections of experimental spectra. The results showed that spectral similarity alone provides limited discriminatory power and is often insufficient for confident structural identification. This finding highlights the need to combine multiple sources of information during annotation workflows. Subsequently, the ambiguity of individual fragment ions was investigated using publicly available spectral databases. The analysis demonstrated that not all fragments contribute equally to compound discrimination and that fragment ambiguity strongly depends on database composition and chemical diversity. These findings indicate that the number of matched fragments alone is insufficient for assessing annotation confidence.

Overall, this thesis demonstrates that the major challenge in modern LC–HRMS-based NTA is not only the detection of chemicals but also the extraction and interpretation of chemical information. The results show that current workflows cover only a limited portion of environmentally relevant chemical space and that both molecular and spectral representations strongly influence the information that can be obtained from detected compounds. By investigating chemical space coverage, feature prioritization, molecular representation, spectral similarity, and fragment-level informativeness, this work contributes

to a better understanding of the opportunities and limitations of LC-HRMS-based NTA and supports the development of more data-driven approaches for characterization of the exposome chemical space.

Nederlandse samenvatting

De voortdurende groei in de productie en het gebruik van chemische stoffen heeft geleid tot steeds complexere mengsels van chemicaliën in milieu- en biologische monsters. Naast doelbewust geproduceerde stoffen dragen ook talrijke transformatieproducten, metabolieten en onzuiverheden bij aan het chemische exposoom. Het karakteriseren van deze stoffen en het beoordelen van hun potentiële effecten op het milieu en de menselijke gezondheid vormen een grote uitdaging vanwege de enorme omvang en diversiteit van de chemische ruimte.

Vloeistofchromatografie gekoppeld aan hoge-resolutie tandem-massaspectrometrie (LC–HRMS/MS) is uitgegroeid tot een van de belangrijkste analytische technieken voor het onderzoeken van complexe monsters. Met name non-target analyse (NTA) maakt het mogelijk een breed scala aan verbindingen te detecteren zonder voorafgaande kennis van hun identiteit. Hoewel NTA in staat is om honderden tot duizenden signalen (features) in één monster te detecteren, kan doorgaans slechts een klein deel daarvan met voldoende zekerheid worden geïdentificeerd. Hierdoor blijft een aanzienlijke kloof bestaan tussen de detectie van chemicaliën en het begrip van hun eigenschappen en betekenis.

Het doel van dit proefschrift was het verbeteren van het inzicht in de dekking van de chemische ruimte en de extractie van chemische informatie binnen LC–HRMS-gebaseerde NTA-workflows. Dit werd onderzocht aan de hand van studies naar de dekking van de chemische ruimte, moleculaire en spectrale representaties, feature-

prioriteitstelling en structuurannotatie.

Allereerst werd een meta-analyse uitgevoerd van LC–HRMS-gebaseerde NTA-studies die tussen 2017 en 2023 zijn gepubliceerd, om de mate waarin de chemische ruimte momenteel wordt afgedekt binnen milieuonderzoek met NTA te evalueren. De resultaten lieten zien dat slechts een klein deel van de geschatte chemische ruimte wordt vertegenwoordigd door de gerapporteerde verbindingen. Dit onderstreept de beperkte dekking van bestaande workflows en benadrukt de noodzaak van uitgebreidere analytische en computationele strategieën.

Vervolgens werd de uitdaging het stellen van prioriteiten voor onbekende features onderzocht vanuit zowel een moleculair als een experimenteel perspectief. Op moleculair niveau werden verschillende strategieën geëvalueerd voor het combineren van niet-gehashte moleculaire fingerprints binnen meerdere regressie- en classificatietaken. De resultaten toonden aan dat het combineren van complementaire fingerprints de voorspellende prestaties verbetert, wat het belang van moleculaire representaties binnen cheminformatica onderstreept. Op experimenteel niveau werd een strategie voor prioriteitstelling ontwikkeld die chromatografische en fragmentatie-informatie rechtstreeks koppelt aan categorieën van aquatische toxiciteit, zonder dat identificatie van de verbinding noodzakelijk is. De ontwikkelde modellen presteerden vergelijkbaar met structuurgebaseerde benaderingen, waarmee werd aangetoond dat relevante toxicologische informatie direct uit LC–HRMS-metingen kan worden afgeleid.

Het laatste deel van dit proefschrift richtte zich op de interpretatie van tandemmassaspectrometrische gegevens voor structuurannotatie. Eerst werden twintig MS^2-gelijkenisstatistieken systematisch geëvalueerd met behulp van grote verzamelingen experimentele spectra. De resultaten lieten zien dat spectrale overeenkomst op zichzelf slechts een beperkt onderscheidend vermogen bezit en vaak onvoldoende is voor een betrouwbare structurele identificatie. Dit benadrukt de noodzaak om meerdere informatiebronnen te combineren binnen annotatieworkflows. Vervolgens werd de ambiguïteit van individuele fragmenten onderzocht met behulp van openbaar beschikbare spectrale databases. De analyse toonde aan dat niet alle fragmenten in gelijke mate bijdragen aan het onderscheiden van verbindingen en dat de ambiguïteit van fragmenten sterk afhankelijk is van de samenstelling van de database en de chemische diversiteit daarvan. Deze bevindingen laten zien dat het aantal overeenkomende fragmenten op zichzelf onvoldoende is om de betrouwbaarheid van een annotatie te beoordelen.

Samenvattend laat dit proefschrift zien dat de grootste uitdaging binnen moderne LC–HRMS-gebaseerde NTA niet alleen ligt in het detecteren van chemicaliën, maar vooral in het extraheren en interpreteren van chemische informatie. De resultaten tonen aan dat huidige workflows slechts een beperkt deel van de milieurelevante chemische ruimte bestrijken en dat zowel moleculaire als spectrale representaties een sterke invloed hebben op de informatie die uit gedetecteerde verbindingen kan worden verkregen. Door de dekking van de chemische ruimte, feature-prioriteitstelling, moleculaire representaties, spectrale overeenkomst en de informatiewaarde van individuele fragmenten te onderzoeken, draagt dit onderzoek bij aan een beter begrip van de mogelijkheden en beperkingen van LC–HRMS-gebaseerde NTA en ondersteunt het de ontwikkeling van meer datagedreven benaderingen voor de karakterisering van de chemische ruimte van het exposoom.

Publications

Publications covered by the thesis (in order of the chapters):

1. Viktoriia Turkina*, Tobias Hulleman*, Jake W. O'Brien, Aleksandra Chojnacka, Kevin V.Thomas, Saer Samanipour, "Critical Assessment of the Chemical Space Covered by LC–HRMS Non-Targeted Analysis," *Environ. Sci. Technol.*, vol. 57, no. 38, pp. 14101-14112, 2023.
   *equal contribution
   Covered by Chapter 2.

2. Viktoriia Turkina, Melanie R.W. Messih, Etienne Kant, Jelle T. Gringhuis, Annemieke Petrignani, Garry Corthals, Jake W. O'Brien, Saer Samanipour, "Enhancing predictive modeling with molecular fingerprint fusion strategies," *Digit. Discov.*, vol. 5, pp. 2232-2242, 2026.
   Covered by Chapter 3.

3. Viktoriia Turkina, Jelle T. Gringhuis, Sanne Boot, Annemieke Petrignani, Garry Corthals, Antonia Praetorius, Jake W. O'Brien, Saer Samanipour, "Prioritization of Unknown LC-HRMS Features Based on Predicted Toxicity Categories," *Environ. Sci. Technol.*, vol. 59, no. 16, pp. 8004-8015, 2025.
   Covered by Chapter 4.

4. Viktoriia Turkina, Quinten van Erp, Annemieke Petrignani, Garry Corthals, Jake W. O'Brien, Saer Samanipour, "From Spectra to Local Networks: Evaluating MS^2 Similarity Metrics," *Anal. Chem.*, vol. 98, no. 24, pp. 17937 - 17947, 2026.
   Covered by Chapter 5.

5. Viktoriia Turkina, Lapo Renai, Annemieke Petrignani, Garry Corthals, Jake W. O'Brien, Saer Samanipour, "Quantifying Fragment-level Ambiguity in LC-HRMS/MS Spectral Databases," *ChemRxiv*, 2026.
   The manuscript is in preparation.
   Covered by Chapter 6.

   Publications not covered by the thesis (in chronological order):

6. Clara Granzotto, Amra Aksamija, Gerjen H. Tinnevelt, Viktoriia Turkina, Ken Sutherland, "New insight from MALDI-TOF MS and multivariate data analysis on the botanical origin of polysaccharide-based paint binders in ancient Egypt," *Anal. Methods*, vol. 16, pp. 2959-2971, 2024.

7. Alexandros Nikolopoulos, Denice van Herwerden, Viktoriia Turkina, Anneli Kruve, Melissa Baerenfaenger, Saer Samanipour, "Ionization efficiency prediction of electrospray ionization mass spectrometry analytes based on molecular fingerprints and cumulative neutral losses," *Journal of Cheminformatics*, vol. 17, no. 183, 2025.

8. Lapo Renai, Viktoriia Turkina, Tobias Hulleman, Alexandros Nikolopoulos, Andrea F.G. Gargano, Elvio D. Amato, Massimo Del Bubba, Saer Samanipour, "A Novel Chemical-Space-Dependent Strategy for Compound Selection in Non-target LC-HRMS Method Development Using Physicochemical and Structural Data," *Environ. Sci. Technol. Lett.*, vol. 12, no. 9, pp. 1162-1168, 2025.

9. Hiu-Lok Ngan, Viktoriia Turkina, Denice van Herwerden, Hong Yan, Zongwei Cai, Saer Samanipour, "Machine Learning for Enhanced Identification Probability in RPLC/HRMS Nontargeted Workflows" *Anal. Chem.*, vol. 97, no. 33, pp. 18028-18035, 2025.

10. Lapo Renai, Viktoriia Turkina, Aleksandra Chojnacka, Andrea F. G. Gargano, Saer Samanipour, "Measurable Feature Prediction for Estimating Chemical Space

Coverage in LC–ESI–HRMS Nontargeted Analysis," *Anal. Chem.*, vol. 98, no. 10, pp. 7637-7643, 2026.

11. Lapo Renai, Federica Calabrò, Viktoriia Turkina, Pradeep Dewapriya, Kevin V Thomas, Stefano Papazian, Saer Samanipour, "Understanding and Reporting PFAS Exposure: Wide-scope Retrospective Suspect Screening of High-resolution Mass Spectrometry Data from Exposomics Studies," *ChemRxiv*, 2026. (under review in Environ. Sci. Technol.)

12. Tobias Hulleman, Saer Samanipour, Viktoriia Turkina, Jinglong Li, Cassandra Rauert, Elvis Okoffo, Kevin Thomas, Jake O'Brien, "LC-HRMS Known and Unknown Feature Prioritisation Based on Machine Learning-Predicted Ready Biodegradability," *ChemRxiv*, 2026. (under review in Anal. Chem.)

Author contributions

CHAPTER 1: General introduction

Contribution†	Authors
Viktoriia Turkina	Conceptualization, Writing - original draft
Saer Samanipour	Conceptualization, Writing - review & editing

CHAPTER 2: Critical assessment of the chemical space covered by LC-HRMS non-targeted analysis

Contribution†	Authors
Viktoriia Turkina	Methodology, Investigation, Visualization, Formal analysis, Writing - original draft
Tobias Hulleman	Methodology, Investigation, Visualization, Formal analysis, Writing - review & editing
Jake W. O'Brien	Conceptualization, Funding acquisition, Supervision, Writing - review & editing
Aleksandra Chojnacka	Investigation, Writing - review & editing
Kevin V. Thomas	Writing - review & editing
Saer Samanipour	Conceptualization, Supervision, Methodology, Investigation, Resources, Funding acquisition, Writing - review & editing

CHAPTER 3: Enhancing Predictive Modeling with Molecular Fingerprint Fusion Strategies

Contribution†	Authors
Viktoriia Turkina	Conceptualization, Project administration, Methodology, Investigation, Visualization, Formal analysis, Writing - original draft
Melanie R.W. Messih	Investigation, Writing - review & editing
Etienne Kant	Investigation, Writing - review & editing
Jelle T. Gringhuis	Investigation, Writing - review & editing
Annemieke Petrignani	Funding acquisition, Supervision, Writing - review & editing
Garry Corthals	Writing - review & editing
Jake W. O'Brien	Conceptualization, Funding acquisition, Supervision, Writing - review & editing
Saer Samanipour	Conceptualization, Supervision, Methodology, Investigation, Resources, Funding acquisition, Writing - review & editing

CHAPTER 4: Prioritization of unknown LC-HRMS features based on predicted toxicity categories

Contribution†	Authors
Viktoriia Turkina	Conceptualization, Project administration, Methodology, Investigation, Visualization, Formal analysis, Writing - original draft
Jelle T. Gringhuis	Investigation, Visualization, Writing - review & editing
Sanne Boot	Investigation, Writing - review & editing
Annemieke Petrignani	Funding acquisition, Supervision, Writing - review & editing
Garry Corthals	Writing - review & editing
Jake W. O'Brien	Conceptualization, Funding acquisition, Supervision, Writing - review & editing
Saer Samanipour	Conceptualization, Supervision, Methodology, Investigation, Resources, Funding acquisition, Writing - review & editing

CHAPTER 5: From Spectra to Local Networks: Evaluating MS^2 Similarity Metrics

Contribution†	Authors
Viktoriia Turkina	Conceptualization, Project administration, Methodology, Investigation, Visualization, Formal analysis, Writing - original draft
Quinten van Erp	Methodology, Investigation
Annemieke Petrignani	Funding acquisition, Supervision, Writing - review & editing
Garry Corthals	Writing - review & editing
Jake W. O'Brien	Conceptualization, Funding acquisition, Supervision, Writing - review & editing
Saer Samanipour	Conceptualization, Supervision, Methodology, Investigation, Resources, Funding acquisition, Writing - review & editing

CHAPTER 6: Quantifying Fragment Ambiguity in LC-HRMS/MS Spectral Databases

Contribution†	Authors
Viktoriia Turkina	Conceptualization, Project administration, Methodology, Investigation, Visualization, Formal analysis, Writing - original draft
Lapo Renai	Methodology, Investigation, Writing - review & editing
Annemieke Petrignani	Funding acquisition, Supervision, Writing - review & editing
Garry Corthals	Writing - review & editing
Jake W. O'Brien	Funding acquisition, Supervision, Writing - review & editing
Saer Samanipour	Conceptualization, Supervision, Methodology, Investigation, Resources, Funding acquisition, Writing - review & editing

CHAPTER 7: Conclusion and outlook

Contribution†	Authors
Viktoriia Turkina	Conceptualization, Writing - original draft
Saer Samanipour	Conceptualization, Writing - review & editing

† Following the Contributor Role Taxonomy (CRediT), see `https://credit.niso.org/`.

Acknowledgements

It still feels unimaginable to me to see this book completed.

I would first like to express my gratitude to the members of my defense committee: **Prof. Arian van Asten**, **Dr. Frederik Béen**, **Dr. Andrea Bruno**, **Prof. Bernd Ensing**, **Prof. Paul Ferguson**, **Prof. Juliane Hollender**, and **Prof. Katrien Keune**. Thank you for accepting the invitation to be part of my committee and for taking the time to read and evaluate this thesis. Your work has inspired me throughout my PhD, and I feel privileged to have the opportunity to discuss my work with you.

I would not have been able to reach this point without the many people who guided me scientifically, shared their life wisdom, supported me when I thought things would never work out, and celebrated the victories along the way. I have to admit that my expectations of doing a PhD were quite different from reality. But it has definitely been an experience that reshaped the way I think, see, and learn. Four years ago, I arrived in Amsterdam full of excitement and fear about what was ahead. I was welcomed by wonderful people, and I could not value that more.

Saer, I cannot imagine how this book would look without your guidance. We had endless conversations about projects, interpreting results, and generating new ideas. You were always available when I needed your advice, which helped me enormously in finishing this thesis on time. Thank you for all the opportunities you gave me: presenting our work at international conferences, collaborating with researchers from different

parts of the world, and even working in Brisbane. I could not be more grateful.

Jake, thank you for your kindness, positivity, and openness towards everyone. Thank you for your guidance throughout my PhD and especially during my time at the University of Queensland. My time there changed my perspective on work and life in ways I had not expected. I was overwhelmed by the kindness, laughter, and support. I will always remember that experience warmly.

Garry, thank you for always having your door open and for taking the time to talk whenever I needed it. Our conversations meant a lot to me and made some difficult days a little easier.

Annemieke, thank you for being such an inspiring example of how to manage a million responsibilities at once, and for sharing your perspective on science, people, and life. I always left our conversations inspired.

Aleksandra, I am grateful for your limitless wisdom and energy. Thank you for always being there, and, of course, for all the plants. They definitely made my life greener!

I am also very grateful to the rest of the AC group. **Alina**, **Andrea**, **Arian**, **Bob**, **Maarten**, **Pascal**, **Peter**, **Tom**, thank you for sharing your knowledge and experience. **Annika**, we shared struggles and supported each other, but we also enjoyed morning tea, afternoon walks, and many beautiful moments together. Thank you for your energy, ideas, and support. **Joshka**, you are one of the most inspirational people I have met. I admire your strength, wisdom, perseverance, and kindness. **Nino**, thank you for your endless support and enthusiasm, which helped me keep going whenever I felt stuck. And thank you for our Dutch training sessions over a cup of coffee. **Ziran**, thank you for teaching me a bit of Chinese and for your curiosity, openness, and strong personality. **Irene**, the most stylish person in the room, thank you for the smiles, coffee breaks, conversations, and all your support with my applications and preparation for the defence. **Maria**, thank you for our talks about everything, work , life, adventures, and travels. Thank you for always believing in me and encouraging me along the way. **Denice**, Thank you for sharing your knowledge with me, introducing me to Julia, and guiding me through the world of LC–HRMS-based non-target analysis. And, of course, thank you for the rubber duck, which will now stay with Sanne. **Sanne**, thank you for your kindness and for always welcoming new people into our office, making it a more gezellig place.

Alejandro, **Althea**, **Gerben**, **Giulia**, **Jelle**, **Jeremy**, **Leon**, **Luca**, **Maartje**, **Mahsa**, **Masa**, **Merel**, **Mirjam**, **Rianne**, **Rick**, **Stef**, **Thomas**, **Tijmen**, **Wouter**, and **Xander**, thank you for being incredible colleagues. I cannot imagine these four years without you. Each of you brought ideas, inspiration, laughter, or simply made an ordinary day at work more enjoyable.

I was lucky to supervise brilliant Master's and Bachelor's students: **Adan**, **Deirdre**, **George**, **Gijs**, **Jelle**, **Lisa**, **Melanie**, and **Quinten**. It was an absolute privilege to work with you, and I learned a lot from each of you. Thank you for your curiosity, enthusiasm, independence, and all the ideas you brought to our projects. Supervising you was an important part of my own learning experience, and I could not have asked for a better group of students.

Alex, **Etienne**, **Jiani**, and **Tobias**, we started our projects around the same time and went through that part of the journey together. I was so lucky to share it with you. Thank you for the support, jokes, fun, and adventures, and for making those months so much more enjoyable.

Ioanna and **Skylar**, thank you for inviting me to be part of your project. I left every meeting with you feeling inspired. I am grateful for your curiosity, openness, and kindness, and for always making me feel welcome.

Jim and **Marco**, thank you for sharing your thoughts and ideas on my projects, surviving a PhD, and what comes after. I have always looked up to you, and I am so grateful that I could always turn to you for advice.

I would like to thank my paranymphs for agreeing to take on this role. **Lapo**, thank you for everything: for being the best conference buddy, a supportive colleague, and a wonderful friend. Thank you for our small adventures, long conversations, your scientific opinions and ideas, your wisdom, and also for always keeping things optimistic. Your motivation and work ethic inspired me every day. **Anna**, how lucky I was to meet you at my very first HIMS symposium. Thank you for all our philosophical conversations, shared books, countless lunches, adventures, explorations of the Netherlands and beyond, long walks, and all your support throughout these years. And thank you both for helping me organize my defence.

Finally, **Kirill**, thank you for supporting my decision to start this journey in the first place.

Thank you for moving to Amsterdam with me and building a new life here together. You supported me through my lowest moments and celebrated my achievements even more than I did. Thank you for reminding me how strong I can be when I need to be, and for always finding something, big or small, to surprise me and make life much brighter.

Looking back at the person who arrived in Amsterdam four years ago, excited and afraid of what was ahead, I could never have imagined everything these years would bring. There are so many more people I would like to thank for inspiring me along the way, broadening my horizons, and showing me different perspectives on science, life, and the world. The journey behind this book was shaped by all of you. Thank you for being part of it.

"Research is what I'm doing when I don't know what I'm doing."

— Wernher von Braun

