ChemComm
Chemical Communications

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: L. Renai, J. Heemskerk, F. Béen and S. Samanipour, *Chem. Commun.*, 2026, DOI: 10.1039/D6CC02811J.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/chemcomm

View Article Online
DOI: 10.1039/D6CC02811J

Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Chemical Space Blind Spots: How Chromatographic Selectivity Dictates Chemical Measurability and Coverage of LC-HRMS comprehensive analysis

Lapo Renai,*[a‡] Jens Heemskerk,[a‡] Frederic Béen,[b,c] and Saer Samanipour[a,d,e]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

Liquid chromatography high-resolution mass spectrometry enables broad chemical detection by comprehensive accurate mass to charge ratio measurement of the components in complex samples; yet analytical design constrains chemical space measurability. A meta-analysis of 236 methods and over 75,000 measured compounds reveals strong convergence toward reversed-phase separations, limiting the coverage of sample chemical diversity. This "measurability trap" narrows the observable chemical space and can lead to the underrepresentation of many environmentally and biologically relevant compounds.

The chemical space refers to the wide collection of all existing and plausible chemical structures from an organic chemistry perspective, also including the chemicals relevant to human and environmental exposure.[1,2] Modern analytical chemistry strives for a holistic, wide-scope view of chemical space, and liquid chromatography high-resolution mass spectrometry (LC–HRMS) has become a cornerstone of comprehensive screening methods (known as non-targeted analysis), enabling in principle the simultaneous detection of thousands (known and unknown) compounds in environmental and biological samples.[3] While often perceived as unbiased, the effective chemical space accessible to LC–HRMS is intrinsically shaped by analytical design, which constrains the expansion of measurability and consequently contributes to limit the discovery rate for novel structures. In particular, chromatographic selectivity, defined by the interactions of known analytes (analytical standards) with stationary phase and mobile phase, acts as a primary driver of measurability in LC–HRMS analysis. Only the compounds that are successfully retained are efficiently ionized, and ultimately detected under given experimental conditions, define the measurable chemical space.[4,5] Although the non-targeted vision favors generic mobile phase conditions—such as broad, shallow gradient elution programs and limited use of modifiers to maximize peak capacity (often exceeding 1000 chemical features in the separation domain)—interactions with the stationary phase inevitably impose selectivity.[6] As a result, measurability varies across LC–HRMS methods, systematically prioritizing the detection of some chemical classes over others.

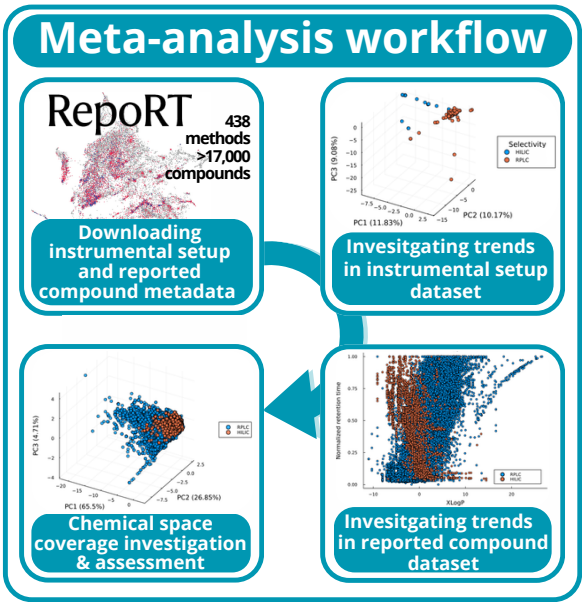

Fig. 1 Meta-analysis workflow evaluating the chemical space coverage of the curated LC methods available in the RepoRT repository.

While HRMS acquisition can register thousands of signals within a single analysis, only a subset of these features reflects

[a] Van 't Hoff Institute for Molecular Sciences (HIMS), University of Amsterdam, 1090 GD, Amsterdam, the Netherlands; E-mail: l.renai@uva.nl
[b] Amsterdam Institute for Life and Environment, Vrije Universiteit Amsterdam, 1081HV Amsterdam, The Netherlands
[c] KWR Water Research Institute, 3433BB Nieuwegein, The Netherlands
[d] UvA Data Science Center, University of Amsterdam, Amsterdam
[e] Queensland Alliance for Environmental Health Sciences (QAEHS), 20 Cornwall Street, Woolloongabba, QLD, 4102, Australia
‡ These authors contributed equally to this work

View Article Online
DOI: 10.1039/D6CC02811J

compounds that are effectively retained and detected under the chosen chromatographic conditions, and therefore carries meaningful chemical information. This selectivity–measurability bias can create a false perception (i.e., a measurability "trap") of analytical comprehensiveness, particularly among non-specialist end-users, despite underlying physicochemical constraints, with important implications for exposure assessment. To clearly communicate the actual LC trade-off in holistically capturing the chemical space, we performed a meta-analysis on a large data repository storing methods and retention time information (RepoRT) on small molecules, including exposure-relevant chemicals and metabolites (**Figure 1**).[7] RepoRT currently covers 438 method entries and over 17,083 unique compounds measured across 49 different LC stationary phases under variable mobile phase conditions (e.g., eluent combination, modifiers, and flow rates) compiled from publicly available datasets and peer-reviewed studies. Importantly, many of the reported chemicals were measured in a targeted manner using authentic reference standards, providing reliable chromatographic information across diverse LC setups. Thus, poorly retained compounds (e.g., sugars and organic acids) under incompatible stationary phases are reported alongside successful retention under alternative separations. Such dataset representatively defines the practical boundaries of accessible chemical space: targeted methods can probe the extremes of measurability, whereas non-targeted coverage depends on effective retention under the method's selectivity. To systematically investigate chemical space coverage in LC, both the variability of analytical methods and the physicochemical diversity of the measured compounds must be considered simultaneously. Accordingly, the compiled collection from RepoRT was organized into two complementary datasets: one describing homogeneous and comparable LC instrumental configurations (i.e., refined method metadata, n=236) and one capturing the chemical descriptors of analytes (retention time entries n=78,226) retained under those analytical methods (section S1 of the *Supporting information*). Together, these datasets reflect how method design and chemical properties jointly define the observable chemical space, although they do not represent all theoretically achievable selectivity modes and compounds. Method meta-data reported eight distinct column chemistry types across the setups, classified by USP code (**Figure 2a**; see also **Table S1**). The USP system categorizes LC columns according to stationary-phase chemistry.[8] L1 (C18) dominates the dataset, accounting for 78% of all setups (n=186), followed by L122 and L11 (≃8% each; n=19 and 18). All remaining column types individually represent only ≃1–2%. This distribution confirms the strong predominance of reversed-phase LC (RPLC) selectivity in RepoRT. With C18, phenyl, C8, and pentafluorophenyl phases collectively representing 89% of setups, retention is largely governed by hydrophobic interactions, biasing the measurable chemical space toward moderately polar and non-polar compounds (partition coefficient (XLogP) between -1 and 6).[4,9,10] Only 11% of methods employ hydrophilic interaction chromatography (HILIC) stationary phases (bare silica, zwitterionic, alkylamide), suggesting that retention data of polar and very polar analytes are underrepresented. Operational parameters also exhibit high convergence, typically utilizing 100-150 mm columns, standard UHPLC flow rates (0.2-0.4 mL/min),[11] and aqueous/organic gradients with 0.1% formic acid(**Figure S1**). Such homogeneity reinforces a systematic bias toward RPLC-compatible compounds. This also reflects in the targeted scope of the reported methods, with 90.3% of setups reporting fewer than 500 analytes, despite seeking higher theoretical single-run capacity (**Figure 2b**).

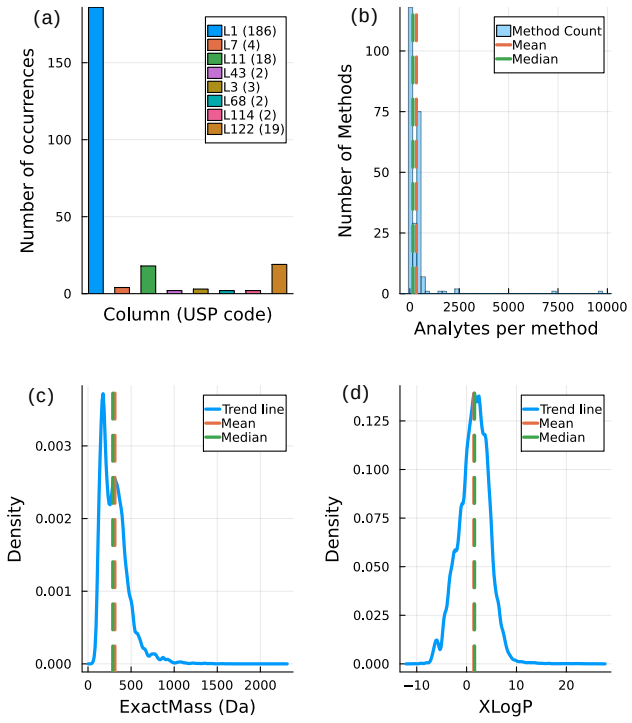

Fig. 2 Overview of the curated RepoRT datasets, showing (a) column stationary phases type USP code distribution, (b) number of analytes per method, and distributions of (c) exact mass and (d) XLogP for each chemical entry across methods.

This suggests that current methodological uniformity leaves much of the theoretical peak capacity and chemical space unexploited. However, a closer look at the distribution of the >78k measured compounds reveals an unexpected trend. Given that RepoRT reports mainly small molecules (mean exact mass 312 Da; **Figure 2c**), the XlogP range covered by the majority of compounds measured across setups is remarkably evenly distributed between polar and nonpolar structures (-10<XLogP<10; **Figure 2d**). How, then, can such a substantial fraction of polar and hydrophilic compounds appear to be captured under predominantly reversed-phase conditions? Examination of a representative subset of frequently analyzed compounds in RepoRT (**Figure S2a**), several highly polar molecules—such as hexose (XLogP=-2.6), mannitol (XLogP=-3.1), and quinic acid (XLogP=-2.4)—are reported as retained under RPLC setups. Such observations may be representative of applications in targeted analysis, as well as of a missing or incomplete exclusion of early-eluting features in non-targeted protocols. Nevertheless, applying a dead volume threshold to filter out poorly retained species (section S2 of the *Supporting information*, **Figure S2b**) substantially reduces the apparent RPLC coverage for polar compounds. The presence of

View Article Online
DOI: 10.1039/D6CC02811J

these analytes in RPLC datasets indicates that non-targeted chemical coverage in the reversed-phase domain can be easily overestimated, despite HILIC being a more robust method for retaining these highly polar compounds. While such retention data are valuable for defining method boundaries, from a non-targeted perspective, unknown compounds eluting in the dead volume or undergoing breakthrough are unlikely to be reliably detected, as poor retention results in low-quality, noisy MS signals and limited discovery potential.[12] To explore relationships between methods and chemical coverage, the two RepoRT datasets were analyzed by principal component analysis (PCA, section S3 of the *Supporting information*). The RPLC- and HILIC-based setups generated two distinct clusters in the PC scores' plot described by a moderate explained variance (31.1%, **Figure 3a**). This trend is consistent with the contribution of the first three components, largely driven by stationary-phase typology, with clear separation between C18 and zwitterionic columns along PC1 and phenyl-based RPLC along PC3 (**Figure S3** and **S4a**). Particle size and flow rate further contribute, with larger particles associated with HILIC and higher flow rates with RPLC under UHPLC conditions. PC2 is mainly influenced by eluent composition, delineating RPLC through acidic aqueous phases with strong organic modifiers (e.g., acetonitrile). K-means clustering ((**Figure S4b**) and centroids similarity heatmap (**Figure S5**) were used to interpret the variables driving this limited separation. Cluster 1 (n=19) mainly comprises L122 columns and HILIC-specific eluents. Cluster 2 (n=209) confirms the dominant RPLC group, including C18, C8, and phenyl phases, and shows broader internal variability (low centroid similarity). Clusters 3–5 are distinguished by alternative stationary phases, the use of unconventional organic modifiers (e.g., isopropanol, acetone), and different buffer systems (ammonium formate or phosphate). Expectedly, the reported compounds across setups highlight that vast majority of the RepoRT-represented chemical space is dominated by RPLC (**Figure 3b**, RPLC dead-volume entries removed). HILIC compounds occupy a partially distinct but strongly overlapping region, indicating that both modes predominantly capture similar physicochemical domains. Most chemical variability is captured along PC1 (65.5%), which positively correlates the increase in exact mass with the increase in sites generating polar interactions (acid-base descriptors and topological polar surface area (TPSA)), but inversely with XLogP (**Figure S6**). This confirms what has been previously demonstrated on unrealistic chemical coverage under RPLC conditions. PC2 (26.85%) further refines this distribution by capturing the combined variation of exact mass and XLogP, but it does not substantially resolve RPLC and HILIC chemical space overlap. This convergence likely reflects methodological constraints, such as the limited flexibility (i.e., less tunable retention behavior) of broad-gradient HILIC methods, resulting in repeated sampling of the same physicochemical regions rather than true orthogonal expansion of measurability. A better view of these constraints is provided by normalized retention vs XLogP and TPSA (**Figure S7**), showing that HILIC captures a large fraction of semi-polar and moderately apolar compounds, resulting in substantial overlap with RPLC within the central polarity domain. This depicts how HILIC is often implemented as a complementary "inverse" of RPLC (i.e., switching mobile phase composition) without fully exploiting its distinct separation mechanisms.

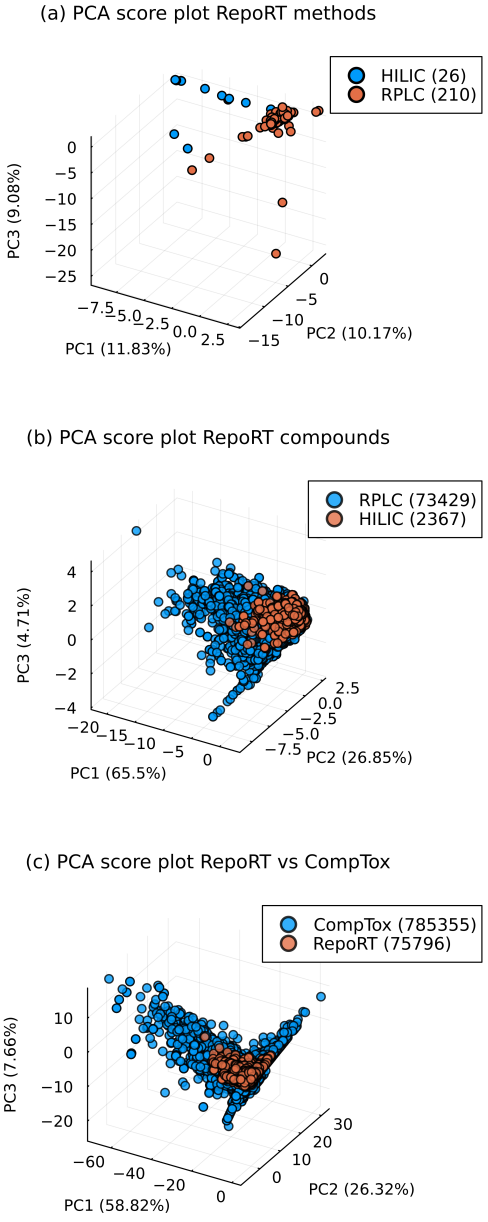

Fig. 3 Coverage by PCA score plots based on five molecular descriptors (ExactMass, XLogP, HBondDonorCount, HBondAcceptorCount, and TPSA). (a) RepoRT method coverage showing separation between HILIC (n = 26) and RPLC (n = 210) entries. (b) RepoRT chemical space colored by chromatographic selectivity (RPLC, n = 73,429; HILIC, n = 2,367). (c) Overlay of the RepoRT dataset (n = 75,796) and the CompTox dataset (n = 785,355), illustrating relative chemical space coverage. Percent variance explained is shown on the axes.

Rather than extending measurability, both selectivity modes concentrate on the intermediate descriptor space. Exact-mass distributions reinforce this pattern: RPLC spans a broad mass range, including >1000 Da compounds, whereas HILIC is largely confined below 1000 Da, regardless of polarity (**Figure S8**). Overall, no substantial expansion of chemical coverage is observed between RPLC and HILIC. In principle, HILIC would be expected

View Article Online
DOI: 10.1039/D6CC02811J

to shift measurability toward highly polar chemicals; yet, such a pronounced displacement is not evident, due in part to the imbalanced methods' reporting, the limited representation of optimized HILIC data, and biases toward available analytical standards also contribute to this trend. To contextualize the chemical space covered by curated LC methods, the RepoRT compounds were projected against the U.S. EPA CompTox Chemistry Dashboard (≃800k chemicals representing an approximation of the exposome chemical space) in the same physicochemical descriptor space (**Figure 3c** & **Figure S9**). The RepoRT compounds occupy only a confined subregion of the broader CompTox chemical space, **Figure 3c**. While substantial overlap exists in the central PC domain, large areas of CompTox characterized by high polarity (high TPSA and H-bond capacity), extreme hydrophobicity (high XLogP), and very large molecular weights remain entirely unrepresented. Assuming that RepoRT is a good sample of chemical LC-HRMS measurability, the detectable chemical space does not cover the maximum sample diversity, but results in a projection constrained by a poorly exploited selectivity. Due to methodological convergence, analyses repeatedly capture well-characterized regions of chemical space, while others remain largely inaccessible.[1] Although combining RPLC with HILIC is often proposed as a strategy to enhance orthogonality, our meta-analysis suggests that, within the currently reported methods, this expansion remains modest. Currently, the available data on HILIC do not substantially displace coverage toward the highly polar domain.[13] Although incorporating additional chromatographic modes (e.g., SFC or IC) can extend chemical space coverage, the resulting gains remain incremental relative to the vast theoretical chemical universe, with no combination of current approaches achieving comprehensive measurability due to RP-centered compound variability, inter-platform correlation biases, and implementation incompatibilities for such orthogonal multidimensional workflows.[14] This constraint is not purely physicochemical but also methodological: analytical practice is biased toward compounds that can be confirmed with reference standards. As a result, reported chemical space largely reflects known compounds, while unknown features, defining the frontier of measurability, remain underrepresented. Rather than pursuing unreliable comprehensive coverage, method-specific measurability domains should be explicitly defined and quantified, and chemical space coverage considered alongside sensitivity and mass accuracy as a key performance metric (e.g., by predicting fractional coverage and mapping the measurable structural/physicochemical property boundaries).[5,13] Further future best practices should prioritize systematic reporting of unknown features and the development of continuously updated repositories capturing evolving tentative structures beyond currently recognized compounds.[15] Only by documenting both what is observed and what remains unseen, non-targeted analysis and exposomics can move beyond the measurability trap toward a genuinely exploratory strategy. Such chemical coverage-directed expansion of chromatographic diversity and repository may redefine these measurable domains and modify the trends currently observed in LC–HRMS chemical-space accessibility. This work was supported by the EU MSCA Postdoctoral Fellowship 2023 (Grant No. 101150312).

Author Contributions

L.R.(first & corresponding author): funding acquisition, conceptualization, investigation, methodology, writing - original draft; J.H.: investigation, formal analysis; F.B.: supervision, writing - review & editing, S.S.: supervision, writing - review & editing.

Conflicts of interest

There are no conflicts to declare.

Data availability

Supporting information (SI) is provided as a separate file. Datasets and meta-analysis code are available at `https://doi.org/10.6084/m9.figshare.31553716`.

Notes and references

1 S. Samanipour, L. P. Barron, D. van Herwerden, A. Praetorius, K. V. Thomas and J. W. O'Brien, *JACS Au*, 2024, **4**, 2412–2425.
2 B. L. Milman and I. K. Zhurkovich, *TrAC Trends in Analytical Chemistry*, 2017, **97**, 179–187.
3 K. E. Manz, A. Feerick, J. M. Braun, Y.-L. Feng, A. Hall, J. Koelmel, C. Manzano, S. R. Newton, K. D. Pennell and B. J. e. a. Place, *Journal of exposure science & environmental epidemiology*, 2023, **33**, 524–536.
4 T. Hulleman, V. Turkina, J. W. O'Brien, A. Chojnacka, K. V. Thomas and S. Samanipour, *Environmental Science & Technology*, 2023, **57**, 14101–14112.
5 L. Renai, V. Turkina, T. Hulleman, A. Nikolopoulos, A. F. Gargano, E. D. Amato, M. Del Bubba and S. Samanipour, *Environmental Science & Technology Letters*, 2025.
6 J. Hollender, E. L. Schymanski, L. Ahrens, N. Alygizakis, F. Béen, L. Bijlsma, A. M. Brunner, A. Celma, A. Fildier, Q. Fu *et al.*, *Environmental Sciences Europe*, 2023, **35**, 75.
7 F. Kretschmer, E.-M. Harrieder, M. A. Hoffmann, S. Böcker and M. Witting, *nature methods*, 2024, **21**, 153–155.
8 K. Huynh-Ba and R. C. Moreton, *Specification of Drug Substances and Products*, Elsevier, 2025, pp. 185–204.
9 F. Menger, P. Gago-Ferrero, K. Wiberg and L. Ahrens, *Trends in Environmental Analytical Chemistry*, 2020, **28**, e00102.
10 T. Reemtsma, U. Berger, H. P. H. Arp, H. Gallard, T. P. Knepper, M. Neumann, J. B. Quintana and P. d. Voogt, *Mind the gap: Persistent and mobile organic compounds water contaminants that slip through*, 2016.
11 S. Fekete, J. Schappler, J.-L. Veuthey and D. Guillarme, *TrAC Trends in Analytical Chemistry*, 2014, **63**, 2–13.
12 B. Ng, N. Quinete and P. R. Gardinali, *Science of the Total Environment*, 2020, **713**, 136568.
13 L. Renai, V. Turkina, A. Chojnacka, A. F. G. Gargano and S. Samanipour, *Analytical Chemistry*, 2026, **98**, 7637–7643.
14 J. Zweigle, M. Schlusener, J. Flottmann, T. Bader, N. H. Vidkjær, U. E. Bollmann, J. H. Christensen and S. Tisler, *Analytical Chemistry*, 2025, **97**, 25099–25110.
15 L. Renai, F. Calabrò, V. Turkina, P. Dewapriya, K. V. Thomas, S. Papazian and S. Samanipour, *ChemrXiv Preprint*, 2026.

View Article Online
DOI: 10.1039/D6CC02811J

Data availability

Supporting information (SI) is provided as a separate file. Datasets and meta-analysis code are available at https://doi. org/10.6084/m9.figshare.31553716.