Environmental Science & Technology LETTERS

Open Access

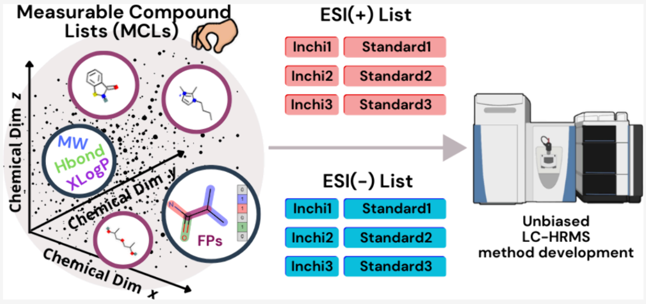
A Novel Chemical-Space-Dependent Strategy for Compound Selection in Non-target LC-HRMS Method Development Using Physicochemical and Structural Data

Lapo Renai,* Viktoriia Turkina, Tobias Hulleman, Alexandros Nikolopoulos, Andrea F. G. Gargano, Elvio D. Amato, Massimo Del Bubba, and Saer Samanipour*

Cite This: *Environ. Sci. Technol. Lett.* 2025, 12, 1162−1168

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: The virtual chemical space of substances, including emerging contaminants relevant to the environment and exposome, is rapidly expanding. Non-targeted analysis (NTA) by liquid chromatography−high-resolution mass spectrometry (LC-HRMS) is useful in measuring broad chemical space regions. Internal standards are typically used to optimize the selectivity and sensitivity of NTA LC-HRMS methods, assuming a linear relationship between structure and behavior across all analytes. However, this assumption fails for large, heterogeneous chemical spaces, narrowing measurable coverage to structurally similar compounds. We present a data-driven strategy for unbiased sampling of candidate structures for NTA LC-HRMS method development from extensive chemical spaces, such as the U.S. EPA's CompTox (>1 million chemicals). The workflow maximizes physicochemical/structural diversity using precomputed PubChem descriptors (e.g., molecular weight, XLogP) and grants LC-HRMS compatibility thanks to predicted mobility and ionization efficiency from molecular fingerprints. The resulting measurable compound lists (MCLs) provide broad, heterogeneous coverage for NTA method development, validation, and boundary assessment. Applied to the CompTox space, the approach yielded MCLs with greater chemical coverage and broader predicted LC-HRMS applicability than conventional "watch list" contaminants, offering a robust framework for enhancing NTA's measurable chemical space while preserving diversity.

KEYWORDS: *Non-target Analysis, Chemical Space, Exposomics, Emerging Contaminants, Mobility, Ionization Efficiency, Liquid Chromatography, Mass Spectrometry*

1. INTRODUCTION

The concept of chemical space has been introduced to depict the virtual regions occupied by all the existing and probable molecules according to their structural and physiochemical properties.[1,2] There is no unified vision of the entire chemical universe, and even if one existed, its high dimensionality and continuous expansion would limit any practical applications. For instance, based on physicochemical, bioactivity, bioavailability, and toxicity descriptors, the virtual space of small organic molecules is estimated to contain over 10^{60} compounds.[3,4] Consequently, a "practical" chemical (sub)space is typically chosen based on the compounds present in the sample matrix being analyzed. For example, the exposome space of emerging contaminants that humans encounter throughout their lives is addressed when dealing with environment-relevant samples (e.g., water samples).[5,6]

In theory, the comprehensive measure of a selected chemical subspace, exposome included, is enabled by non-target analysis (NTA) approaches, relying on data generated via high-resolution mass spectrometry (HRMS) hyphenated with separation techniques such as liquid chromatography (LC).[7,8] Yet, the measurable regions of a selected chemical space are limited to LC-HRMS-detectable compounds and depend on the quality of NTA study design,[4] including sample preparation and chromatographic analysis.[6,9]

These analytical steps impact the selectivity and sensitivity of recorded signals, affecting the composition of measured space.[10,11] To optimize selectivity (e.g., analyte recovery and peak capacity) and sensitivity (i.e., HRMS signal quality for precursor and fragmented ions), analytical NTA workflows use internal standards that represent the relevant chemical subspace.[12] In exposome analysis, chemicals of environmental concern (CECs) included in the European monitoring and/or

Received: July 29, 2025
Revised: August 12, 2025
Accepted: August 13, 2025
Published: August 18, 2025

https://doi.org/10.1021/acs.estlett.5c00759

ENTACT initiative lists are often used.[13−16] This strategy assumes linear relationships between properties (e.g., LogP and molecular weight) and the retention/ionization behaviors of internal standards relevant to the chemical subspace.[12] While valid for targeted analysis of specific compounds, this assumption fails for NTA of a heterogeneous chemical subspace due to the complex (non-linear) nature of the chemicals and samples involved.[11,17] As an example, Figure 1 depicts the distribution of ∼800,000 exposure-relevant chemicals from the CompTox database,[18] highlighting overlaps between EU monitoring (n = 62) and ENTACT lists (n = 1019). While these subgroups show a linear trend, the overall CompTox distribution is irregular.

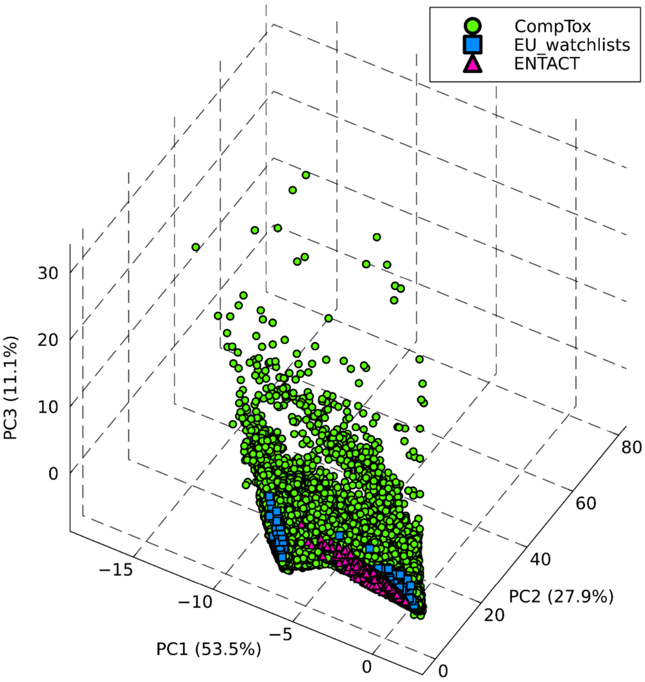


Figure 1. Score plot of the principal component analysis (PCA) on chemical structures within the CompTox database ($n \approx$ 800k) and the overlap with the shared compounds listed in European water monitoring program (n = 62) and the prioritized list of the ENTACT initiative (n = 1019). PCA was carried out on compounds' physicochemical properties extracted from PubChem (e.g., molecular weight and XLogP) and six calculated elemental mass defects (e.g., CO, CCl, and CN), which were used as model variables. See Section S5 of the Supporting Information for a 2D plot.

Thus, the selection of a few internal standards with limited chemical coverage can bias NTA method development. This restriction means that, for LC-HRMS, only a small set of chemically similar compounds may meet the necessary detection and identification criteria, which limits the measured subspace and decreases the discovery rate within the chosen chemical space.[4,11]

To minimize such bias in NTA analysis, it is crucial to select standards that accurately represent the desired chemical subspace and are LC-HRMS compatible, as well as managing their quantity and costs.[19] In this context, we present a data-driven and unbiased approach for the comprehensive and reproducible selection (or "sampling") of chemicals within a selected chemical subspace (e.g., exposomics). This method aims to generate measurable compound lists (MCLs) that serve two primary purposes: (i) maximizing the measurable space and (ii) establishing the applicability and detection boundaries of the LC-HRMS NTA methods. We utilize the exposome chemical space, approximated by the CompTox database. Chemical selection incorporates precomputed physicochemical properties, including predictions of environmental mobility and ionization efficiency (IE) as descriptors to identify candidate substances relevant for environmental monitoring and compatible with LC-HRMS analysis.[20−23]

2. MATERIALS AND METHODS

2.1. Workflow for MCL Sampling. The overall workflow for computing the chemical subspace and extracting MCLs starts from the conversion of CompTox dataset canonical SMILES into a set of six non-hashed molecular fingerprints (FPs),[24] followed by the collection of relevant physicochemical properties and the calculation of six elemental mass defects (EMDs),[25] predicted mobility classes,[22] and logarithmic-scale ionization efficiency (LogIE) values.[26] Full details are reported in Section S1 of the Supporting Information.

The chemical space is represented and analyzed by using PCA to identify regions for MCL sampling. Filters (see Section 2.2) were applied to select LC-HRMS-compatible candidates in both ionization modes.

The full workflow is depicted in Figure S1, including references to calculations and data availability.

2.2. Selection and Applicability of MCLs. Compound descriptors related to the CompTox dataset were obtained according to Sections S1.2 and S1.3 of the Supporting Information. PCA was conducted on 13 variables with the *ScikitLearn.jl* package, assigning numeric codes for mobility classification (i.e., "Non-mobile" = 1, "Mobile" = 2, and "Very mobile" = 3). Data were mean-centered and scaled before PCA for comparability. A symmetric gridding was applied to the dataset score plot to sample candidates across the selected chemical space (Table S1). Then candidates were extracted, retaining only the eligible structures according to the following criteria: (i) to avoid the inclusion of chemicals with high physicochemical and structural similarity, Jaccard distances were calculated using PCA variables to retain structures with distance values >0.15 (i.e., a Roger/Tanimoto score >0.85);[27] (ii) to ensure the inclusion of ESI ionizable structures, the dataset was filtered according to LogIE >3.5 for positive ionization mode (ESI(+)) and to LogIE <1.5 and hydrogen bond donor count >0 for negative ionization mode (ESI(−)). Since the LogIE prediction models were trained on LogIE values from ESI(+) data, using a scale with methyl benzoate as anchor compound,[28] for structures that are potentially ionizable in ESI(−), LogIE <1.5 was chosen as the threshold for lower probability of positive charge stabilization, also considering the presence of hydrogen bond donor functions that promote negative ionization. LogIE thresholds can be adjusted on (i) the initial dataset (chemical space) involved and (ii) the required ionization polarity.

Filtered structures were grouped by mobility class and then ranked to select a user-defined number of ESI(+) candidates by descending LogIE, and ESI(−) candidates by descending LogIE <1.5.

The applicability of sampled MCLs was validated in terms of chemical coverage and retention behavior in comparison with the list of monitored CECs in the European Union water framework (Table S2).[14−16] To assess chemical space coverage, besides PCA plots, the ClassyFire chemical taxonomy tool was used to identify the chemical class related to each structure (using InChIKeys) included in the MCLs.[29] Retention behavior

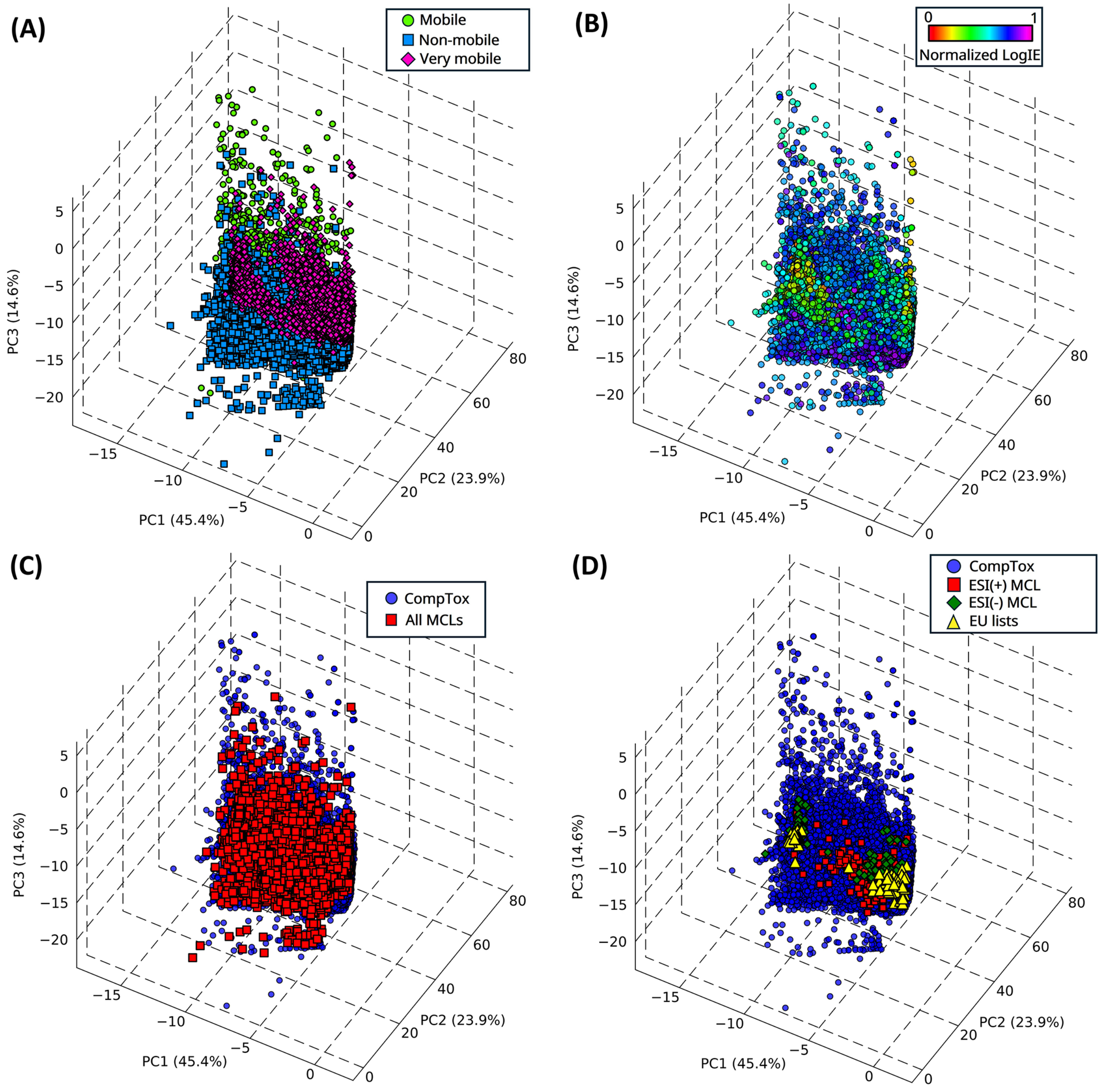


Figure 2. PCA score plots of CompTox structures (n = 785,294) highlighting (A) predicted mobility, (B) predicted ionization efficiency (LogIE normalized scale), (C) sampled chemicals for MCLs (red squares, n = 17,743), and (D) sampled candidates for ESI(+) MCL (red squares, n = 150) and ESI(−) MCL (green diamonds, n = 150) in comparison with chemicals monitored under the EU water framework (Table S2). See Section S5 of the Supporting Information for grayscale plots.

was investigated by means of retention index (RI) predictions using the *RIprediction.jl* package developed by van Herwerden et al.,[17] providing supplementary retention classification related to RPLC subspace (i.e., −1 = "outside", 0 = "maybe", 1 = "inside" RPLC domain). After candidate selection, to assess the availability of purchasable standards inside the obtained MCLs, patent and literature counts for each structure were extracted using PubChem CIDs.

3. RESULTS AND DISCUSSION

3.1. CompTox Space Distribution. PCA was performed to visualize the distribution of the CompTox chemical space (n = 785,294), as displayed in Figure S2. This dataset effectively balances the chemical subspace's structural diversity with the known physicochemical distribution for a clear demonstration of the proposed approach.

The first three PCs explain about 84% of the variance in the original data (Figure S2A), representing the CompTox chemical subspace (Figure S2B). This spatial representation, now including mobility and LogIE predictions, shows no distinct trends, similar to Figure 1.

Based on explained variance, the loadings define a complex 3D space of structural and physicochemical features. PC1 scores/loadings decrease with molecular weight, XLogP, H-

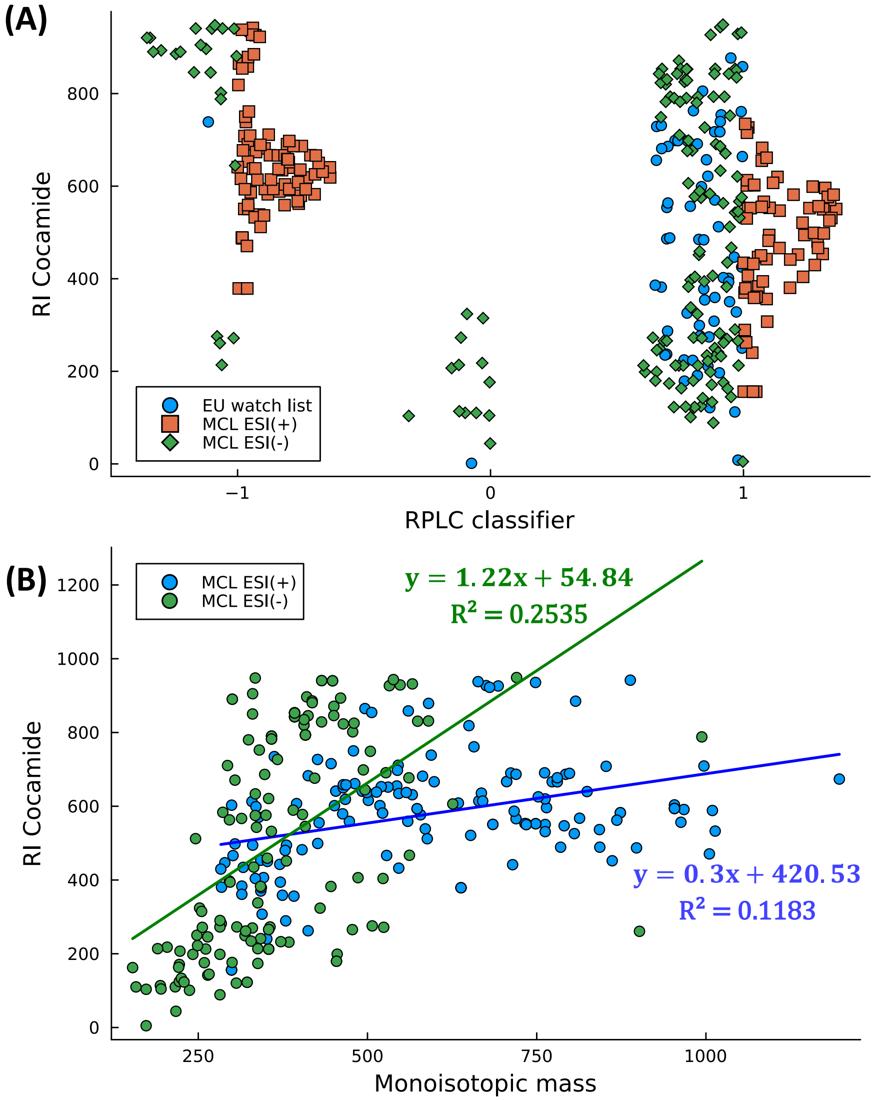


Figure 3. Predicted retention index (RI) values based on the cocamide scale for ESI(−) and ESI(+) MCLs plotted against (A) retention classification related to RPLC subspace (−1 = "outside", 0 = "maybe", 1 = "inside", vs EU monitored chemicals) and (B) monoisotopic mass. Linear regression trend lines and equations summarize the distinct characteristics of the two distributions.

bonding features, and TPSA but increase with structure-related variables (mobility, LogIE, and EMDs) (Figure S2C). PC2 shows a partial reverse trend, with higher scores linked to greater polarity and molecular weight. In PC3, higher mobility corresponds to lower ionization efficiency, molecular weight, and XLogP. This is due to LogIE's dependence on the positive ionization scale, which is reduced by electronegative groups that enhance molecular interactions and mobility.[30]

Mobility and LogIE variables indeed exhibit interesting patterns in the CompTox chemical space, as displayed by the PC spaces in Figures 2A and 2B. The distribution of "Non-mobile" structures is generally localized in a lower region (>molecular weight and XLogP) of the subspace, aligning with the increase of molecular weight and hydrophobicity,[20,31] also suggesting the compatibility of this chemical space region with the reversed-phase chromatographic domain and high ESI(+) IE.

Instead, LogIE seems to be negatively affected by the increase of mobility, as the ionizable moieties stabilizing positive charges gradually decrease with molecular weight, other than the increase of electronegative functions more prone to ionize in ESI(−).[32] The presence of highly mobile structures suggests that the CompTox subspace includes chemicals beyond the reversed-phase LC domain, highlighting the need to consider additional selectivity mechanisms in NTA for its measurability. Including structures with varied mobility makes MCLs effective in defining measurable space boundaries under specific analytical conditions.

This complex scenario underlines the need for "data-driven" sampling, collecting representative structures for MCLs throughout the entire chemical subspace to preserve the original physicochemical variability involved in the CompTox database.

3.2. MCL Selection and Validation. Our approach aims to maximize the coverage of diverse structures out of the thousands encompassed in the chemical subspace, which was achieved through the symmetric gridding procedure (Table S1) and the use of Jaccard distances. Figure 2C shows the outcome of the MCL selection workflow, highlighting in the PC scores plot the selected structures (red squares, n = 17,743) that are homogeneously distributed in all the PC space according to the applied criteria, also reducing the risk of oversampling from high-density areas within the scores space compared to a pure random sampling. Additionally, the use of LogIE filters helped to retain only chemicals highly compatible with ESI(+) and ESI(−), thus already highlighting a wide space of measurable LC-HRMS NTA chemicals. The ESI(−) compatibility criteria applied to the CompTox dataset included structures containing electronegative and resonance-stabilizing functions promoting a negative charge (see atomic FPs in the *Data preprocessed descriptors* at DOI: 10.6084/m9.figshare.28788143.v3). The use of ESI IE-based criteria restricts MCL sampling for apolar or neutral compounds more compatible with APCI, "biasing" MCL extended chemical coverage.[33,34] However, in the CompTox space, the threshold LogIE <1.5 included structures like bile acids that are compatible with both ESI and APCI.

To better scale the selected structures for a feasible analytical standard selection, it is advisable to pick a specific number of candidate structures (e.g., $n = 50$) from each mobility category. These candidates are sorted in descending order based on the maximum LogIE values for ESI(+) MCL and on LogIE values <1.5 for ESI(−) MCL (supplementary data available at DOI: 10.6084/m9.figshare.28788143.v3). The number of MCL candidates can be tuned depending on the selected chemical space and required chemical coverage.

3.2.1. Chemical Coverage of MCLs.

Figure 2D displays the result of the PC scores overlap by the ESI(+) and ESI(−) MCLs within the CompTox chemical space and the CEC structures belonging to EU monitoring lists (Table S2), since there are no alternative approaches available for comparison. MCLs provided a greater coverage of the PC space compared to the EU-monitored chemicals, supporting the rationale behind the use of MCLs for LC-HRMS NTA method development.

The classification density distribution for MCL ESI(+) and ESI(−) structures ($n = 300$) vs the EU CECs by the ClassyFire tool (see section 2.2) is summarized in Figure S3.

MCLs encompass a broader range of chemical classes than EU lists due to the uniform sampling of CompTox structures. While there is some overlap with EU CECs (e.g., piperazines and diphenylmethanes), the latter are skewed on specific classes of chemicals (e.g., alkyl fluorides), demonstrating how class-dependent internal standards can bias NTA method development and chemical coverage.

3.2.2. Chromatographic and Mass Domain of MCLs.

Alongside broad chemical coverage, MCLs need to reflect diverse retention behaviors, thanks to mobility class prediction trained on chromatographic data, including gradient and organic modifier content.[22]

Predicted RI values were calculated to demonstrate the range of the chromatographic domain coverage for the candidate structures of both ESI(+) and ESI(−) MCLs. Figure 3 shows the results of the RI predictions based on the cocamide scale versus RPLC classes and monoisotopic masses from both MCLs and EU CECs. The sampled structures exhibited a wide RI coverage ($5 < \mathrm{RI} < 900$), being nonetheless fully aligned with the coverage of EU chemicals "inside" the RPLC domain, with only a few exceptions for ESI(−) MCL candidates with RI ≈ 900 (Figure 3A). Among EU CECs, only metformin and guanylurea were classified as "maybe inside" and "outside" RPLC domains, whereas MCLs included several candidates in these alternative chromatographic domains, thus stepping outside the linearity assumption limiting comprehensive chemical measurements. Recalling the discussion in section 3.1, MCLs can assist NTA coverage under different LC conditions (e.g., hydrophilic interaction and supercritical fluid chromatography).[8,17] Figure S4 supports this evidence, showing some of the MCL candidates with available experimental records (RepoRT database)[35] in alternative chromatographic setups beyond RP C18, including phenyl and HILIC stationary phases.

Additionally, the MCL candidates have been analyzed for their coverage of the mass domain in relation to the predicted RIs (Figure 3B). The distribution of the monoisotopic masses of the combined MCLs covers the 100−1200 Da range. This outcome aligns with the PCA trend of molecular weights (section 3.1). Furthermore, the masses of the MCL structures exhibit weak linear trends when compared to RIs ($R^2 \approx 0.1$−0.2), supporting our hypothesis that the reduced structure−retention relationships enhance the chemical coverage for NTA method development.

3.3. Availability of MCL Subset.

A final consideration for applying the selected MCLs is the availability of the structures as reference standards for LC-HRMS analysis. Table S3 shows the results on the ESI(+) and ESI(−) structures ($n = 300$) according to the patent and literature references extracted from the PubChem database. 72% ($n = 217$) of the selected structures possess records of patent registrations and/or literature citations, encouraging their potential availability as purchasable standards. These candidates preserved an optimal chemical space coverage of the PC score plot of the CompTox dataset (Figure S5), fully in line with the complete MCL lists (Figure 2D).

The availability of analytical standards depends on the user, method, and target chemical space. If a standard is unavailable or more candidates are needed, the PCA score matrix can guide the selection of alternative structures.

4. ENVIRONMENTAL IMPLICATION

Sampling MCLs across a vast chemical (sub)space of interest can strongly reduce the "biased coverage" in NTA method development, providing reliable internal standard lists stretching (beyond the RPLC linearity assumption) and defining the measurable chemical space under multiple LC-HRMS experimental conditions. Furthermore, it applies to any chemical subspace, from exposomics to metabolomics, with an available molecular representation.

Contextually, by utilizing key structural and physicochemical variables, such as mobility and ionization efficiency, MCL selection maintains the diversity of the CompTox space and yields lists compatible with LC-HRMS analysis in both ionization modes.

Moreover, MCL subsets based on mobility and ionization efficiency cover a significant range of the CompTox space regarding the mass range, predicted retention index, and structural variability, exceeding the chemical classes outlined in the European "watch lists" for water monitoring.

In the specific context of the exposome chemical subspace, represented here by the CompTox dataset, we consider MCLs to be effective tools for understanding and expanding the chemical coverage of NTA methods in identifying unknown or undetected CECs. MCLs not only have the potential to enhance the rate of CEC discovery but also can assist users in assessing the boundaries of chemical space, thereby reducing the risk of false positive detections in environmental analysis.

■ ASSOCIATED CONTENT

Data Availability Statement

The input and output datasets retrieved in this study are available as .csv files and can be found at 10.6084/m9.figshare.28788143.v3. The code that was used to perform the calculations and MCL selection is available at https://bitbucket.org/laporen/mcl_selection_workflow/src/main/. The packages used for variable collection and prediction are also available: PubChemCrawler.jl - https://github.com/JuliaHealth/PubChemCrawler.jl; Mobility_prediction.jl - https://github.com/tobihul/Mobility_prediction; IE_prediction.jl - https://github.com/pockos56/IE_prediction.jl; and ClassyFireR - https://rdrr.io/cran/classyfireR/f/vignettes/Getting-Started.Rmd.

SI Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.estlett.5c00759.

> S1, dataset creation; S2, principal component analysis of the CompTox dataset; S3, chemicals on European monitoring lists; S4, MCL selection and validation; and S5, appendix (PDF)

AUTHOR INFORMATION

Corresponding Authors

Lapo Renai − *Van't Hoff Institute for Molecular Sciences (HIMS), University of Amsterdam, 1090 GD Amsterdam, The Netherlands;* orcid.org/0000-0002-0738-7901; Email: l.renai@uva.nl

Saer Samanipour − *Van't Hoff Institute for Molecular Sciences (HIMS), University of Amsterdam, 1090 GD Amsterdam, The Netherlands; Queensland Alliance for Environmental Health Sciences (QAEHS), Woolloongabba, QLD 4102, Australia; UvA Data Science Center, University of Amsterdam, 1012 WP Amsterdam, The Netherlands;* orcid.org/0000-0001-8270-6979; Email: s.samanipour@uva.nl

Authors

Viktoriia Turkina − *Van't Hoff Institute for Molecular Sciences (HIMS), University of Amsterdam, 1090 GD Amsterdam, The Netherlands;* orcid.org/0000-0002-7153-2333

Tobias Hulleman − *Queensland Alliance for Environmental Health Sciences (QAEHS), Woolloongabba, QLD 4102, Australia; ARC Training Centre for Hyphenated Analytical Separation Technologies (HyTECH), Queensland Alliance for Environmental Health Sciences (QAEHS), Woolloongabba, QLD 4102, Australia;* orcid.org/0009-0001-5439-9702

Alexandros Nikolopoulos − *Van't Hoff Institute for Molecular Sciences (HIMS), University of Amsterdam, 1090 GD Amsterdam, The Netherlands;* orcid.org/0009-0004-5832-1197

Andrea F. G. Gargano − *Van't Hoff Institute for Molecular Sciences (HIMS), University of Amsterdam, 1090 GD Amsterdam, The Netherlands;* orcid.org/0000-0003-3361-7341

Elvio D. Amato − *KWR Water Research Institute, 3433 PE Nieuwegein, The Netherlands*

Massimo Del Bubba − *Department of Chemistry, University of Florence, 50019 Sesto Fiorentino, Florence, Italy;* orcid.org/0000-0002-6326-6549

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.estlett.5c00759

Notes

A preprint of this work was published on ChemRxiv:10.26434/chemrxiv-2025-zv4h0-v2 (July 30, 2025).

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The presented research has been supported by funding from the European Union under the program HORIZON.1.2 - Marie Skłodowska-Curie Actions (MSCA) Postdoctoral Fellowships 2023-BEspace 101150312 (10.3030/101150312). S.S. and V.T. thank ChemistryNL for financial support. S.S. is thankful to the UvA Data Science Center for their support via the "accelerate program".

REFERENCES

(1) Arús-Pous, J.; Awale, M.; Probst, D.; Reymond, J. L. Exploring Chemical Space with Machine Learning. *Chimia* **2019**, *73*, 1018−1023.

(2) Reymond, J. L. Chemical Space as a Unifying Theme for Chemistry. *J. Cheminform.* **2025**, *17*, 6 DOI: 10.1186/s13321-025-00954-0.

(3) Reymond, J. L. The Chemical Space Project. *Acc. Chem. Res.* **2015**, *48* (3), 722−730.

(4) 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**, *4*, 2412−2425.

(5) Lai, Y.; Koelmel, J. P.; Walker, D. I.; Price, E. J.; Papazian, S.; Manz, K. E.; Castilla-Fernández, D.; Bowden, J. A.; Nikiforov, V.; David, A.; Bessonneau, V.; Amer, B.; Seethapathy, S.; Hu, X.; Lin, E. Z.; Jbebli, A.; McNeil, B. R.; Barupal, D.; Cerasa, M.; Xie, H.; et al. High-Resolution Mass Spectrometry for Human Exposomics: Expanding Chemical Space Coverage. *Environ. Sci. Technol.* **2024**, *58*, 12784−12822.

(6) Manz, K. E.; Feerick, A.; Braun, J. M.; Feng, Y. L.; Hall, A.; Koelmel, J.; Manzano, C.; Newton, S. R.; Pennell, K. D.; Place, B. J.; Godri Pollitt, K. J.; Prasse, C.; Young, J. A. Non-Targeted Analysis (NTA) and Suspect Screening Analysis (SSA): A Review of Examining the Chemical Exposome. *J. Exposure Sci. Environ. Epidemiol.* **2023**, *33*, 524−536.

(7) Milman, B. L.; Zhurkovich, I. K. The Chemical Space for Non-Target Analysis. *TrAC - Trends Anal. Chem.* **2017**, 179−187, DOI: 10.1016/j.trac.2017.09.013.

(8) Black, G.; Lowe, C.; Anumol, T.; Bade, J.; Favela, K.; Feng, Y. L.; Knolhoff, A.; Mceachran, A.; Nuñez, J.; Fisher, C.; Peter, K.; Quinete, N. S.; Sobus, J.; Sussman, E.; Watson, W.; Wickramasekara, S.; Williams, A.; Young, T. Exploring Chemical Space in Non-Targeted Analysis: A Proposed ChemSpace Tool. *Anal. Bioanal. Chem.* **2023**, *415* (1), 35−44.

(9) Huang, X.; Zhai, W.; Su, W.; Yang, Z.; Liang, W.; Wang, P.; Ruan, T.; Jiang, G. Exploration of Chemical Space Covered by Nontarget Screening Based on the Prediction of Chemical Substances Amenable to LC-HRMS Analysis. *Environ. Sci. Technol. Lett.* **2025**, *12*, 661.

(10) Hajeb, P.; Zhu, L.; Bossi, R.; Vorkamp, K. Sample Preparation Techniques for Suspect and Non-Target Screening of Emerging Contaminants. *Chemosphere* **2022**, *1*, 132306 DOI: 10.1016/j.chemosphere.2021.132306.

(11) 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**, *26*, 14101−14112.

(12) Hollender, J.; Schymanski, E. L.; Ahrens, L.; Alygizakis, N.; Béen, F.; Bijlsma, L.; Brunner, A. M.; Celma, A.; Fildier, A.; Fu, Q.; Gago-Ferrero, P.; Gil-Solsona, R.; Haglund, P.; Hansen, M.; Kaserzon, S.; Kruve, A.; Lamoree, M.; Margoum, C.; Meijer, J.; Merel, S. M. NORMAN Guidance on Suspect and Non-Target Screening in Environmental Monitoring. *Environ. Sci. Europe* **2023**, *35*, 75 DOI: 10.1186/s12302-023-00779-4.

(13) Commission Recommendation (EU) 2022/1431 on the Monitoring of Perfluoroalkyl Substances in Food. Aug 24, 2022. https://eur-lex.europa.eu/eli/reco/2022/1431/oj/eng

(14) European Parliament Legislative Resolution on the Proposal for a "Directive of the European Parliament and of the Council Concerning Urban Wastewater Treatment". April 10, 2024. http://data.europa.eu/eli/C/2023/250/oj.

(15) Directive (EU) 2020/2184 of the European Parliament and of the Council on the Quality of Water Intended for Human Consumption. December 16, 2020. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32020L2184.

(16) Commission Implementing Decision (EU) 2022/1307 establishing a watch list of substances for Union-wide monitoring in the field of water policy. July 22, 2022. https://eur-lex.europa.eu/eli/dec_impl/2022/1307/oj/eng.

(17) van Herwerden, D.; Nikolopoulos, A.; Barron, L. P.; O'Brien, J. W.; Pirok, B. W. J.; Thomas, K. V.; Samanipour, S. Exploring the Chemical Subspace of RPLC: A Data Driven Approach. *Anal. Chim. Acta* **2024**, *1317*, 342869.

(18) Williams, A. J.; Grulke, C. M.; Edwards, J.; McEachran, A. D.; Mansouri, K.; Baker, N. C.; Patlewicz, G.; Shah, I.; Wambaugh, J. F.; Judson, R. S.; Richard, A. M. The CompTox Chemistry Dashboard: A Community Data Resource for Environmental Chemistry. *J. Cheminform.* **2017**, *9*, 61.

(19) Fisher, C. M.; Miele, M. M.; Knolhoff, A. M. Community Needs and Proposed Solutions for a Broadly Applicable Standard/QC Mixture for High-Resolution Mass Spectrometry-Based Non-Targeted Analysis. *Anal. Chem.* **2025**, *97*, 5424.

(20) Arp, H. P. H.; Hale, S. E. Assessing the Persistence and Mobility of Organic Substances to Protect Freshwater Resources. *ACS Environ. Au* **2022**, *16*, 482−509.

(21) Aalizadeh, R.; Nikolopoulou, V.; Alygizakis, N.; Slobodnik, J.; Thomaidis, N. S. A Novel Workflow for Semi-Quantification of Emerging Contaminants in Environmental Samples Analyzed by LC-HRMS. *Anal. Bioanal. Chem.* **2022**, *414* (25), 7435−7450.

(22) Hulleman, T.; Samanipour, S.; Haddad, P. R.; Rauert, C.; Okoffo, E. D.; Thomas, K. V.; O'brien, J. W. Machine Learning for Predicting Environmental Mobility Based on Retention Behaviour. *ChemRxiv Preprint* **2025**, DOI: 10.26434/chemrxiv-2025-xl6xl.

(23) Wang, W. C.; Amini, N.; Huber, C.; Kull, M.; Kruve, A. Active Learning Improves Ionization Efficiency Predictions and Quantification in Nontargeted LC/HRMS. *Anal. Chem.* **2025**, *97*, 13131.

(24) Turkina, V.; Messih, M. R. W.; Kant, E.; Gringhuis, J.; Petrignani, A.; Corthals, G.; O'Brien, J. W.; Samanipour, S. Molecular Fingerprints Optimization for Enhanced Predictive Modeling. *ChemRxiv Preprint* **2024**, DOI: 10.26434/chemrxiv-2024-zr2vr.

(25) Sleno, L. The Use of Mass Defect in Modern Mass Spectrometry. *J. Mass Spectrom.* **2012**, *47* (2), 226−236.

(26) 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. *ChemRxiv Preprint* **2025**, DOI: 10.26434/chemrxiv-2025-dc9gd.

(27) Safizadeh, H.; Simpkins, S. W.; Nelson, J.; Li, S. C.; Piotrowski, J. S.; Yoshimura, M.; Yashiroda, Y.; Hirano, H.; Osada, H.; Yoshida, M.; Boone, C.; Myers, C. L. Improving Measures of Chemical Structural Similarity Using Machine Learning on Chemical-Genetic Interactions. *J. Chem. Inf. Model.* **2021**, *27*, 4156−4172.

(28) Oss, M.; Kruve, A.; Herodes, K.; Leito, I. Electrospray Ionization Efficiency Scale of Organic Compound. *Anal. Chem.* **2010**, *82* (7), 2865−2872.

(29) Djoumbou Feunang, Y.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; Greiner, R.; Wishart, D. S. ClassyFire: Automated Chemical Classification with a Comprehensive Computable Taxonomy. *J. Cheminform.* **2016**, *8* (1), 1−20.

(30) Kruve, A.; Kaupmees, K.; Liigand, J.; Leito, I. Negative Electrospray Ionization via Deprotonation: Predicting the Ionization Efficiency. *Anal. Chem.* **2014**, *86* (10), 4822−4830.

(31) Ji, W. X.; Tian, Y. C.; Li, A.; Gu, X. M.; Sun, H. F.; Cai, M. H.; Shen, S. Q.; Zuo, Y. T.; Li, W. T. Unravelling Relationships between Fluorescence Spectra, Molecular Weight Distribution and Hydrophobicity Fraction of Dissolved Organic Matter in Municipal Wastewater. *Chemosphere* **2022**, *308*, 136359 DOI: 10.1016/j.chemosphere.2022.136359.

(32) Nguyen, T. B.; Nizkorodov, S. A.; Laskin, A.; Laskin, J. An Approach toward Quantification of Organic Compounds in Complex Environmental Samples Using High-Resolution Electrospray Ionization Mass Spectrometry. *Anal. Methods* **2013**, *5* (1), 72−80.

(33) Singh, R. R.; Chao, A.; Phillips, K. A.; Xia, X. R.; Shea, D.; Sobus, J. R.; Schymanski, E. L.; Ulrich, E. M. Expanded Coverage of Non-Targeted LC-HRMS Using Atmospheric Pressure Chemical Ionization: A Case Study with ENTACT Mixtures. *Anal. Bioanal. Chem.* **2020**, *412* (20), 4931−4939.

(34) Bonnefille, B.; Karlsson, O.; Rian, M. B.; Raqib, R.; Parvez, F.; Papazian, S.; Islam, M. S.; Martin, J. W. Nontarget Analysis of Polluted Surface Waters in Bangladesh Using Open Science Workflows. *Environ. Sci. Technol.* **2023**, *57* (17), 6808−6824.

(35) Kretschmer, F.; Harrieder, E. M.; Hoffmann, M. A.; Böcker, S.; Witting, M. RepoRT: A Comprehensive Repository for Small Molecule Retention Times. *Nat. Methods* **2024**, *1*, 153−155.

CAS BIOFINDER DISCOVERY PLATFORM™
PRECISION DATA FOR FASTER DRUG DISCOVERY
CAS BioFinder helps you identify targets, biomarkers, and pathways
Unlock insights
CAS
A division of the American Chemical Society