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

Lapo Renai,[*,†] Jens Heemskerk,[†] Frederic Béen,[‡,¶] and Saer Samanipour[*,†,§,‖]

†*Van 't Hoff Institute for Molecular Sciences (HIMS), University of Amsterdam, 1090 GD, Amsterdam, the Netherlands*

‡*Amsterdam Institute for Life and Environment, Vrije Universiteit Amsterdam, 1081HV Amsterdam, The Netherlands*

¶*KWR Water Research Institute, 3433BB Nieuwegein, The Netherlands*

§*UvA Data Science Center, University of Amsterdam, Amsterdam*

‖*Queensland Alliance for Environmental Health Sciences (QAEHS), 20 Cornwall Street, Woolloongabba, QLD, 4102, Australia*

E-mail: l.renai@uva.nl; s.samanipour@uva.nl

List of Figures

List of Tables

S1. Data acquisition and curation

The RepoRT repository was downloaded from GitHub (`https://github.com/michael witting/RepoRT/tree/master`, and the available experimental metadata was processed using the Julia 1.11.4 programming language.This collection was divided into two separate datasets, one containing the LC instrumental setup metadata and one containing chemical descriptors on the analytes that were retained by these setups.

Instrumental setup metadata were extracted from the metadata ".tsv" files associated with each analytical method available in the raw data repository. Only methods reporting a complete set of key chromatographic parameters—namely column name, USP column code, column length, internal diameter, particle size, column temperature, and flow rate—were retained, resulting in a total of 236 unique LC setups. To facilitate comparative analysis, USP column codes were one-hot encoded, yielding a standardized representation of stationary-phase selectivity.

In parallel, chemical structure information was obtained by extracting compound identities and retention times from the corresponding rtdata.tsv files for each included method. For all retained analytes (n = 75,796), precomputed physicochemical descriptors—including exact mass, predicted XLogP, hydrogen-bond donor and acceptor counts, and topological polar surface area (TPSA)—were retrieved from PubChem using the *PubChemCrawler.jl* package. The complete code for dataset curation is available at `https://doi.org/10.6084/m9.figshare.31553716` (*Meta-analysis_markdown.ipynb*).

Because RepoRT is derived primarily from published targeted and semi-targeted analytical studies, the dataset inherently reflects methodological and reporting biases toward compounds that are routinely measured, commercially available as analytical standards, or compatible with commonly used LC–HRMS workflows. Therefore, the present study does not aim to define the absolute theoretical limits of chromatographic selectivity or LC–HRMS measurability, but rather to evaluate how analytical practice shapes the experimentally accessible chemical space.

S2. RepoRT methods and compounds overview

Table S1: Overview of the USP column codes for the 236 RepoRT methods, along with their occurrence and stationary-phase description. USP classification of LC columns categorizes columns based on stationary phase characteristics, assigning an "L" number that corresponds to chemical features such as the nature of the bonded phase and functional groups connected to the silica surface

Selectivity	USP code	Occurrence	Description
RPLC (89%)	L1	186	Octadecylsilane (C18) chemically bonded to porous or non-porous silica or ceramic microparticles or superficially porous particles or a monolithic rod.
	L7	4	Octylsilane (C8) chemically bonded to totally or superficially porous silica particles or a monolithic silica rod.
	L11	18	Phenyl groups chemically bonded to porous silica particles or superficially porous particles or a monolithic silica rod.
	L43	2	Pentafluorophenyl groups chemically bonded to silica particles or superficially porous particles by a propyl spacer.
HILIC (11%)	L3	3	Porous silica particles or superficially porous particles or a monolithic silica rod.
	L68	2	Spherical porous silica covalently modified with alkyl amide groups, not endcapped.
	L114	2	Sulfobetaine graft-polymerized to totally or superficially porous silica, packed with densely bonded zwitterionic groups with 1:1 charge balance.
	L122	19	Sulfobetaine graft-polymerized to totally or superficially porous hydrophilic polymer particles, packed with densely bonded zwitterionic groups with 1:1 charge balance.

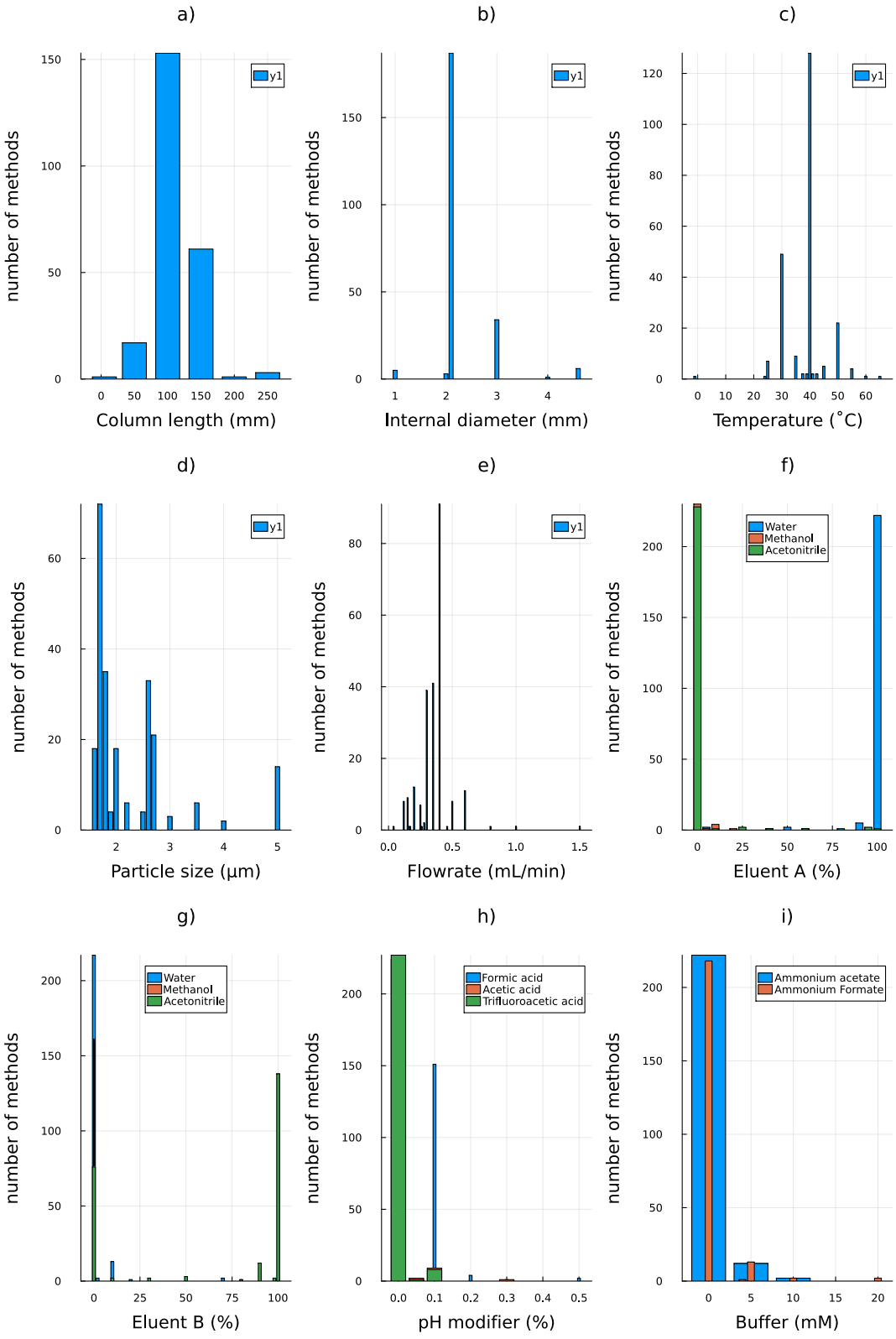


Figure S1: Key method descriptors for the RepoRT homogeneous LC setups (n = 236): (a) column length, (b) internal diameter, (c) temperature, (d) particle size, (e) flowrate, (f) eluent A composition, (g) eluent B composition, (h) pH modifier, and (i) buffer.

To evaluate the influence of poorly retained compounds on the apparent chemical-space coverage of LC methods, an additional exploratory filtering step was applied to the RepoRT retention-time dataset. Compounds eluting within the chromatographic dead-volume region were excluded using a retention-time threshold of 2 min. This value was selected as a pragmatic average threshold considering the predominant column dimensions (mainly 100–150 mm) and flow-rate conditions (typically 0.2–0.4 mL min^{-1}) represented across the curated RepoRT methods (Figure S1). Compound occurrence frequencies were subsequently recalculated separately for the complete and retention-filtered datasets by grouping entries according to PubChem CID and associated chromatographic selectivity labels (Figure S2). The complete code for retention time exclusion is available at `https://doi.org/10.6084/m9.figshare.31553716` (*Meta-analysis_markdown.ipynb*). This comparison was used to investigate how exclusion of poorly retained species affected the apparent overlap between RPLC and HILIC chemical-space coverage.

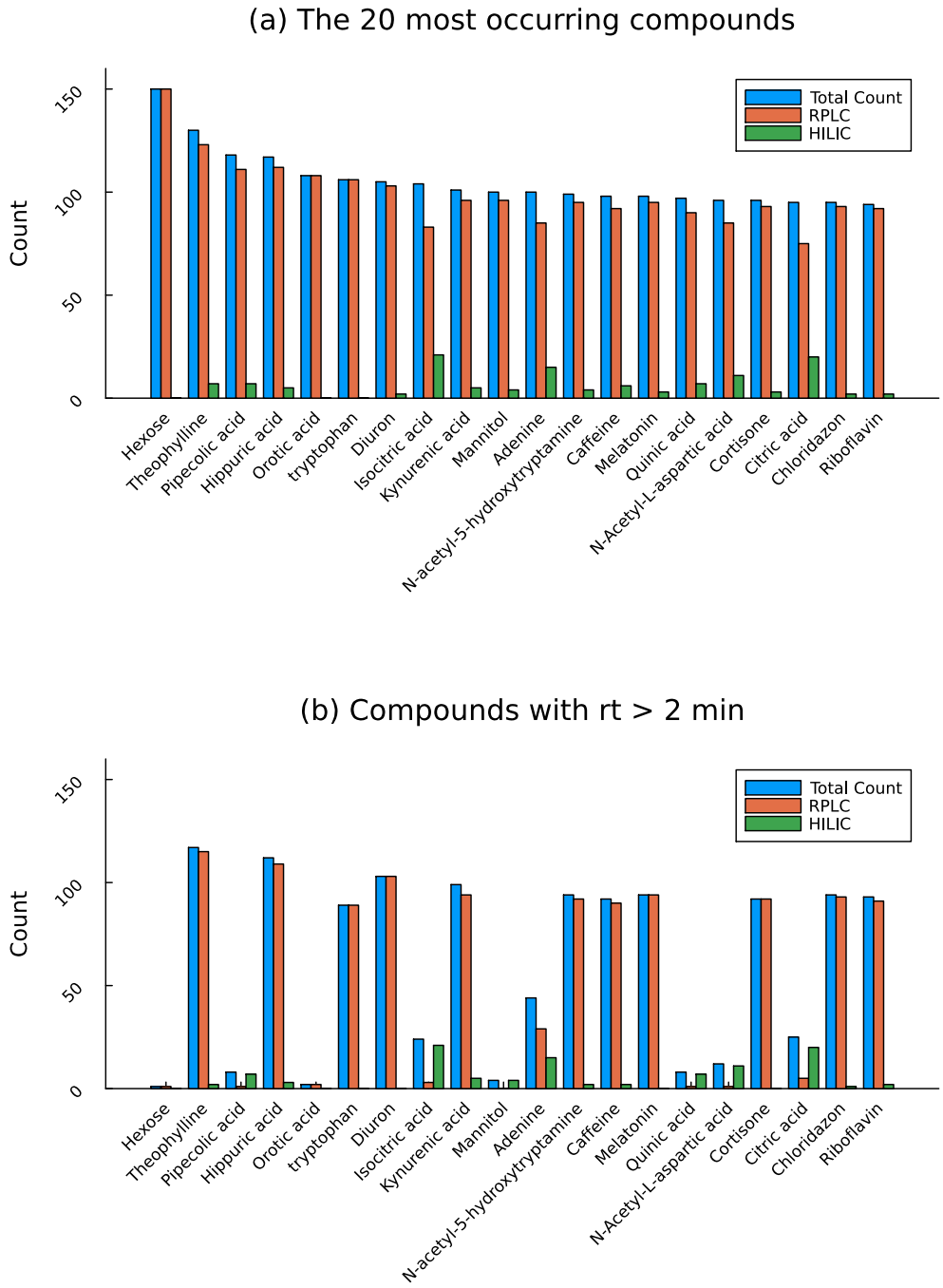


Figure S2: Bar plots showing (a) the 20 most occurring compounds within the RepoRT repository and (b) these same compounds but excluding retention times less than 2 min, while indicating the total count and the count associated with the different selectivity's (RPLC vs HILIC)

S3. RepoRT multivariate analysis

The relationships between 236 LC methods and 75,796 reported compounds were investigated using PCA for dimension reduction and visualization of orthogonal components described by the original variables that capture the most variance in the data. Prior to PCA, all datasets were mean-centered and autoscaled to unit variance. This preprocessing was performed by subtracting the mean value of each variable and dividing by its standard deviation. The complete code used for PCA is available at `https://doi.org/10.6084/m9.figshare.31553716` (*Meta-analysis_markdown.ipynb*). Autoscaling was necessary because the investigated descriptors and chromatographic parameters span substantially different numerical ranges and variances (e.g., exact mass, TPSA, XLogP, flow rate, and particle size). Without normalization, variables with larger absolute magnitudes or variances would disproportionately dominate the PCA model and bias the resulting component structure. The applied preprocessing therefore ensured that all variables contributed comparably to the variance structure of the dataset, allowing the PCA to capture relative relationships between chromatographic selectivity and physicochemical descriptors rather than differences driven purely by scale magnitude.

Method metadata

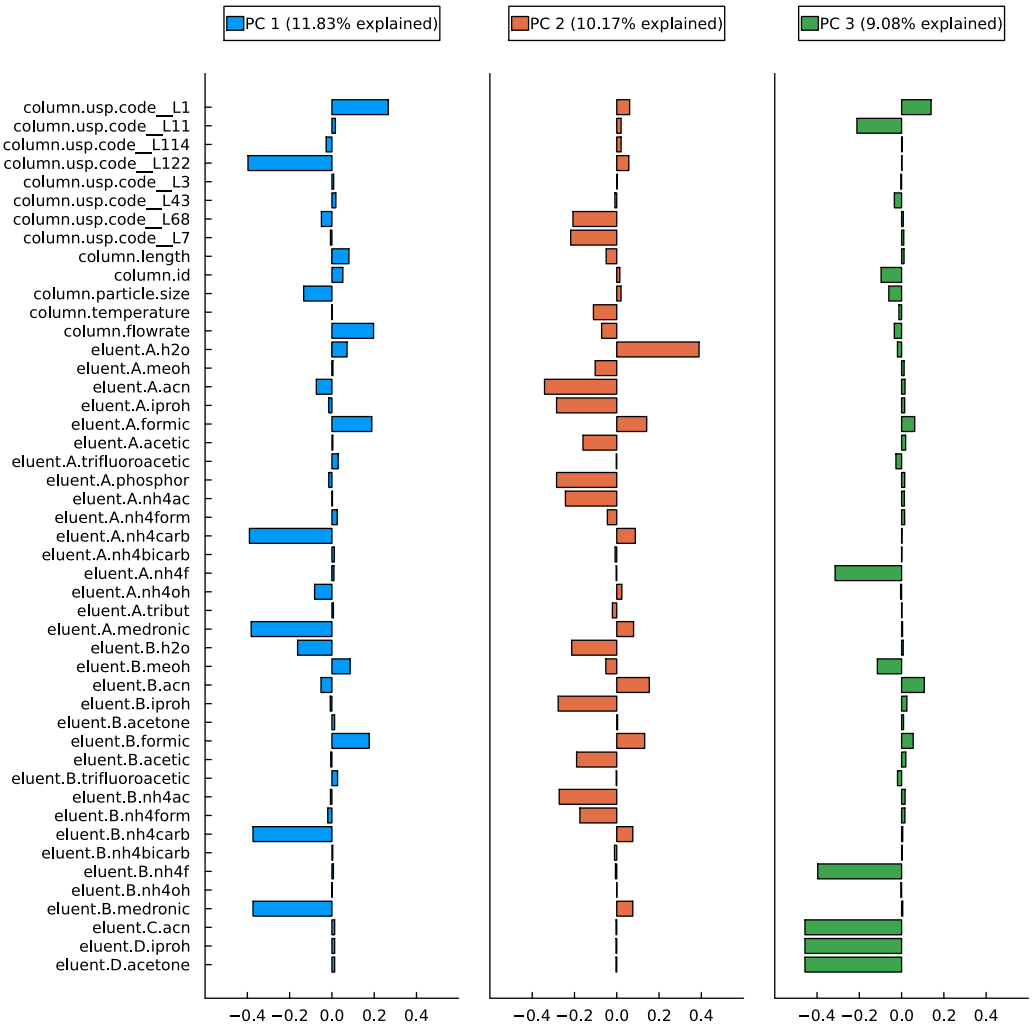
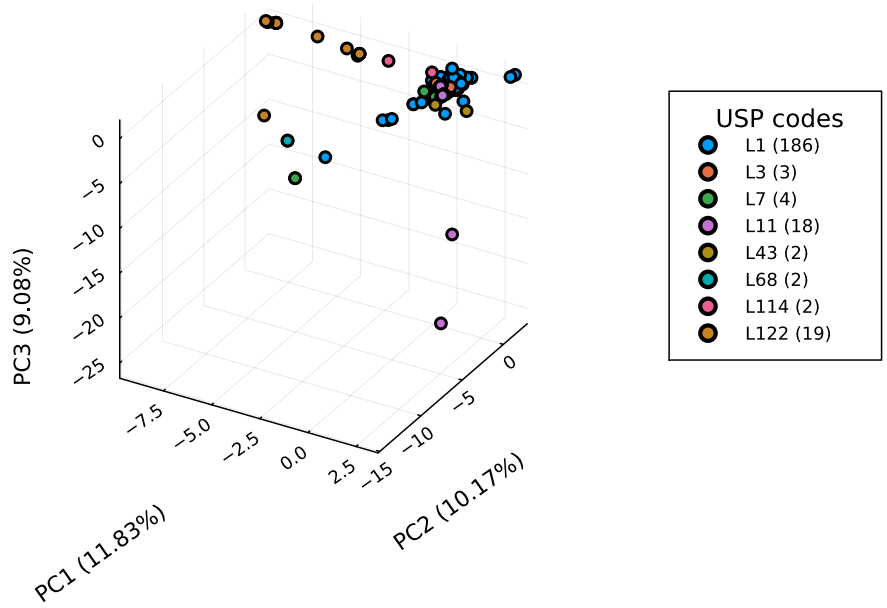


Figure S3: Bar plots of the loadings for the first three principal components used to describe the LC setups.

(a) PCA LC columns RepoRT methods



(b) PCA k-means RepoRT methods

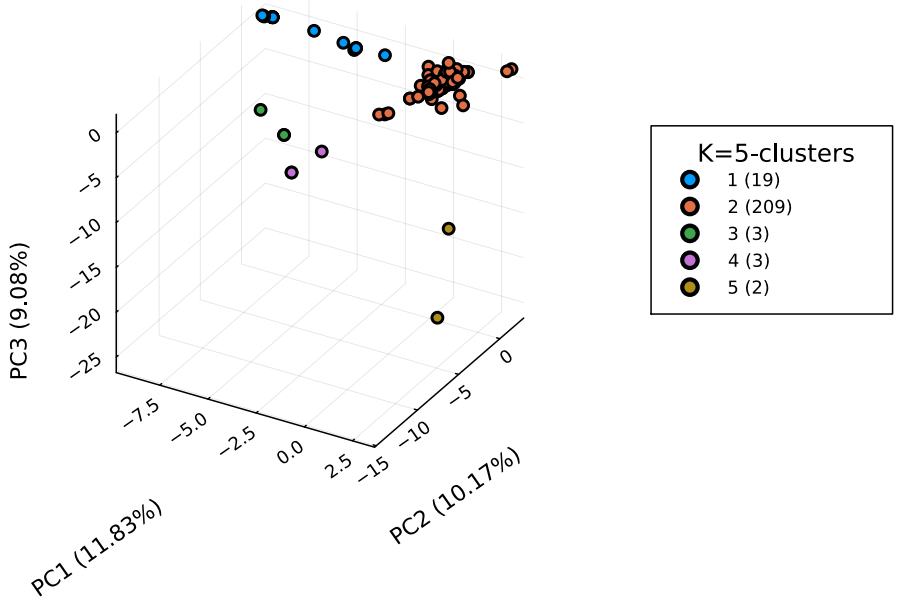


Figure S4: PCA scores for the first three components of the LC setups grouped by (a) USP column code and (b) k-means (k=5).

The first three principal components collectively explained 31.1% of the total variance in the LC setup metadata (PC1 = 11.83%, PC2 = 10.17%, PC3 = 9.08%). PC1 was primarily associated with stationary-phase selectivity, particularly the separation between dominant reversed-phase configurations (L1, i.e. C18) and zwitterionic HILIC columns (L122), together with contributions from flow rate and acidic mobile-phase modifiers (Figure S3). PC2 was mainly influenced by eluent composition and buffer selection, distinguishing aqueous/organic gradient systems commonly associated with RPLC workflows. PC3 further captured secondary variations related to alternative stationary phases and less frequently used organic modifiers.

K-means clustering (k = 5) was applied as an exploratory approach to identify recurring groups of chromatographic conditions within the RepoRT dataset. The resulting clusters reflected dominant methodological trends rather than discrete or strictly separated analytical categories. In particular, Cluster 2 represented the large majority of conventional RPLC methods (n = 209), while smaller clusters corresponded to less frequent stationary phases, alternative buffer systems, or unconventional solvent compositions. Because of the strong imbalance between RPLC and HILIC reporting frequencies within RepoRT, clustering results should be interpreted qualitatively as indicators of methodological convergence rather than as statistically independent classes. The centroid heatmap (Figure S5) was therefore included primarily to visualize recurring combinations of chromatographic parameters and their relative similarities across reported workflows.

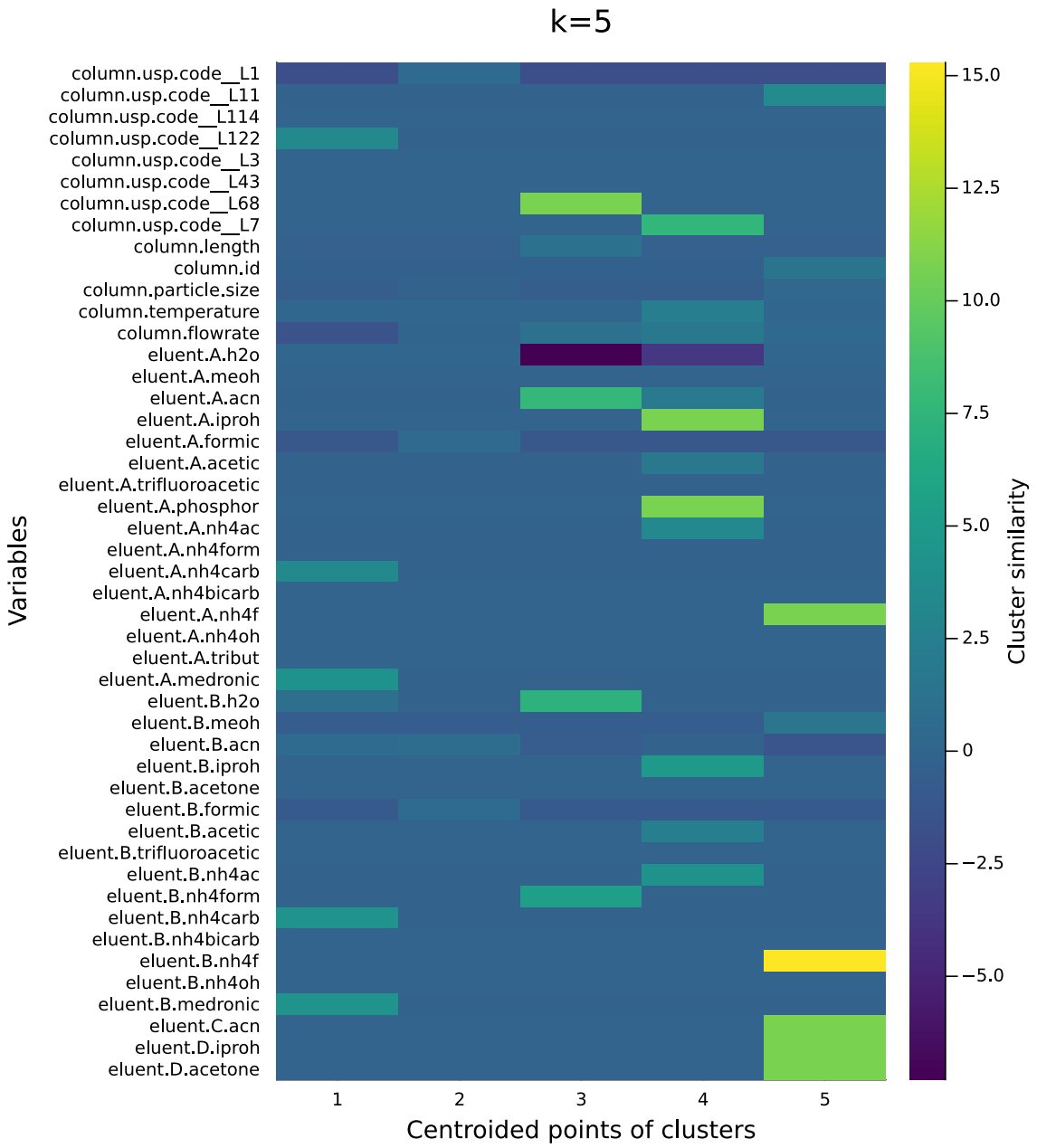


Figure S5: A heatmap of the most centroided points within each one of the five clusters (shown in Figure S4b) against all variables used to describe the LC setups reported in the RepoRT repository, highlighting the variation and similarity between these clusters

Compound metadata

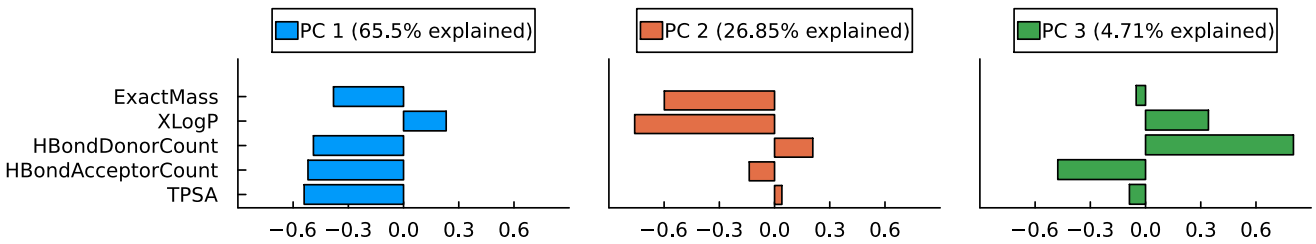


Figure S6: Bar plots of the loadings for the first three principal components used to describe the reported compound dataset.

The substantial overlap observed between RPLC and HILIC entries in the PCA score distributions should not be interpreted as evidence that the two chromatographic modes are intrinsically equivalent in separation capability. Inspection of the loading vectors indicated that PC1 was primarily driven by physicochemical descriptors associated with hydrophobicity and molecular size, including XLogP, exact mass, and topological polar surface area (TPSA), which collectively separated highly polar compounds from more hydrophobic analytes. PC2 showed stronger contributions from hydrogen-bond donor/acceptor counts and ionization-related descriptors, reflecting secondary polarity and interaction effects relevant to both chromatographic retention and electrospray ionization compatibility. Because many routinely analyzed compounds occupy intermediate regions of these descriptor spaces, substantial overlap between RPLC and HILIC score distributions was observed. It is worth mentioning that this overlap partially reflects also a combination of factors associated with current analytical practice, including the predominance of broad-gradient generic methods, the limited representation of highly optimized HILIC workflows within RepoRT, and reporting biases toward compounds routinely measurable by LC–HRMS.

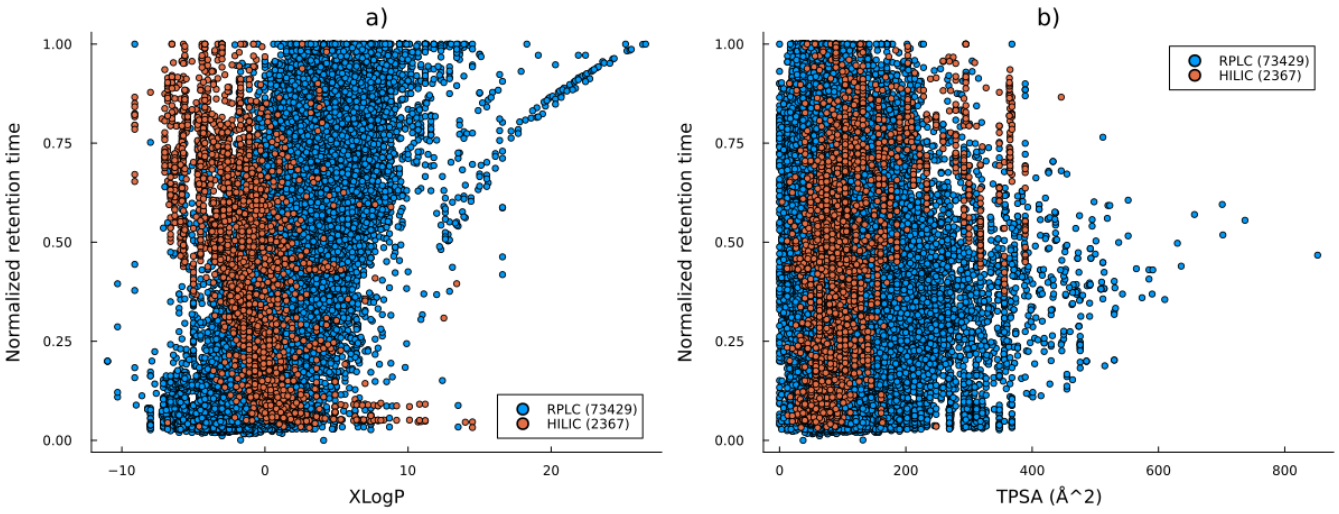


Figure S7: Trend of the normalized retention time vs (a) XLogP and (b) TPSA for the reported compounds, grouped by selectivity

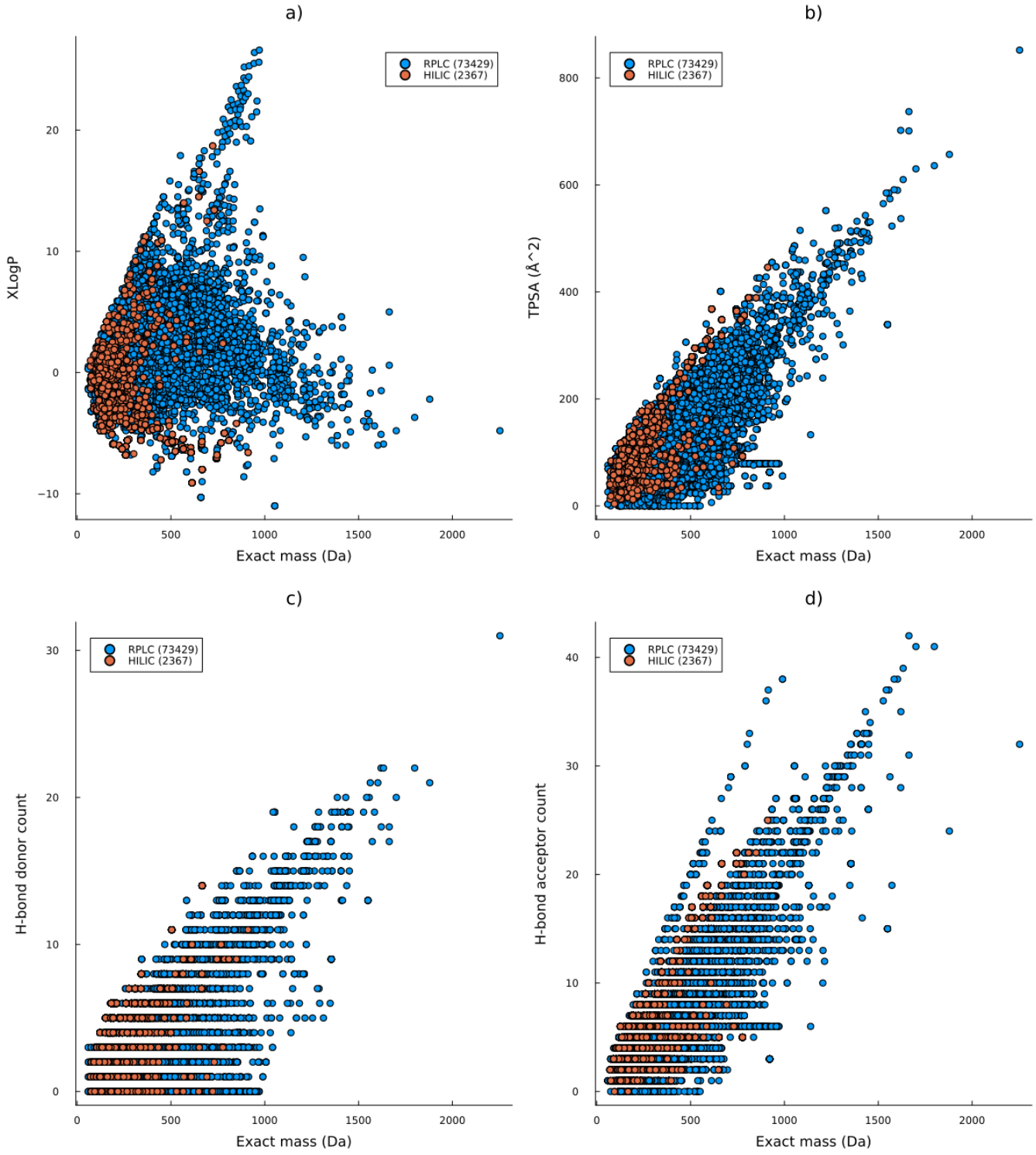


Figure S8: Exact mass distributions vs (a) XLogP, (b) TPSA, (c) H-bond donor and (d) H-bond acceptor counts for all reported compounds, grouped by selectivity.

RepoRT vs CompTox chemical coverage

The CompTox Chemicals Dashboard dataset was downloaded at `https://comptox.epa.gov/dashboard/chemical-lists` (release version 2.3, accession date: March 2024). The dataset was subsequently refined to retain unique, structure-resolved compounds with canonical SMILES and unique InChIKeys, excluding salts, mixtures, duplicate entries, and records lacking complete structural information. This curation resulted in a final dataset of approximately 800k unique chemical structures.

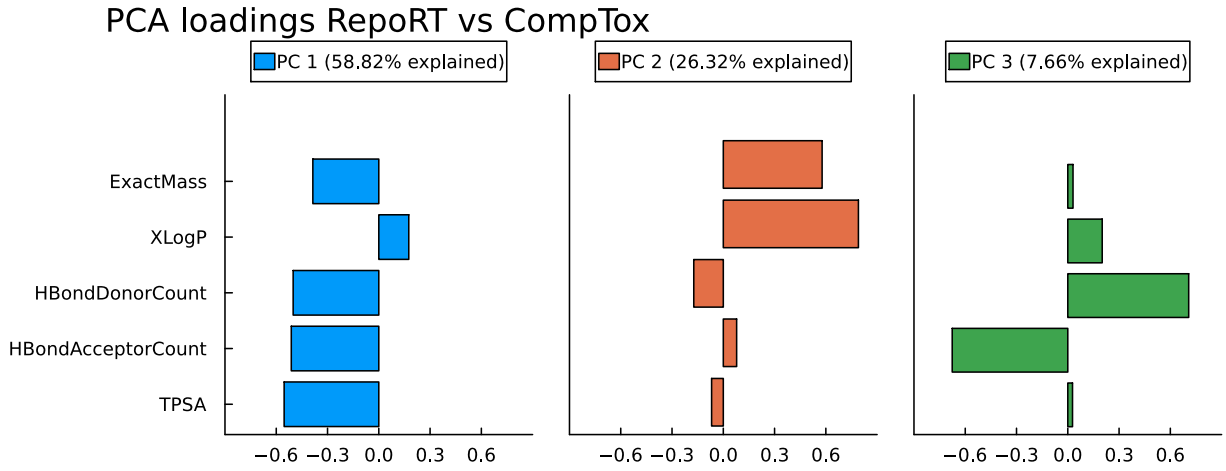


Figure S9: Bar plots of the loadings for the first three principal components used to describe the chemical coverage of the RepoRT compound dataset vs CompTox databse (≃800k chemical entries).