scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Retention time dataset for heterogeneous molecules in reversed–phase liquid chromatography

Yan Zhang[1,2,3], Fei Liu[1]✉, Xiu Qin Li[2,3], Yan Gao[2,3], Kang Cong Li[2,3] & Qing He Zhang[2,3]✉

Quantitative structure–property relationships have been extensively studied in the field of predicting retention times in liquid chromatography (LC). However, making transferable predictions is inherently complex because retention times are influenced by both the structure of the molecule and the chromatographic method used. Despite decades of development and numerous published machine learning models, the practical application of predicting small molecule retention time remains limited. The resulting models are typically limited to specific chromatographic conditions and the molecules used in their training and evaluation. Here, we have developed a comprehensive dataset comprising over 10,000 experimental retention times. These times were derived from 30 different reversed-phase liquid chromatography methods and pertain to a collection of 343 small molecules representing a wide range of chemical structures. These chromatographic methods encompass common LC setups for studying the retention behavior of small molecules. They offer a wide range of examples for modeling retention time with different LC setups.

Background & Summary

Liquid chromatography (LC) coupled to mass spectrometry (MS) is extensively used for both targeted and untargeted analysis in many fields[1–4]. LC−based separation aids in distinguishing isomeric and isobaric molecules, resulting in cleaner fragmentation spectra, and improves detection of low−abundance molecules by minimizing ionization competition[5]. In addition, chromatographic retention time (RT) provides crucial identification data, especially for molecules with indistinct mass/spectra but differing RTs[6]. However, the use of RT in identification workflows is often limited by the lack of reference standards and the inconsistent RT across different chromatographic methods (CMs), which affects the availability of comprehensive datasets.

Predicting RT for specific molecules within a given CM has become a popular alternative[7–9]. In untargeted metabolomics, the use of quantitative structure−retention relationship (QSRR) strategies predicts RT for potential candidates, reducing false positives[10,11]. However, the need for different QSRR models for different LC setups complicates this approach[12–16]. Strategies to address these complexities include using universal retention indices for different CMs[11,17], mapping RTs from one CM to another[5,18,19] and integrating chromatographic descriptors into QSRR models[20,21]. While current public datasets cover diverse CMs[22–24], the limited molecular overlap remains a challenge in modeling LC setups and their RTs.

We developed a dataset of Multiple Chromatographic Methods–based Retention Time (MCMRT), which contains over 10,000 experimental RT entries for a set of 343 small molecules from 30 different CMs[25]. These molecules were carefully selected to represent various chemical classes and exhibit a wide range of physicochemical properties, effectively mimicking the diverse chemical space encountered in reverse–phase (RP) LC analyses. The CMs were tailored to reflect common LC setups in untargeted analyses, incorporating different C18 columns, gradient profiles, mobile phases, additives, etc. The extensive molecular overlaps among the CMs

[1]Key Laboratory of Groundwater Conservation of MWR, China University of Geosciences, Beijing, 100083, People's Republic of China. [2]Division of Chemical Metrology and Analytical Science, National Institute of Metrology, Beijing, 100029, People's Republic of China. [3]Key Laboratory of Chemical Metrology and Applications on Nutrition and Health for State Market Regulation, Beijing, 100029, China. ✉e-mail: feiliu@cugb.edu.cn; zhangqh@nim.ac.cn

in MCMRT make it easier to transfer machine learning models between different LC setups, enhancing their practical applicability in predicting RTs under various chromatographic conditions.

Methods

Chemicals and regents. The 343 small molecules were sourced from various suppliers, including LGC Standards and Wellington Laboratories in Canada, Sigma–Aldrich in the USA, Dr. Ehrenstorfer in Germany, and several institutions in China like the National Institutes for Food and Drug Control and the National Institute of Metrology, alongside other providers such as Altascientific and J&K Scientific. These molecules were classified according to their ionization efficiency and chemical class. This classification facilitated their distribution into eight mixtures, with concentrations adjusted to span from 500 μg/mL to 20 mg/mL. The solubility and stability of each standard were taken into account when selecting suitable solvents for the preparation of these mixtures, which were then stored at −20 °C until analysis. Before use, the mixtures were combined to form a final mixture at a concentration of 20 μg/mL for all molecules. Exceptions were made for molecules with lower ionization efficiency, which received concentration adjustment to ensure their effective detection. Detailed information on the sourcing of these molecules and the methodology for mixture preparation is available in Table S1 (see supplementary xlsx file).

HPLC–grade reagents acetonitrile (ACN), methanol (MeOH), acetone and formic acid (FA) were supplied by Merck (Germany). Analytical–grade ammonium acetate and formate were purchased from Sigma-Aldrich (USA). Ultrapure water was prepared with a Milli-Q IQ 7000 system, also supplied by Merck (Germany).

Instrumental analysis. The analytical experiments were carried out using a Vanquish UHPLC system (Thermo Fisher Scientific, USA) coupled to an Orbitrap Q–Exactive Plus mass spectrometer (Thermo Fisher Scientific, USA) operated by TraceFinder V4.1 Software. All analyses were performed in both positive and negative ionization modes, utilizing a comprehensive full mass scan. The instrumental parameters were set as follows: two scan ranges covering 80–400 Da and 350–1600 Da, with a high resolution of 70,000; an Automatic Gain Control (AGC) target set as 1e6; a maximum injection time of 100 ms; sheath gas flow at 40; auxiliary gas at 8; and sweep gas at 1. The spray voltage was meticulously calibrated to 2.5 kV, with the heater temperature maintained at 350 °C and the Capillary Temperature at 250 °C, complemented by an RF–Lens setting of 60.

To ensure utmost precision in mass measurements, a tuning mix was injected at the onset of each CM run for calibration purposes. A detailed outline of the LC setups, including 30 different CMs, is provided in Table 1 and Table S2 (see supplementary xlsx file). All retention data for each CM were collected in a single day, with three replicate analyses.

Determination of retention time. The determination of RT was conducted using Xcalibur V4.3 software. First, the exact mass-to-charge (m/z) ratios of potential adducts for each molecule were calculated, e.g., $[M+H]^+$, $[M+H]^{2+}$, $[M+NH_4]^+$, $[M+Na]^+$, $[M-H]^-$, and $[M+HCO_2]^-$. Then, these adducts were used to extract the associated chromatographic peaks, allowing for a mass deviation of 5 ppm. To ensure accurate RT determinations, all RT values for the molecules were carefully determined through manual assessment. In cases where the m/z ratio of one adduct of a molecule (e.g., $[M+H]^+$) differed by less than 5 ppm from another adduct of a different molecule (e.g., $[M+Na]^+$), the RTs of all possible adducts were carefully combined to confirm the correct RTs. For isomeric or isobaric molecules, separate standard solutions for each molecule were analyzed to accurately determine their distinct RT values.

Data Records

Repository and data overview. The dataset is publicly accessible through the Science Data Bank[25] at https://doi.org/10.57760/sciencedb.15823. It is organized into 30.xlsx files, each corresponding to a unique CM run. Each file contains two worksheets. The first worksheet in each file is dedicated to RT data, where molecules are identified using isomeric SMILES strings encoded to represent their molecular structures. To ensure consistency, all SMILES strings adhere to the PubChem standardization procedure[26]. RT data for all observed molecules were recorded in MCMRT, including those with RTs close to the dead time. The RT values provided are the averages of three replicate analyses. Additionally, the relative standard deviation (RSD) between the three replicates is included to indicate method variability and support data quality. Furthermore, the repository offers extensive molecular data, including InChI codes, IUPAC names, MCMRT numbers, CAS numbers, PubChem numbers, and chemical formulas. The second worksheet provides comprehensive chromatographic information, including details on data sources, instruments used, analytical columns, temperatures, mobile phases, gradient profiles, runtimes, flow rates, and dead times used to calculate retention factors. Retention factors are also provided in the first worksheet. This thorough documentation ensures the dataset's robustness and utility for researchers.

Data description. The MCMRT repository currently houses 10,073 RT entries, encompassing 343 unique molecules and 30 different CMs. These CMs utilized RP columns, specifically six different C18 columns with varying dimensions (50–150 × 2.1–4.6 mm) and particle sizes (1.7–5 μm). Except for the Thermo Hypersil GOLD column (100 × 2.1 mm, 1.9 μm) and the Acclaim 120 C18 column (4.6 × 150 mm, 5 μm), all columns were new at the time of use. To ensure proper equilibration, two blank gradient runs were performed prior to each CM run. Among the published datasets[22], the most frequently utilized columns were the Waters ACQUITY UPLC BEH C18 and Waters ACQUITY UPLC HSS T3, both included in MCMRT. The gradient profiles were designed with both single and multi–slopes, employing either isocratic or gradient flow rates ranging from 0.2 to 1 mL/min. While constant flow rates are more common in RPLC, gradient flow rates were included to explore their potential effects on RTs. This approach was inspired by the work of Gago-Ferrero *et al.*[24] who introduced flow rate variations in their CMs, creating a widely used dataset for suspect and non-target screening of environmental samples[10,11,27]. Total run times for these methods varied from 10 to 100 min. The column temperatures were varied

CM	Molecules	Column specifications	Mobile phase A	Mobile phase B	Run Time (min)
1	335	ACQUITY PRIMER HSS T3 (100 × 2.1 mm, 1.8 μm)	90% H_2O + 10% MeOH + 0.01% FA + 5 mM NH_4COOH	MeOH + 0.01% FA + 5 mM NH_4COOH	10
2	335	Acclaim RSLC 120 C18 (100 × 2.1 mm, 2.2 μm)	90% H_2O + 10% MeOH + 0.01% FA + 5 mM NH_4COOH	MeOH + 0.01% FA + 5 mM NH_4COOH	15
3	335	Acclaim RSLC 120 C18 (100 × 2.1 mm, 2.2 μm)	90% H_2O + 10% MeOH + 0.01% FA + 5 mM NH_4COOH	MeOH + 0.01% FA + 5 mM NH_4COOH	21
4	335	Thermo Hypersil GOLD (100 × 2.1 mm, 1.9 μm)	90% H_2O + 10% MeOH + 0.01% FA + 5 mM NH_4COOH	MeOH + 0.01% FA + 5 mM NH_4COOH	21
5	335	ACQUITY BEH C18 (100 × 2.1 mm, 1.7 μm)	90% H_2O + 10% MeOH + 0.01% FA + 5 mM NH_4COOH	MeOH + 0.01% FA + 5 mM NH_4COOH	21
6	335	ACQUITY PRIMER HSS T3 (100 × 2.1 mm, 1.8 μm)	90% H_2O + 10% MeOH + 0.01% FA + 5 mM NH_4COOH	MeOH + 0.01% FA + 5 mM NH_4COOH	21
7	335	Acclaim RSLC 120 C18 (100 × 2.1 mm, 2.2 μm)	90% H_2O + 10% MeOH + 0.01% FA + 5 mM NH_4COOH	MeOH + 0.01% FA + 5 mM NH_4COOH	30
8	335	Acclaim RSLC 120 C18 (100 × 2.1 mm, 2.2 μm)	90% H_2O + 10% MeOH + 0.01% FA + 5 mM NH_4COOH	MeOH + 0.01% FA + 5 mM NH_4COOH	30
9	335	Thermo Hypersil GOLD (100 × 2.1 mm, 1.9 μm)	90% H_2O + 10% MeOH + 0.01% FA + 5 mM NH_4COOH	MeOH + 0.01% FA + 5 mM NH_4COOH	30
10	335	Acclaim RSLC 120 C18 (100 × 2.1 mm, 2.2 μm)	90% H_2O + 10% MeOH + 0.01% FA + 5 mM NH_4COOH	MeOH + 0.01% FA + 5 mM NH_4COOH	30
11	335	ACQUITY BEH C18 (100 × 2.1 mm, 1.7 μm)	90% H_2O + 10% MeOH + 0.01% FA + 5 mM NH_4COOH	MeOH + 0.01% FA + 5 mM NH_4COOH	30
12	335	Acclaim RSLC 120 C18 (100 × 2.1 mm, 2.2 μm)	90% H_2O + 10% MeOH + 0.01% FA + 5 mM NH_4COOH	MeOH + 0.01% FA + 5 mM NH_4COOH	45
13	335	ACQUITY PRIMER HSS T3 (100 × 2.1 mm, 1.8 μm)	90% H_2O + 10% MeOH + 0.01% FA + 5 mM NH_4COOH	MeOH + 0.01% FA + 5 mM NH_4COOH	60
14	335	ACQUITY UPLC HSS T3 (2.1 × 50 mm,1.8 μm)	90% H_2O + 10% MeOH + 0.01% FA + 5 mM NH_4COOH	MeOH + 0.01% FA + 5 mM NH_4COOH	60
15	335	Acclaim 120 C18 (4.6 × 150 mm, 5 μm)	90% H_2O + 10% MeOH + 0.01% FA + 5 mM NH_4COOH	MeOH + 0.01% FA + 5 mM NH_4COOH	60
16	335	ACQUITY PRIMER HSS T3 (100 × 2.1 mm, 1.8 μm)	90% H_2O + 10% MeOH + 0.01% FA + 5 mM NH_4COOH	MeOH + 0.01% FA + 5 mM NH_4COOH	100
17	335	Thermo Hypersil GOLD (100 × 2.1 mm, 1.9 μm)	H_2O + 0.1% FA + 4 mM NH_4COOH	MeOH + 0.1% FA + 4 mM NH_4COOH	14
18	335	Thermo Hypersil GOLD (100 × 2.1 mm, 1.9 μm)	H_2O + 0.1% FA + 4 mM NH_4COOH	MeOH + 0.1% FA + 4 mM NH_4COOH	21
19	335	Acclaim RSLC 120 C18 (100 × 2.1 mm, 2.2 μm)	90% H_2O + 10% MeOH + 0.01% FA + 5 mM NH_4COOH	ACN	21
20	330	ACQUITY PRIMER HSS T3 (100 × 2.1 mm, 1.8 μm)	H_2O + 0.1% FA	ACN + 0.1% FA	20
21	330	Thermo Hypersil GOLD (100 × 2.1 mm, 1.9 μm)	H_2O + 0.1% FA	ACN + 0.1% FA	21
22	330	ACQUITY PRIMER HSS T3 (100 × 2.1 mm, 1.8 μm)	H_2O + 0.1% FA	ACN + 0.1% FA	45
23	330	ACQUITY UPLC HSS T3 (2.1 × 50 mm,1.8 μm)	H_2O + 0.1% FA	ACN + 0.1% FA	45
24	330	Acclaim 120 C18 (4.6 × 150 mm, 5 μm)	H_2O + 0.1% FA	ACN + 0.1% FA	45
25	343	ACQUITY PRIMER HSS T3 (100 × 2.1 mm, 1.8 μm)	H_2O + 5 mM NH_4CH_2COOH	ACN + 5 mM NH4CH2COOH	20
26	343	ACQUITY UPLC HSS T3 (2.1 × 50 mm,1.8 μm)	H_2O + 5 mM NH_4CH_2COOH	ACN + 5 mM NH4CH2COOH	20
27	343	ACQUITY PRIMER HSS T3 (100 × 2.1 mm, 1.8 μm)	H_2O + 5 mM NH_4CH_2COOH	ACN + 5 mM NH4CH2COOH	20
28	343	ACQUITY UPLC HSS T3 (2.1 × 50 mm,1.8 μm)	H_2O + 5 mM NH_4CH_2COOH	ACN + 5 mM NH4CH2COOH	30
29	343	Acclaim 120 C18 (4.6 × 150 mm, 5 μm)	H_2O + 5 mM NH_4CH_2COOH	ACN + 5 mM NH4CH2COOH	30
30	343	Acclaim RSLC 120 C18 (100 × 2.1 mm, 2.2 μm)	H2O + 5 mM NH_4COOH	ACN + 5 mM NH_4COOH	21

Table 1. Chromatographic conditions, source and number of included molecules for CMs used in this study.

between 30 °C and 45 °C to optimize separation efficiency. Regarding the mobile phases, 18 CMs utilized a water/MeOH (90:10, v/v) mixture for mobile phase A, 12 utilized water for mobile phase A, 24 used MeOH for mobile phase B, and 6 chose ACN for mobile phase B. While ACN generally offers higher efficiency, we used MeOH in most CMs based on initial experiments indicating that RT variations were more influenced by additives than the solvent itself. This choice was also guided by the work of Gago-Ferreroa *et al.*[24], who used MeOH in their CMs. Preferred mobile phases included water with 0.1% formic acid (weak phase) and either acetonitrile or MeOH with 0.1% formic acid (strong phase)[22]. MCMRT also explores various mobile phase compositions, optimized

with different additives such as 0.01% formic acid with 5 mM ammonium formate, 0.1% formic acid with 4 mM ammonium formate, 0.1% formic acid, 5 mM ammonium formate, and 5 mM ammonium acetate. These mobile phase compositions were referenced from existing published datasets[11,14,23,24,28], facilitating the comparison and integration of new data with historical data for better understanding and utilization. An analysis of representative chromatographic parameters in the repository highlights the significant influence of column selection and mobile phase compositions on RTs and peak orders[29]. Detail information about the instrumental and chromatographic conditions are described in Table 1 and Table S2 (see supplementary xlsx file).

The molecules in MCMRT span diverse chemical classes and exhibit a broad range of octanol/water partition coefficients (log Kow −8.1 to 11.6) and molecular weights (89 to 1449 Da) (Fig. 1a). They encompass 11 ClassyFire groups at the superclass level[30], including benzenoids (27.7%), organic acids and derivatives (20.4%), organoheterocyclic compounds (18.7%), lipids and lipid-like molecules (9.9%), phenylpropanoids and polyketides (7.6%), organohalogen compounds (7.3%), organic oxygen compounds (3.5%), organosulfur compounds (1.2%), organic nitrogen compounds (1.2%), organophosphorus compounds (1.2%), and other compounds (1.5%). Figure 1b,c provide an overview of the elemental composition within these molecules, showcasing a diversity of elements (C, H, O, N, P, S, Cl, Br, F, and I). The METLIN dataset contains 80,038 molecules and covers seven similar superclasses[23]. Additionally, Gago-Ferreroa *et al.*'s dataset (referred to as CM 03 P) includes retention time data for 1820 emerging pollutants, such as pesticides, pharmaceuticals from different therapeutic categories, illicit drugs, industrial chemicals, and transformation products, representing a diverse set of chemical structures[24]. However, compared to these datasets, MCMRT includes some unique compound classes, such as organophosphorus flame retardants and perfluoro and polyfluoro organic compounds, which are absent in both the METLIN and CM 03 P datasets. METLIN focuses on metabolomics and aims to include molecules likely to be found in human samples, which explains the absence of certain classes. In contrast, MCMRT aims to provide broad coverage of chemical structures, including those not typically found in human samples. MCMRT also includes several pairs of isomers, further enhancing its utility in various analytical applications. A full list of these molecules is provided in Table S3 (see supplementary xlsx file), with their common name, IUPAC name, InChI, SMILES, PubChem number, CAS number, formula, Molecular Weight, predicted log K_{ow} and superclass.

Among the 343 diverse molecules in MCMRT, eight environmental hormones were detected exclusively in non-acidic mobile phases (CMs 25–30). These hormones include bisphenol A, bisphenol B, bisphenol F, 4-octylphenol, 4-nonylphenol, diethylstilbestrol, hexestrol, and estriol. These compounds primarily ionize in negative ion mode, exhibiting significant responses. The presence of acidic additives in mobile phases likely suppresses their ionization efficiency, resulting in detection limits not being met at the used concentration levels in acidic mobile phases (CMs 01–24). Additionally, five molecules were undetected in mobile phases containing solely acidic additives (CMs 20–24). Among these, one is an environmental hormone whose ionization efficiency may have been further reduced by the high concentration of 0.1% formic acid. The other four molecules—bromopropylate, permethrin, halfenprox, and bifenthrin—primarily responded as $[M+NH4]^+$ or $[M+Na]^+$ ions. In acidic mobile phases, their $[M+NH4]^+$ peaks were not detected, and their $[M+H]^+$ and $[M+Na]^+$ peaks were too weak to be detected. In contrast, the remaining 330 molecules were consistently detected across all CMs (Table S4, see supplementary xlsx file). This significant overlap enables cross-comparison and the study of retention behavior under various chromatographic conditions. Furthermore, MCMRT includes CMs that systematically vary a single chromatographic parameter, providing valuable insights into the effects of these variations. For instance, there are variations in column type between CM 04 and CM 05, mobile phase composition between CM 03, CM 19, and CM 30, running time between CM 01 and CM 13, and gradient profile between CM 09 and CM 10.

Overall, MCMRT serves as a crucial resource for exploring the complex relationship between LC setups and molecular RTs. With its comprehensive coverage of LC setups and systematic variations in chromatographic parameters, this resource is poised to significantly enhance the work of researchers who are exploring the optimization of LC methods or the development of predictive models that incorporate these chromatographic conditions. While replicating all setups may not be practical, MCMRT allows researchers to select the most relevant setups for their studies. This flexibility enables the evaluation of model performance across different chromatographic conditions, thereby enhancing the robustness and applicability of their models. This dataset is expected to play a crucial role in the methodological transition across diverse LC setups, providing valuable references for molecular behavior under various conditions. Such insights are crucial for making customized adjustments to methodologies. Furthermore, MCMRT is positioned to improve the accuracy and reliability of scientific work by enabling the cross-validation of methods, ensuring that the RTs of known compounds are consistent with those recorded in the dataset across different CMs. In its contribution to the broader field, MCMRT aims to promote methodological consistency and uniformity in data reporting by providing a benchmark for RTs across a range of CMs. This initiative is a step toward fostering a more integrated and collaborative scientific community, where shared knowledge leads to collective advancement.

Technical Validation

To ensure the accuracy of the resulting dataset, it was crucial to validate the experimental RTs for various molecules within each CM and to confirm the accuracy of RT relationships across different CMs. Initially, the experimental RTs in MCMRT for three CMs —CM 03, CM 11, and CM 21—were compared with data from other laboratories using the same CMs. Specifically, CM 11 was compared with CM 11 A (data from collaborating laboratory A, Table S5, see supplementary xlsx file), CM 21 was compared with CM 21B (data from collaborating laboratory B, Table S6, see supplementary xlsx file), and CM 03 was compared with CM 03 P (data from Gago-Ferreroa *et al.*[24], Table S7, see supplementary xlsx file). The results showed that CM 11 and CM 11 A had 335 overlapping molecules with RT deviations ranging from 0.03 to 1.21 min and an average deviation of

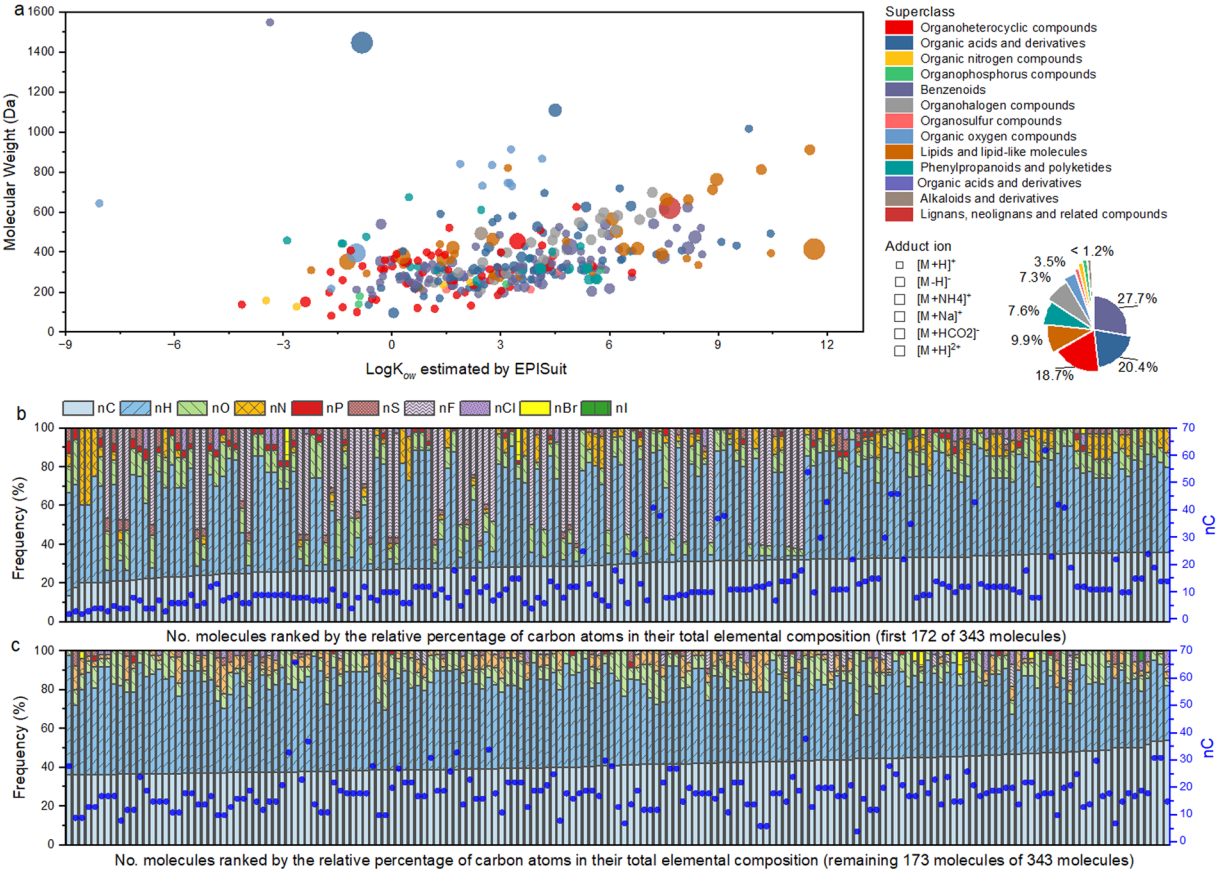

Fig. 1 Chemical diversity of molecules in MCMRT. (**a**) Molecular weight and log K_{ow} predicted by EPISuite for each molecule. Each data point corresponds to one molecule from the mixture; its color indicates the superclass defined by ClassyFire; its size indicates the adduct ion detected by ESI-HRMS. Panels (**b,c**) show the elemental composition of each molecule. Columns are aligned vertically for each individual molecule. The left axis represents the relative abundance of each element, while the right axis represents the absolute number of carbon atoms.

0.65 min (Fig. 2a). For CM 21 and CM 21B, there were 330 overlapping molecules with RT deviations ranging from 0.01 to 0.56 min and an average deviation of 0.14 min (Fig. 2b). CM 03 and CM 03 P had 154 overlapping molecules with RT deviations ranging from 0.01 to 1.21 min and an average deviation of 0.58 min (Fig. 2c). These findings indicate that the same molecules analyzed with identical CMs in different laboratories can result in different RTs. This discrepancy may be attributed to variations in chromatographic systems between laboratories or differences in column conditions. Importantly, the experimental RTs between these pairs of CMs exhibited strong correlations, with R^2 values ranging from 0.996 to 0.999. The correlation between CM 03 and CM 03 P was slightly lower ($R^2 = 0.996$), potentially due to discrepancies in molecule names provided by the online data source for CM 03 P, leading to mismatched RTs. In contrast, the data for CM 11 A and CM 21B were obtained from collaborating laboratories where the methods for determining RTs and defining molecule names were consistent, thereby avoiding such issues. This underscores the importance of standardized RT reporting to minimize discrepancies.

To reduce the low reproducibility of RT data caused by differences in column conditions and LC systems, the retention factor data for each molecule was also provided (Table S8, see supplementary xlsx file). The mathematical form of the retention factor k' is as follows:

$$k' = \frac{RT_x - RT_0}{RT_0}$$

where k' is the retention factor, RT_x is the RT of the molecule, and RT_0 is the dead time. In MCMRT, the RT of 4-Amino-1,2,4-triazole (MCMRT ID 001) was used as the dead time because it is typically not significantly retained on the column in RPLC.

To further enhance data usability and comparability between methods, a set of calibrants and a detailed calibration procedure were recommended in our previous publication[29]. The procedure involves measuring the RTs of the calibrants under both CMs and using these values to establish an RT projection model. For the pairs of CM 11 and CM 11 A, CM 21 and CM 21B, and CM 03 and CM 03 P, 35 molecules were randomly selected from the overlapping molecules based on their RT distribution to construct the RT projection models from CM 11

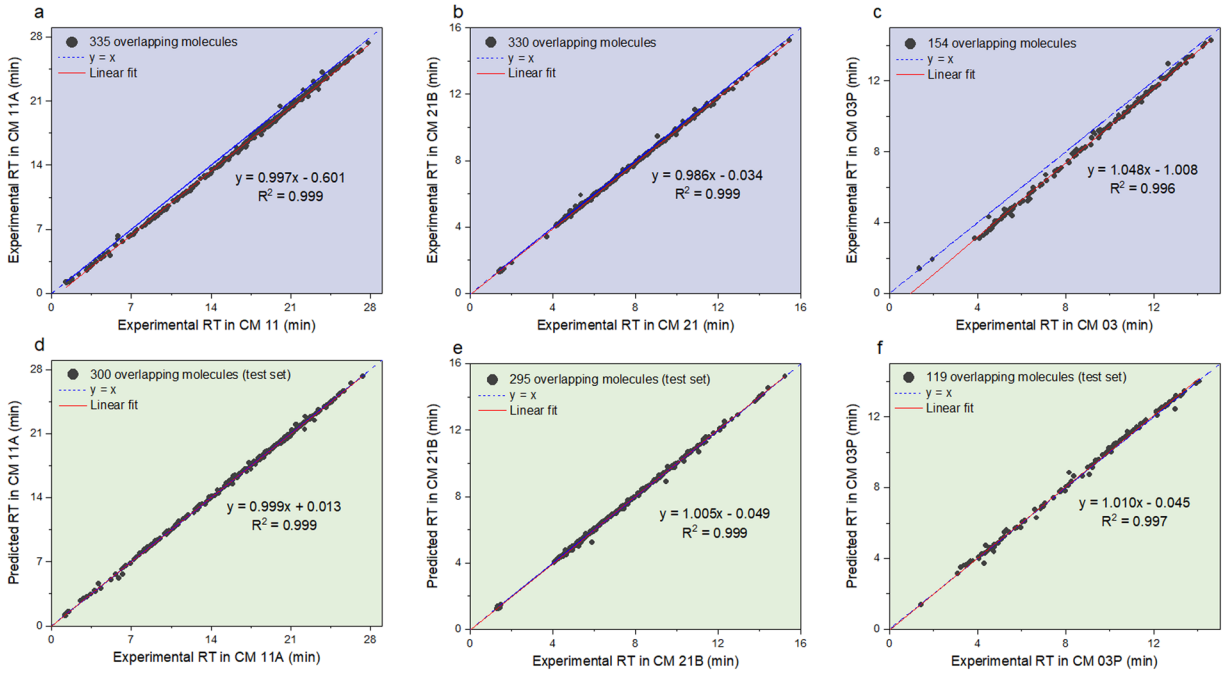

Fig. 2 Interlaboratory validation of retention time data. Panels (**a**–**c**) show the relationship between the experimental retention times of all overlapping molecules in the MCMRT (CM 11, CM 21, and CM 03) and non-MCMRT (CM 11 A, CM 21B, and CM 03 P) datasets. Panels (**d**–**f**) show the relationship between the predicted retention times and experimental retention times of overlapping molecules in the non-MCMRT datasets (CM 11 A, CM 21B, and CM 03 P) after applying the retention time projection model calibration from the MCMRT datasets (CM 11, CM 21, and CM 03).

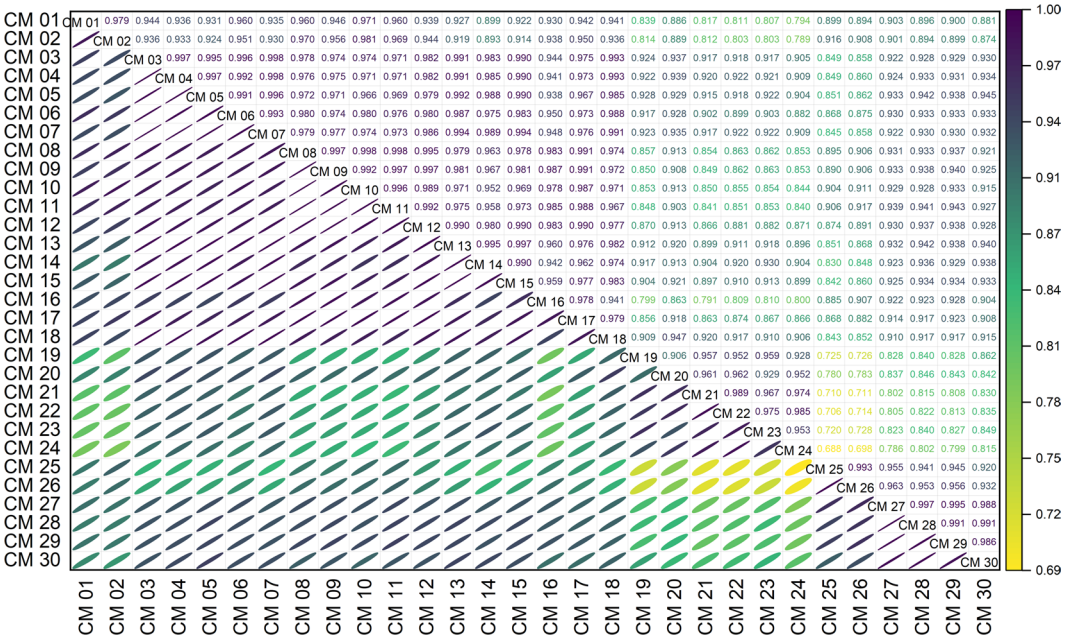

Fig. 3 Relationships of experimental retention times between two different CMs in MCMRT.

to CM 11 A (Fig. 2d), CM 21 to CM 21B (Fig. 2e), and CM 03 to CM 03 P (Fig. 2f). The remaining overlapping molecules' RTs in CM 11, CM 21, and CM 03 were then used to predict their RTs in CM 11 A, CM 21B, and CM 03 P. Comparing the predicted and experimental RTs in these CMs, ~85% of the prediction errors were less than 0.2 min (relative deviation of 3%). This demonstrates that the RT values in CM 11, CM 21, and CM 03 are largely accurate. Information on calibrants and predicted RTs can be found in Tables S5–S7 (see supplementary xlsx file).

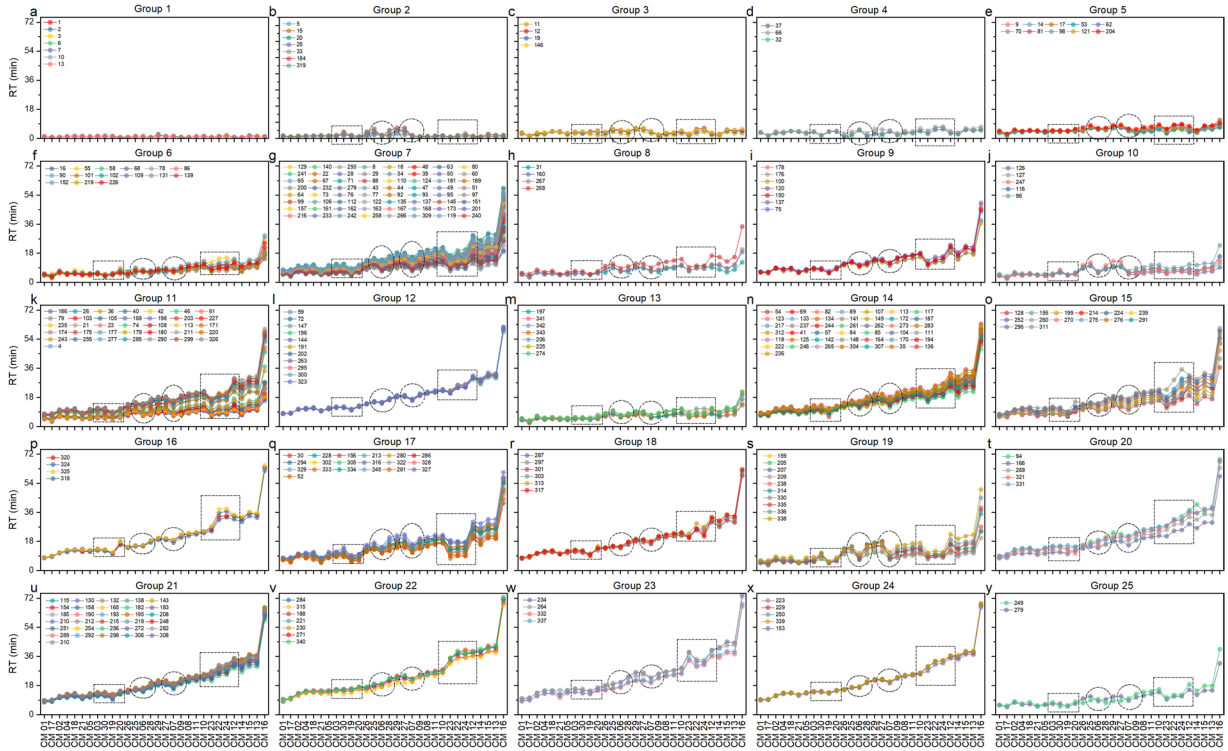

Fig. 4 Experimental retention times of 330 overlapping molecules in 30 different CMs. These molecules are divided into 25 groups based on their retention behavior (panels **a**–**y**). Each retention time profile corresponds to one molecule in MCMRT.

Next, the relationship between experimental RTs in different CMs was demonstrated using the overlapping set of 330 molecules within the MCMRT (Fig. 3). The R^2 values ranged from 0.688 to 0.999, depending on the similarity of LC setups. The R^2 values were highest (0.892–0.998) for CMs with identical mobile phase compositions and decreased slightly for CMs with similar compositions (0.906–0.993), different compositions (0.798–0.947), and very different compositions (0.698–0.860). These results confirm the accuracy of retention time relationships across the 30 CMs.

Finally, the 30 CM-specific datasets for the 330 overlapping molecules in MCMRT were analyzed using a self-organizing mapping (SOM) clustering algorithm, to characterize their retention behavior within each CM and between different CMs. The SOM clustering algorithm is an unsupervised machine learning technique that organizes data into clusters based on similarities, providing a visual and analytical means to detect patterns in high-dimensional datasets[22,31,32]. These molecules were categorized into 25 groups based on their RT variations across the 30 CMs (Fig. 4 and Table S4, see supplementary xlsx file). This categorization revealed distinct clusters of molecules with consistent retention behaviors, regardless of the mobile phase compositions, indicating robust retention properties for certain compound classes. Notably, some molecules exhibited stable retention times across various mobile phases, while others displayed noticeable shifts depending on the presence of specific additives such as formic acid or ammonium formate. This differentiation is crucial for understanding the impact of chromatographic parameters on molecular retention and highlights the need for diverse experimental setups to capture a comprehensive range of retention behaviors. These findings emphasize the importance of including a diverse array of molecules in the dataset to encompass multiple variations in RT. Further details on the methodology and results of this clustering analysis can be found in our previously published work[29]. These insights underline the necessity of comprehensive datasets that incorporate diverse molecular structures and chromatographic conditions to enhance the robustness and applicability of RT predictions.

In summary, our validation and analysis confirmed that the MCMRT dataset accurately determines RTs for a diverse array of molecules within each CM and captures precise RT relationships across different CMs[25]. The inclusion of heterogeneous molecules provides a comprehensive representation of RT variations, making the dataset a valuable resource for developing predictive models and enhancing the reliability of LC-MS analyses across various chromatographic conditions.

Code availability

The source code of RT projection and SOM clustering algorithm was provided in GitHub (https://github.com/Yanzi-Zhang-oss/Post-projection-calibration-of-retention-time-across-liquid-chromatography-setups).

Received: 1 February 2024; Accepted: 14 August 2024;
Published online: 29 August 2024

References

1. Zonja, B., Delgado, A., Pérez, S. & Barceló, D. LC-HRMS suspect screening for detection-based prioritization of iodinated contrast media photodegradates in surface waters. *Environ Sci Technol* **49**, 3464–3472 (2015).
2. Perez de Souza, L., Alseekh, S., Scossa, F. & Fernie, A. R. Ultra-high-performance liquid chromatography high-resolution mass spectrometry variants for metabolomics research. *Nat Methods* **18**, 733–746 (2021).
3. Giese, S. H., Sinn, L. R., Wegner, F. & Rappsilber, J. Retention time prediction using neural networks increases identifications in crosslinking mass spectrometry. *Nat Commun* **12**, 1–11 (2021).
4. Nikolopoulou, V., Aalizadeh, R., Nika, M. C. & Thomaidis, N. S. TrendProbe: Time profile analysis of emerging contaminants by LC-HRMS non-target screening and deep learning convolutional neural network. *J Hazard Mater* **428**, 128194 (2022).
5. Bouwmeester, R., Martens, L. & Degroeve, S. Generalized Calibration across Liquid Chromatography Setups for Generic Prediction of Small-Molecule Retention Times. *Anal Chem* **92**, 6571–6578 (2020).
6. Haddad, P. R., Taraji, M. & Szücs, R. Prediction of Analyte Retention Time in Liquid Chromatography. *Anal Chem* **93**, 228–256 (2021).
7. Randazzo, G. M. *et al*. Prediction of retention time in reversed-phase liquid chromatography as a tool for steroid identification. *Anal Chim Acta* **916**, 8–16 (2016).
8. Creek, D. J. *et al*. Toward Global Metabolomics Analysis with Hydrophilic Interaction Liquid Chromatography-Mass Spectrometry:Improved Metabolite Identification by Retention Time Prediction Darren. *Anal Chem* 8703–8710 (2011).
9. Kern, S., Fenner, K., Singer, H. P., Schwarzenbach, R. P. & Hollender, J. Identification of transformation products of organic contaminants in natural waters by computer-aided prediction and high-resolution mass spectrometry. *Environ Sci Technol* **43**, 7039–7046 (2009).
10. Aalizadeh, R., Nika, M. C. & Thomaidis, N. S. Development and application of retention time prediction models in the suspect and non-target screening of emerging contaminants. *J Hazard Mater* **363**, 277–285 (2019).
11. Aalizadeh, R. *et al*. Development and Application of Liquid Chromatographic Retention Time Indices in HRMS-Based Suspect and Nontarget Screening. *Anal Chem* **93**, 11601–11611 (2021).
12. Zapadka, M. *et al*. An application of QSRR approach and multiple linear regression method for lipophilicity assessment of flavonoids. *J Pharm Biomed Anal* **164**, 681–689 (2019).
13. Barron, L. P. & McEneff, G. L. Gradient liquid chromatographic retention time prediction for suspect screening applications: A critical assessment of a generalised artificial neural network-based approach across 10 multi-residue reversed-phase analytical methods. *Talanta* **147**, 261–270 (2016).
14. Bade, R. *et al*. Suspect screening of large numbers of emerging contaminants in environmental waters using artificial neural networks for chromatographic retention time prediction and high resolution mass spectrometry data analysis. *Science of the Total Environment* **538**, 934–941 (2015).
15. Feng, C. *et al*. Novel Strategy for Mining and Identification of Acylcarnitines Using Data-Independent-Acquisition-Based Retention Time Prediction Modeling and Pseudo-Characteristic Fragmentation Ion Matching. *J Proteome Res* **20**, 1602–1611 (2021).
16. Goryński, K. *et al*. Quantitative structure-retention relationships models for prediction of high performance liquid chromatography retention time of small molecules: Endogenous metabolites and banned compounds. *Anal Chim Acta* **797**, 13–19 (2013).
17. Albaugh, D. R. *et al*. Prediction of HPLC retention index using artificial neural networks and IGroup E-state indices. *J Chem Inf Model* **49**, 788–799 (2009).
18. Stanstrup, J., Neumann, S. & Vrhovšek, U. PredRet: Prediction of Retention Time by Direct Mapping between Multiple Chromatographic Systems. *Anal Chem* **87**, 9421–9428 (2015).
19. Low, D. Y. *et al*. Data sharing in PredRet for accurate prediction of retention time: Application to plant food bioactive compounds. *Food Chem* **357**, (2021).
20. Souihi, A., Mohai, M. P., Palm, E., Malm, L. & Kruve, A. MultiConditionRT: Predicting liquid chromatography retention time for emerging contaminants for a wide range of eluent compositions and stationary phases. *J Chromatogr A* **1666**, (2022).
21. Bride, E., Heinisch, S., Bonnefille, B., Guillemain, C. & Margoum, C. Suspect screening of environmental contaminants by UHPLC-HRMS and transposable Quantitative Structure-Retention Relationship modelling. *J Hazard Mater* **409**, (2021).
22. Kretschmer, F., Harrieder, E.-M., Hoffmann, M. A., Böcker, S. & Witting, M. RepoRT: a comprehensive repository for small molecule retention times. *Nat Methods* https://doi.org/10.1038/s41592-023-02143-z (2024).
23. Domingo-Almenara, X. *et al*. The METLIN small molecule dataset for machine learning-based retention time prediction. *Nat Commun* **10**, 1–9 (2019).
24. Gago-Ferrero, P. *et al*. Wide-scope target screening of >2000 emerging contaminants in wastewater samples with UPLC-Q-ToF-HRMS/MS and smart evaluation of its performance through the validation of 195 selected representative analytes. *J Hazard Mater* **387**, 121712 (2020).
25. Zhang, Y. *et al*. Retention time dataset for heterogeneous molecules in reversed–phase liquid chromatography[DS/OL]. V3. *Science Data Bank* https://doi.org/10.57760/sciencedb.15823 (2024).
26. Hähnke, V. D., Kim, S. & Bolton, E. E. PubChem chemical structure standardization. *J Cheminform* **10**, (2018).
27. Rostkowski, P. *et al*. The strength in numbers: comprehensive characterization of house dust using complementary mass spectrometric techniques. *Anal Bioanal Chem* **411**, 1957–1977 (2019).
28. Parinet, J. Prediction of pesticide retention time in reversed-phase liquid chromatography using quantitative-structure retention relationship models: A comparative study of seven molecular descriptors datasets. *Chemosphere* **275**, (2021).
29. Zhang, Y. *et al*. Generic and accurate prediction of retention times in liquid chromatography by post–projection calibration. *Commun Chem* **7**, 54 (2024).
30. Djoumbou Feunang, Y. *et al*. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J Cheminform* **8**, 1–20 (2016).
31. Ghaseminezhad, M. H. & Karami, A. A novel self-organizing map (SOM) neural network for discrete groups of data clustering. *Applied Soft Computing Journal* **11**, 3771–3778 (2011).
32. Ilbeigipour, S., Albadvi, A. & Akhondzadeh Noughabi, E. Cluster-based analysis of COVID-19 cases using self-organizing map neural network and K-means methods to improve medical decision-making. *Inform Med Unlocked* **32**, 101005 (2022).

Acknowledgements

The authors are grateful for financial support from the National Key Research and Development Program of China [2023YFF0612601] and the National Natural Science Foundation of China [grant number 41731282].

Author contributions

Y.G. and K.C.L. constructed the MCMRT dataset. Y.Z. wrote the manuscript. Q.H.Z., X.Q.L. and Y.Z. conceived the idea and designed the overall research. F.L. and Q.H.Z. supervised the whole project. All authors critically evaluated and approved the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-024-03780-5.

Correspondence and requests for materials should be addressed to F.L. or Q.H.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2024