Correspondence

https://doi.org/10.1038/s41592-023-02143-z

RepoRT: a comprehensive repository for small molecule retention times

Check for updates

Liquid chromatography (LC) is frequently used to separate metabolites and other small molecules. Retention time is the time required for a particular molecule to pass through the LC column. Prediction of retention times is one of the most investigated problems in the field of quantitative structure–property relationships[1]. However, transferable predictions are intrinsically complicated because retention times depend on both compound structure and the chromatographic system used. Despite four decades of development[1,2] and literally hundreds of published machine learning models, prediction of small molecule retention time is still far from everyday use. The resulting models are restricted to the chosen specific chromatographic conditions, but also by the molecules used in training and evaluation. Consequently, few computational methods use retention time for compound annotation, with at best moderate improvements in annotation power[3,4].

We present RepoRT (https://github.com/michaelwitting/RepoRT), a data repository for storing and retrieving retention time data as well as rich metadata on chromatographic conditions. RepoRT currently contains 373 datasets, 8,809 unique compounds, and 88,325 retention time entries measured on 49 different chromatographic columns using varying eluents, flow rates, and temperatures (Supplementary Note 1, Supplementary Fig. 1). Here, 295 datasets were measured on reversed-phase columns and 71 datasets on hydrophilic interaction chromatography (HILIC) columns. Column models that are regularly used in metabolomics[5] are also well covered in RepoRT. The recorded metadata include chromatographic columns, the composition of eluents and gradients, and temperatures. Automated workflows allow processing and standardization of input data. Notably, several datasets were imported from PredRet[6]. To ensure the longevity of RepoRT, we use GitHub for data storage (Supplementary Table 1). Version control allows data to be updated and corrected, and changes are tracked throughout this process.

We have put particular emphasis on making the data 'machine learning-ready'. First, RepoRT provides molecular structures in different formats, including standardized SMILES (simplified molecular input line entry system), isomeric SMILES and molecular fingerprints (Supplementary Note 2). Isomeric SMILES are required to distinguish between diastereomers, as these may differ in retention time (Supplementary Fig. 2). Second, we performed extensive manual curation for cleaning and completion of the available data (Supplementary Note 3). Third, RepoRT provides information about the column model used as real-valued vectors – Tanaka and hydrophobic subtraction model parameters[7,8] describing the column selectivity – allowing a machine learning model to generalize between different columns (Supplementary Note 4, Supplementary Fig. 3, Supplementary Tables 2 and 3). To improve data quality and reduce the number of mislabeled entries, we have developed and integrated data validation steps in the uploading procedure (Supplementary Note 5, Fig. 1e,g). Datasets in which individual chromatographic parameters are systematically varied are presumably highly informative for transferable machine learning. Consequently, we measured 54 such datasets (Supplementary Note 6, Supplementary Table 4). To avoid differences in spelling of column names, we collected more than 45,000 commercially available columns with lengths, particle sizes, inner diameters, and pore sizes (Supplementary Table 5). Finally, using GitHub allows us to tag a certain version of the repository, so that different machine learning models can be trained and evaluated on exactly the same data.

We analyzed the current coverage of columns, conditions and compounds in RepoRT (Supplementary Note 7). The sizes of the datasets range from 3 to 2,285 small molecules, with an average of 200 and a median of 81 compounds (Fig. 1c,d). Certain small molecules appear in a huge number of datasets, and 410 can be found in more than 50 datasets. Next, we examined how the small molecules from RepoRT cover the 'universe of small biomolecules'[9]. For both reversed-phase and HILIC column models, compounds in RepoRT already provide comprehensive coverage (Fig. 1a,b). Next, we examined the coverage for some of the largest datasets in RepoRT (Supplementary Fig. 4). The difference in coverage from the whole repository is apparent, illustrating the advantage of repository-scale training data. Frequently occurring compounds are spread across a large portion of the space of biomolecules (Supplementary Fig. 5). Finally, we examined the distribution of compounds in RepoRT using two other methods[9]. For compound classes, we observe that RepoRT almost perfectly covers the 'universe of known biomolecules' (Supplementary Table 6). Similarly, the natural product likeness score does not reveal bias of the data (Fig. 1f).

Numerous entries in RepoRT are currently lacking information on stereochemistry. This limits the potential of machine learning models, in particular for diastereomers. We invite providers of reference datasets to include this information in the future, or to update existing datasets. For reversed-phase columns, we argue that RepoRT already offers the data required to train transferable machine learning models. Unfortunately, the situation is worse for HILIC: not only is the column chemistry substantially more diverse, but also we do not have broadly available parameters to generalize between columns. The chromatographic separation in HILIC is driven by multiple types of interaction (partition, van-der-Waals, electrostatic interaction) and no clear separation mechanism can be established, making transferable predictions for HILIC challenging.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Data are freely available from the dedicated GitHub repository (https://github.com/michaelwitting/RepoRT). Source data are provided with this paper.

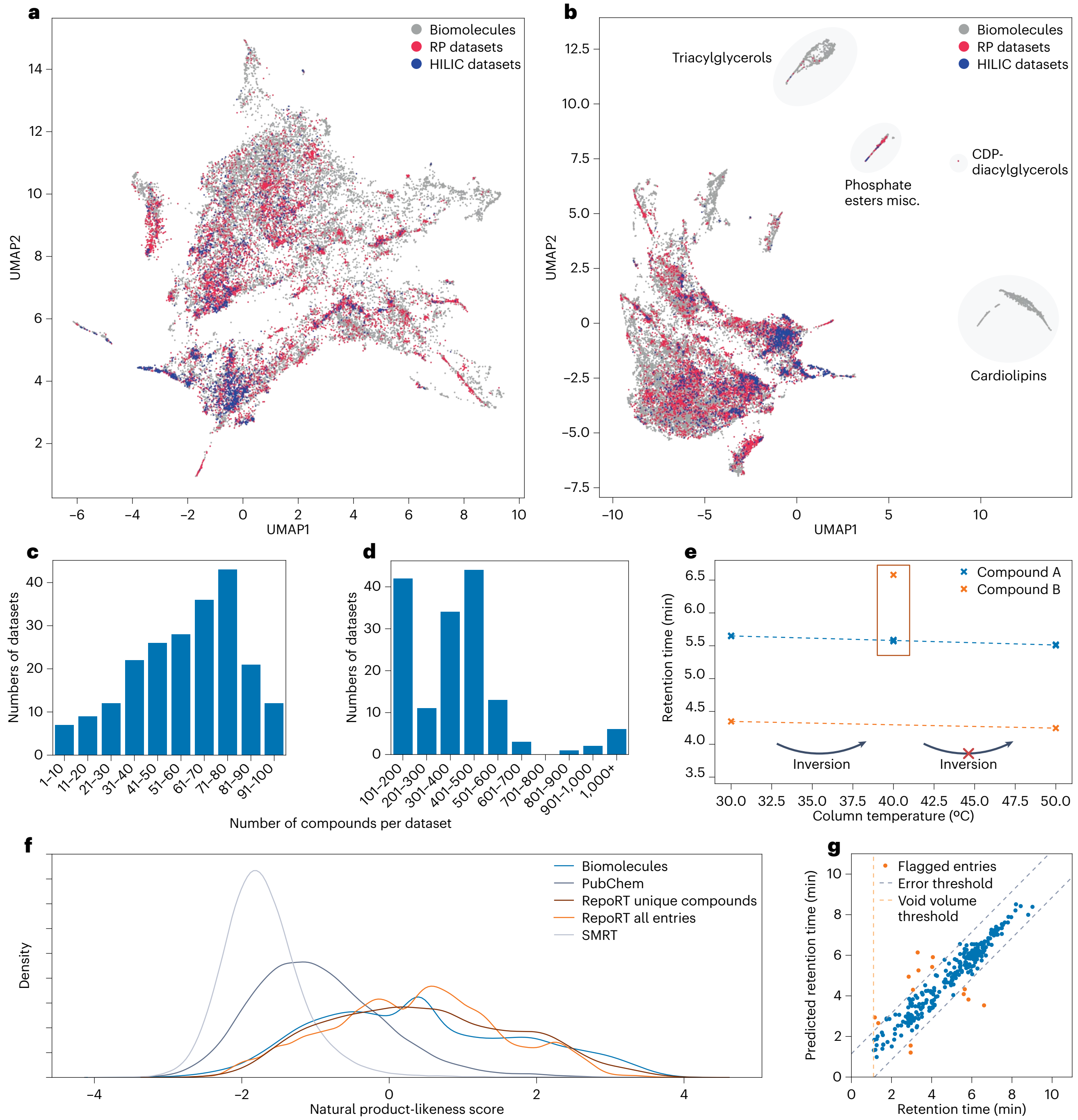

Fig. 1 | RepoRT, a repository for small molecule retention times. a,**b**, Uniform manifold approximation and projection (UMAP) plots of the coverage of molecular structures of biological interest. As described in ref. 9, molecular structures from the RepoRT repository are projected onto the space of known biomolecular structures. RepoRT structures are colored by column type (reverse phase (RP) versus HILIC). Structures from the small molecule retention time (SMRT) dataset are excluded[10] (Supplementary Fig. 1). Projections are shown excluding (**a**) and including (**b**) outlier lipid clusters. CDP, cytidine diphosphate. **c**,**d**, Distribution of the number of small molecule entries per dataset, excluding SMRT. **c**, Datasets that contain 1–100 compounds. **d**, Datasets that contain 100 or more compounds. The number of entries per dataset ranges between 3 and 2,285. **e**, Illustration of a data validation method, based on quantitative structure–property relationships models for an individual dataset. Entries with large errors between predicted and reported retention times in cross-validation are flagged for manual inspection. **f**, Distributions of natural product-likeness scores. We show kernel density estimates for compounds contained in RepoRT excluding SMRT, both for unique compounds and for all entries. For comparison, we show scores for biomolecular structures, PubChem structures, and SMRT. **g**, Illustration of a data validation method. For systematic measurements spanning multiple datasets, where a single chromatographic parameter is varied, double changes in retention order between compound pairs are detected.

Code availability

All code is freely available from the dedicated GitHub repository (https://github.com/michaelwitting/RepoRT).

Fleming Kretschmer [1,6], **Eva-Maria Harrieder**[2,6], **Martin A. Hoffmann** [1,5], **Sebastian Böcker** [1] ✉ & **Michael Witting** [2,3,4] ✉

[1]Chair for Bioinformatics, Institute for Computer Science, Friedrich Schiller University Jena, Jena, Germany. [2]Research Unit Analytical BioGeoChemistry, Helmholtz Zentrum München, Neuherberg, Germany. [3]Metabolomics and Proteomics Core, Helmholtz Zentrum München, Neuherberg, Germany. [4]Chair of Analytical Food Chemistry, TU München, Freising, Germany. [5]Present address: Bright Giant GmbH, Jena, Germany. [6]These authors contributed equally: Fleming Kretschmer, Eva-Maria Harrieder.
✉e-mail: sebastian.boecker@uni-jena.de; michael.witting@helmholtz-muenchen.de

Published online: 8 January 2024

References

1. Héberger, K. *J. Chromatogr. A* **1158**, 273–305 (2007).
2. Witting, M. & Böcker, S. *J. Sep. Sci.* **43**, 1746–1754 (2020).
3. Ruttkies, C., Schymanski, E. L., Wolf, S., Hollender, J. & Neumann, S. *J. Cheminformatics* **8**, 3 (2016).
4. Bach, E., Szedmak, S., Brouard, C., Böcker, S. & Rousu, J. *Proc. European Conference on Computational Biology* (ECCB 2018), i875–i883 (2018).
5. Harrieder, E.-M., Kretschmer, F., Böcker, S. & Witting, M. *J. Chromatogr. B* **1188**, 123069 (2022).
6. Stanstrup, J., Neumann, S. & Vrhovšek, U. *Anal. Chem.* **87**, 9421–9428 (2015).
7. Kimata, K. et al. *J. Chromatogr. Sci.* **27**, 721–728 (1989).
8. Snyder, L. R., Dolan, J. W. & Carr, P. W. *J. Chromatogr. A* **1060**, 77–116 (2004).
9. Kretschmer, F., Seipp, J., Ludwig, M., Klau, G. W. & Böcker, S. Preprint at *bioRxiv* https://doi.org/10.1101/2023.03.27.534311 (2023).
10. Domingo-Almenara, X. et al. *Nat. Commun.* **10**, 5811 (2019).

Acknowledgements

F.K., M.A.H., and S.B. are supported by Deutsche Forschungsgemeinschaft (BO 1910/23). F.K. and S.B. are supported by the Ministry for Economics, Sciences and Digital Society of Thuringia (Framework ProDigital, DigLeben-5575/10-9). E.-M.H. and M.W. are supported by Deutsche Forschungsgemeinschaft (MW 4382/10-1). We thank everybody who publicly shared their retention time data. We thank J. Büscher for providing and uploading unpublished data. We express particular thanks to J. Stanstrup, who allowed us to integrate data from PredRet into RepoRT.

Author contributions

S.B. and M.W. designed the research. F.K., M.A.H., and M.W. implemented the repository. E.-M.H. manually curated datasets. F.K. and S.B. developed methods for automated error detection. E.-M.H. measured the systematically varying chromatographic parameters of the datasets. F.K. and M.A.H. implemented methods. E.-M.H. and M.W. performed a statistical analysis of the repository content. F.K., E.-M.H., S.B., and M.W. wrote 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/s41592-023-02143-z.

Peer review information *Nature Methods* thanks Joshua Rabinowitz and Juho Rousu for their contribution to the peer review of this work.

nature portfolio

Corresponding author(s): Sebastian Böcker, Michael Witting

Last updated by author(s): Jul 16, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

☒ □ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement

☒ □ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☒ □ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☒ □ A description of all covariates tested

☒ □ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

☒ □ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

☒ □ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.

☒ □ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☒ □ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☒ □ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	For recording the 54 datasets where parameters were systematically varied, Agilent MassHunter by Agilent Technologies, Inc. was used. No further preprocessing software was used.
Data analysis	For bounded kernel density estimation, we used the R package 'bde'. For UMAP plots, we used the Python package 'umap'. For Natural Product-likeness score computations, we used RDKit.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Data is freely available from the dedicated GitHub repository https://github.com/michaelwitting/RepoRT.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	N/A
Data exclusions	N/A
Replication	N/A
Randomization	N/A
Blinding	N/A

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
☒	□ Clinical data
☒	□ Dual use research of concern

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging