RepoRT: A comprehensive repository for small molecule retention times

Fleming Kretschmer[1,6], Eva-Maria Harrieder[2,6], Martin A. Hoffmann[1,5], Sebastian Böcker[1,7], and Michael Witting[2,3,4,7]

[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] Currently at Bright Giant GmbH, Jena, Germany
[6] Shared first authors
[7] Corresponding authors, sebastian.boecker@uni-jena.de and michael.witting@helmholtz-muenchen.de

Abstract Liquid chromatography is frequently employed for the separation of metabolites and other small molecules. Prediction of retention times via machine learning methods can assist compound annotation. Yet, transferable predictions are intrinsically complicated for novel compounds and novel chromatographic conditions because retention times depend both on compound structure and the employed chromatographic system. We present RepoRT, the first repository for retention time data. RepoRT presently contains 373 datasets, 8809 unique compounds, and 88,325 retention time entries measured on 49 different chromatographic columns using varying eluents, flow rates, and temperatures. We put particular effort on making RepoRT "machine learning-ready": We performed an extensive manual curation, cleaning and completion of the available data; we developed automated methods for data validation during upload; we collected more than 45,000 different columns of different vendors with their lengths, particle sizes, inner diameters, and pore sizes to create a database with normalized column names; and, we ensured that features required for transferable predictions, such as parameters numerically describing the selectivity of chromatographic columns, are readily available. For version control and reproducible research, RepoRT is hosted on GitHub.

Introduction

Liquid Chromatography (LC) allows us to separate a complex chemical mixture into its components. It is frequently employed in the life sciences for the separation of molecules of biological interest, such as peptides, metabolites, or lipids. *Retention time* is the time a particular molecule requires to pass through the LC column and is dependent on the molecule structure, but also on the applied chromatographic conditions; clearly, this value carries information about the identity of the molecule (e.g. polarity based on the separation mode). Consequently, using machine learning for the prediction of retention time from the structure of a molecule is one of the most investigated problems in the field of Quantitative Structure-Property Relationships (QSPR)[1]. Whereas predicting retention times is relatively straightforward for polymeric biomolecules such as peptides[2] and oligonucleotides[3], it turned out to be an extremely challenging problem for so-called *small molecules* (meaning molecules with a mass smaller than 1000 Dalton). In the following, we will concentrate on small molecules of biological interest, or *biomolecules* for short: Small biomolecules include primary and secondary metabolites, drugs and drug degradation products, as well as toxic small molecules such as tetrodotoxin, the poison of the fugu fish. Analysis of endogenous and exogenous substances in biological samples or even biosystems is nowadays referred to as *metabolomics*.

Despite four decades of development[1,4,5] and literally hundreds of published machine learning models, retention time prediction for small molecules is still far from being in everyday use. The main problem is that retention time is not a property of a compound, but rather a combination of the employed chromatographic column, eluents and their composition, the compound itself, and further used equipment and experimental parameters. Usually, a publication on retention time prediction proceeds by first measuring a small dataset of at most a few hundred compounds on a particular, fixed experimental platform, then training and evaluating a machine learning model using these data. The resulting models are restricted to the chosen specific chromatographic conditions, but also by choice of molecules used in training and evaluation. In 2007, Héberger[1] collected more than 100 publications on the subject but concluded that "a prediction of retention data for not yet measured compounds would be a real gain." This assessment remains true in 2023, and even more so for novel chromatographic conditions.

What is the use of a machine learning model for small molecule retention time prediction that is not restricted to a particular chromatographic condition or a particular set of biomolecules? During the last decade, untargeted metabolomics has gained much momentum: Tandem MS (MS/MS) data from small molecules is annotated by comparison against spectral libraries, or by using so-called *in silico* methods for searching in molecular structure databases. Papers presenting *in silico* methods currently receive hundreds of citations per year[6,7]. However, according to different guidelines, high-confidence annotation or identification of a metabolite or small molecule requires an additional, orthogonal parameter beyond MS/MS[8,9]. Retention time is such an orthogonal parameter, providing information on the polarity of the small molecule. Notably, no generally agreed separation methods have been established in the metabolomics community, and this is extremely unlikely to happen in the future, either. Yet, as noted above, there exist no general models for predicting retention times from molecular structure under arbitrary chromatographic conditions. Consequently, retention time is usually ignored when searching public libraries or when using *in silico* methods. Instead, using retention time is often postponed to a late stage of identification, typically when putatively annotated features are compared against chemical reference standards. The few *in silico* methods that do use retention time for compound annotation report only moderate improvements, if any[7,10–13].

Training a machine learning model requires dedicated training data. Training a widely applicable model for small molecule retention time prediction necessitates going beyond the "single condition, 250 compounds" datasets mentioned above. Some publications combined a handful of datasets, or performed measurements systematically altering chromatographic conditions[14–17]. Some repositories for sharing mass spectrometry (MS) data also record retention times but usually handle this information as a "byproduct". Worse, even if retention time data are available, important chromatographic metadata are often missing, incomplete, or wrong[18]. Hence, data to train machine learning models for transferable prediction of retention times and order, are available in principle but not in practice. Until now, dedicated resources collecting retention time information in a systematic, "machine learning-ready" manner were nonexistent.

Results

We present RepoRT, a data repository for storing and retrieving retention time data as well as rich metadata on the chromatographic conditions. The RepoRT repository currently contains 373 datasets measured on different instruments, in different labs and by different experimentalists. Here, 295 datasets were measured on reversed-phase (RP) columns and 71 datasets on hydrophilic interaction chromatography (HILIC) columns. Few datasets use other column types such as pentafluorophenyl (PFP). Through the diversity of available data, our repository enables the training and evaluation of machine learning models beyond single datasets. Metadata includes chromatographic columns, the composition of eluents and gradients, and temperatures. Automated workflows allow processing and standardization of input data. To improve data quality and reduce the amount of mislabeled entries, we have developed and integrated data verification steps in the uploading procedure. We have put particular emphasis on making the data "machine learning-ready": For example, we provide information about molecular structures in different formats, including standardized SMILES (Simplified Molecular Input Line Entry System) and molecular fingerprints, and we provide information about the employed column model as real-valued vectors (Tanaka and hydrophobic subtraction model parameters[19,20] describing the column selectivity) that can be easily processed by and integrated into machine learning models. Additionally, InChIs (International Chemical Identifier) and InChI-Keys are provided for each small compound. To ensure the longevity of RepoRT, we use GitHub for data storage. Via version control, data can be updated and corrected, and changes are tracked throughout this process. Furthermore, we may tag a certain version of the repository, so that different machine learning models can be trained and evaluated on exactly the same data. This allows developers to compare evaluation results without the need to reevaluate existing models.

In June 2023, RepoRT contained 373 datasets from 16 contributors and several public retention data collections. Data collections include the METLIN's SMRT (small molecule retention time) dataset with 80,038 compounds measured under identical chromatographic conditions[21]; the plant compound datasets of Low *et al.*[16] spanning 24 chromatographic systems; the metabolite library datasets from Folberth *et al.*[22] comprising 57 chromatographic systems measured in positive and negative ion mode; seven metabolite library datasets from Pezzatti *et al.*[23] from four chromatographic systems; the metabolite library datasets from Stoffel *et al.*[24] measured on nine systems; 30 datasets with a metabolite library from Souihi *et al.*[17]; 18 unpublished datasets measured following the protocol from [25]; eight datasets with pesticides from Aalizadeh *et al.*[26]; five datasets from the National Phenome Centre's open platform for LC-MS-based metabolomics[27] spanning three chromatographic systems; four datasets with a metabolite and pesticide library from Huber *et al.*[28] measured on four chromatographic systems; three datasets with different type of lipids from Della Corte *et al.*[29]; and, sixteen dataset from various other publications[30–48]. If metadata was missing on these datasets, additional information was searched in the publications. Finally, 42 datasets were added from the PredRet[14] website (http://predret.org/) but cannot be assigned to an individual publication. Data were uploaded to the repository by the RepoRT maintainers or by the data contributors. If SMILES, InChI or InChI-Key were missing in the raw data, the missing information was again searched and added manually by the RepoRT maintainers. SMRT, being the largest dataset, is very different from all other datasets, not only because it is orders of magnitude larger: Notably, the distribution of molecular structures in SMRT is substantially different from that of both small molecules of biological interest, as well as the other RepoRT datasets, see [49] and Supplementary Fig. 10. In the following, we concentrate on the remaining 372 datasets.

Compounds and molecular structures. The RepoRT repository is organized into individual datasets, each representing one or more LC runs. In case a dataset contains more than one LC-MS run, all runs were measured on the same instrument under identical conditions, presumably using the same physical column or at minimum the same column type, length, diameter, etc., and in close temporal proximity. Each dataset contains information on the measured compounds (small molecules) and their retention times, plus the chromatographic metadata. Each dataset receives a unique RepoRT identifier (ID). Similarly, each compound in the data set receives a unique ID based on the dataset ID and a running number. The identity of the small molecule is minimally recorded via its molecular structure encoded in a SMILES string. RepoRT provides molecular structure information as both a canonical SMILES and an isomeric SMILES. The *canonical* SMILES only represents the molecular structure graph (atom types, connectivity, bond types) but no further structure information; the *isomeric* SMILES additionally provides information on stereochemistry. If only an isomeric SMILES is provided, it is straightforward to compute the corresponding canonical SMILES. Unfortunately, going from canonical SMILES to an isomeric SMILES is often impossible.

See below for a detailed discussion. To consistently represent molecular structures across the repository, all SMILES are standardized using the PubChem standardization procedure[50]. InChI and InChI-Key are also provided for each compound.

We provide additional information for each compound to simplify the prediction of retention times or retention order using the data in the repository; this information is automatically generated from the molecular structure. Firstly, we determine compound classes the compound belongs to, using the ClassyFire taxonomy[51]. Second, we determine and provide molecular descriptors such as XLogP and topological polar surface area using CDK (Chemical Development Kit)[52]. Third, we compute molecular fingerprints, specifically Molecular ACCess System (MACCS)[53], PubChem (CACTVS)[54], and extended-connectivity (ECFP)[55] fingerprints.

Stereoisomerism and canonical vs. isomeric SMILES. We noted above that two types of SMILES are used in RepoRT, namely, canonical and isomeric SMILES. Compounds can have the same constitution, meaning the same atoms and their connectivity, but different configurations, meaning different 3D arrangements of atoms; such compounds are called *stereoisomers*. Canonical SMILES only encode the molecular structure, whereas isomeric SMILES additionally contain information about the 3D arrangement. *Enantiomers* are compounds that are mirror images of each other; examples are L- and D-Tryptophan. Typically, enantiomers behave chemically identically and can only be separated using specialized chiral chromatographic setups. Such chromatographic setups are not used in today's metabolomics and are also not covered in RepoRT. All stereoisomers but enantiomers are *diastereomers*, and do not behave like mirror images of each other, since they do not differ at all stereocenters. An example are L-Isoleucine and L-Alloisoleucine; each of these diastereomers contains two stereocenters, but they only differ at one of them. A subclass of diastereomers are cis-trans isomers, such as oleic acid (cis) and elaidic acid (trans). Diastereomers may or may not be chromatographically separated, depending on the exact chromatographic setup. In particular, cis- and trans-isomers can usually be separated from each other, with the cis-isomers typically eluting first in RP chromatography. A more complicated example of stereoisomers are ursocholic acid, allocholic acid and cholic acid, see Fig. 1. There, changes in the stereochemistry at different positions cause a switch in the 3D structure and therefore, differences in retention times. Different stereoisomers (both enantiomers and diastereomers) have the same canonical SMILES and, hence, cannot be structurally distinguished on this level. Only isomeric SMILES allow us to refer to the correct chemical structure. However, not every compound contains stereocenters; examples are fatty acids such as palmitic acid. For such compounds, the isomeric SMILES is identical to the canonical SMILES.

For measurements, it is important to differentiate if the purchased chemical reference standard is isomerically pure or not. In particular, 1:1 mixtures of enantiomers are called *racemates*; these are considerably cheaper than stereochemically pure chemical standards, since their production and/or purification is less specific. For racemates, canonical SMILES should be used to avoid overreporting; recall that enantiomers cannot be separated by chromatographic systems covered in RepoRT. The same is true for cases of unclear stereochemistry.

For 8.15 % of the entries (compound plus retention time) in RepoRT datasets, the compound has no stereocenters. Next, 10.2 % of the entries correspond to enantiomers. In the following, we concentrate on diastereomers: We find that 81.6 % of RepoRT entries have two or more stereocenters, but only 30.6 % of those are annotated with an isomeric SMILES. Hence, 56.6 % of all entries in RepoRT are missing information potentially important for accurate prediction of retention time. This may be due to unclear stereochemistry of the measured reference compound. Since data in

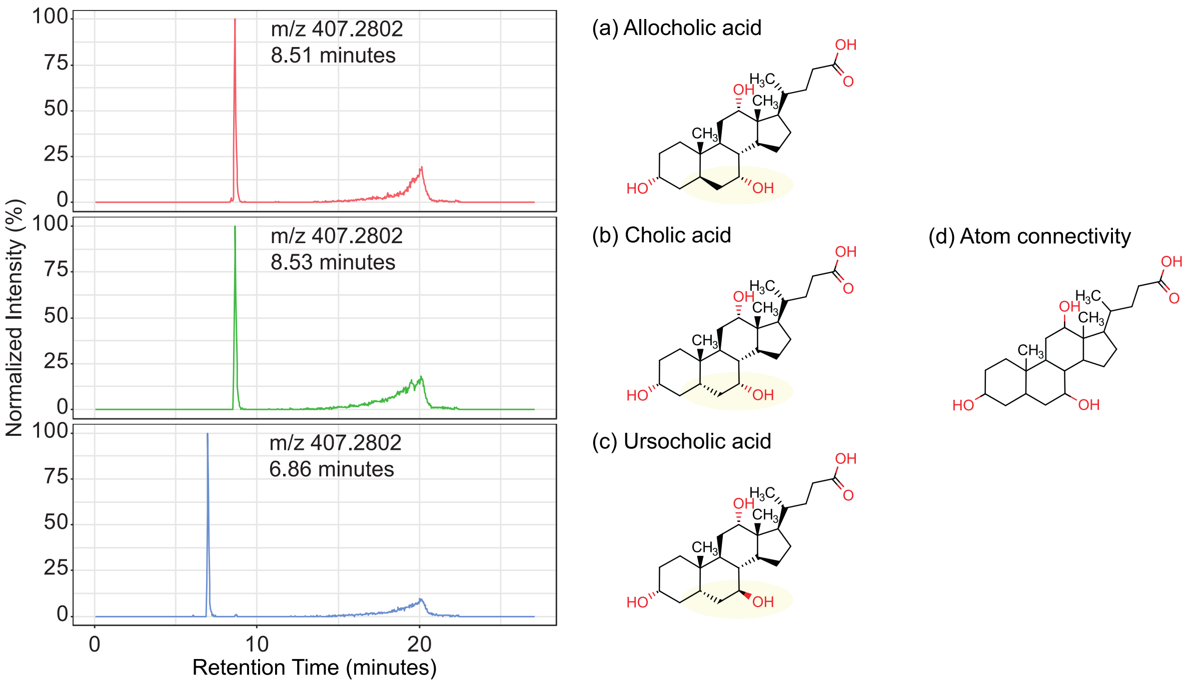

Figure 1. Different diastereomers can have similar or different retention times, depending on the actual 3D structure. As an example, the extracted ion chromatograms and structures of (a) allocholic acid, (b) cholic and (c) ursocholic acid are shown, measured on a Kinetex C18 column (dataset 0229 from RepoRT). Whereas the first two compounds have virtually the same retention time, the third compound elutes almost 2 minutes earlier. These three compounds have different 3D structures, hence different isomeric SMILES; observe the highlighted region. Yet, the atom connectivity is the same, yielding the same canonical SMILES (d).

RepoRT can be updated anytime, isomeric SMILES can and should be added at a later stage in case stereochemically pure standards were measured.

At present, information on stereochemistry must be handled with care both when training and evaluating models for retention time and order prediction. This is unfortunate since stereochemistry can have a substantial impact on retention time, compare to cis-trans isomers. Consequently, missing stereochemistry information limits the quality any machine learning model can possibly reach.

Doublets: Multiple entries for the same compound. For 6.54 % of all entries in RepoRT, two or more retention times were recorded in one dataset with identical SMILES. Here, “identical SMILES” refers to the SMILES string uploaded to the RepoRT database; it may be a canonical or an isomeric SMILES. In the following, we will call such entries *doublets*. (This does not include *duplicate entries* where the *same* retention time and SMILES are recorded in one dataset; those are removed from RepoRT, see Methods for details.) We find that 57.5 % of the doublets in RepoRT are diastereomers where only a canonical SMILES is reported. Here, a possible explanation is that two diastereomers were measured and recorded in the dataset, but that it was unclear from the data which diastereomer resulted in which peak in the elution profile. In cases where cis/trans isomers exist for a particular canonical SMILES, it is reasonable to assume that the cis-isomer elutes before the trans-isomer for reversed-phase separation, see above.

Somewhat surprisingly, for 28.9 % of the doublets, an isomeric SMILES was recorded. Similarly, for 13.6 % of the doublets, a canonical SMILES was recorded but the compound is an enantiomer or does not have any stereocenters. We are not aware of any chemical explanations for the existence of such doublets. We conjecture that a small share of these doublets is due to human errors when recording or uploading the data. Yet, we conjecture that the larger share of these doublets is due the employed experimental setup: In case LC-MS data but no tandem mass spectra are recorded, two features in the LC-MS may both have a mass highly similar to one of the reference compounds. In such cases, one cannot decide which of the features truly corresponds to the reference compounds; it is therefore a reasonable solution to upload retention times for both features.

We stress that doublets (flagged in RepoRT) must be ignored when evaluating the performance of machine learning models. They may be used to train models, though.

Chromatographic conditions, column models and parameters. For each dataset, we provide chromatographic metadata describing the separation system on which the reported retention times were measured. Minimal information is based on the recent report by Harrieder *et al.*[18], and comprises the chromatographic column, its dimensions, the gradient, temperature, flow rate, and the composition of eluents, see below. Further reported properties are particle size of the stationary phase, the United States Pharmacopeial (USP) classification code of the column, as well as estimations for both the void time (based on flow rate and column dimension) and the pH of each eluent. Generally, data from any chromatographic separation mode (e.g. RP, HILIC or other separation modes) can be submitted to RepoRT. To avoid that the columns with the same stationary phase are stored under different names, we have generated a list of standardized names for column models used within RepoRT (Supplementary Tables 2). Considering particle size, pore size, column length and inner diameter in addition to the column name, the list contains more than 25,000 individual column entries.

Integrating a column model via a one-hot encoding would not allow machine learning to generalize between column models. To enable generalization, we provide real-valued representations of the column's separation selectivity, namely Tanaka (six-dimensional) and hydrophobic subtraction model (HSM, five-dimensional) parameters[19,20]. Providing Tanaka and HSM parameters is only possible for column models where these parameters were determined experimentally and reported in the literature (Supplementary Tables 3 and 4). Notably, Tanaka parameters can vary depending on the column's particle size and, to a lesser extent, the pore size. In contrast, column length and inner diameter can be neglected since data is normalized to remove dependency on these factors. Unfortunately, fewer datasets can be assigned Tanaka parameters if we take into account particle size. We found that differences in Tanaka parameters are relatively small for different particle sizes: For those column models where Tanaka parameters are available for multiple particle sizes, the median Euclidean distance between the six-dimensional parameter vectors is 0.26, compared to a median of 3.19 for all column model pairs (Fig 2). To this end, we record Tanaka parameters of a column model with deviating particle size as a substitute in case of missing exact parameters. This fact is reported in RepoRT; corresponding datasets can be used to train models, but evaluations on such datasets should be handled separately.

Eluents are mixtures of water, organic solvents and chemical additives that are used to elute compounds from the stationary phase in the column. The compositions of the eluents are reported as the respective solvents and additives, and their amounts or proportions. Usually, there is an eluent with a high water content and an eluent with a high organic solvent content. Mixing two or more eluents together creates the mobile phase. The percentage of the eluents can be varied over time and is recorded in the *gradient*. In RepoRT, solvents are recorded in volume-percent, while the unit of additives can be freely defined but is typically denoted as percent or mM (millimolar). An example for a composition of an eluent is "90 % water vs. 10 % acetonitrile + 10 mM ammonium formate +

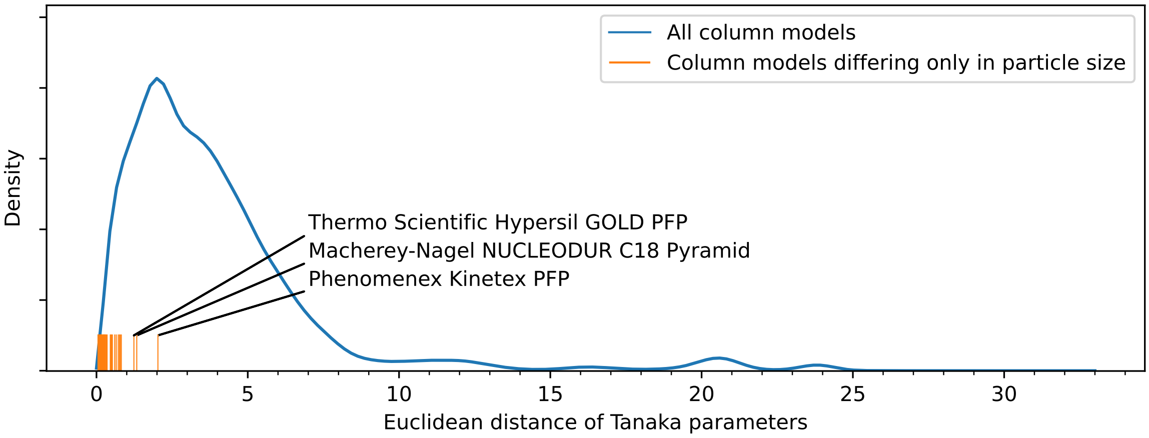

Figure 2. Distribution of difference between column Tanaka parameters. We show the bounded kernel density estimate of Euclidean distances between the six-dimensional Tanaka vectors. We use all 319 column models for which these parameters are available (50,721 pairs). For comparison, we show the exact Euclidean distances of all column model pairs with different particle sizes (38 pairs, orange bars). Only three of these column model pairs have Euclidean distance above 1, as labeled in the plot. The largest distance is observed for two Phenomenex Kinetex PFP columns with particle size 2.6 µm vs. 5 µm and equal pore size.

0.1 % formic acid". RepoRT covers numerous commonly used solvents and additives, but more can be added on demand.

Data validation. To ensure a high quality of the uploaded data and to minimize the number of wrongly annotated samples, we have implemented a series of protocols that allow the user to spot inconsistencies, which in turn may hint at bad data. This is of uttermost importance because our data is highly heterogeneous and stems from many sources. Mislabeling a few compounds or providing wrong chromatographic metadata can make the task of predicting retention time and order extremely challenging. It is understood that compounds flagged by any of the automated methods presented below must be verified manually.

A simple validation is as follows: If identical chromatographic conditions on the same LC-MS instrument in the same laboratory were used, then retention times between, say, positive and negative ionization modes should be very similar. For robustness, we do not verify retention times but rather retention order between pairs of compounds, considering only pairs of compounds with a predefined minimum retention time difference of 30 seconds.

Next, false annotations may be spotted using a simple QSPR model. Here, for an individual dataset, gradient boosting models are applied to predict retention times utilizing computed molecular descriptors as features. Ten-fold cross-validation is used to obtain a predicted retention time for each measured small molecule in a dataset. For each compound, the deviation (prediction error) between the estimated and measured retention time is calculated. Compounds with deviations substantially larger than the expected deviations are flagged as potential false annotations (Fig. 3 a). We have deliberately chosen gradient boosting as a QSPR model with a known tendency to overfit, because dataset-specific bias will also be learned by this model. In doing so, measurements strongly deviating from the biased retention time distribution of a dataset can be detected. However, false positives have to be expected, because a dataset may simply contain a few "outlier structures", having distinct structural features making them unique compared to the rest of the dataset.

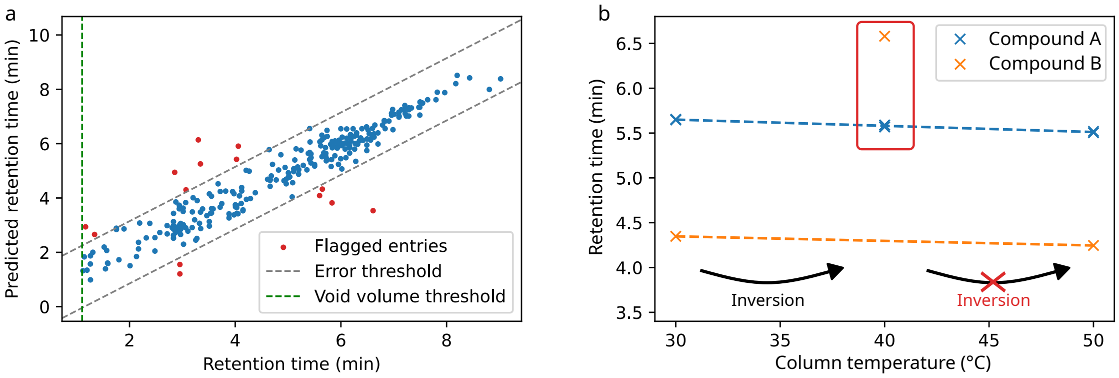

Figure 3. Illustration of two data validation methods. (a) Validation based on QSPR models for individual datasets. Entries with large error between predicted and reported retention times are flagged for manual inspection. (b) For systematic measurements spanning multiple datasets, where a single chromatographic parameter is varied, double changes of retention order between compound pairs are detected. Again, these may hint at wrong annotations. In the example, column temperature is varied, whereas there are multiple measurements and datasets for each temperature value.

For systematic measurements that alter only single parameters, such as temperature or flow rate (see below), a particular type of data validation can be performed. For example, an increase in flow rate causes a decrease in retention times. Column temperature also has an effect on retention behavior, which is compound-specific. When any such parameter is changed, the retention order of two compounds A, B may change. Assume that compound B elutes after A at column temperature 30 °C, but A after B at 40 °C. Now, we have no reasonable explanation how elution order can be again B after A at column temperature 50 °C. Further raising the column temperature should increase the effect we have previously observed, not reverse it (Fig. 3 b). Hence, this is likely the consequence of a false annotation or a wrongly entered value for the retention time. Accordingly, RepoRT allows checking systematic measurements for such double order inversions.

When uploading a dataset, a report is generated. This report contains overview plots indicating the flow rate, gradient and distribution of metabolites over the retention time range. Furthermore, tree maps of the ClassyFire classification on the kingdom, superclass and class level are generated as an overview on the chemical classes covered in the dataset. This information can help track problems and errors in the uploaded data.

Systematic variation of chromatographic parameters. When populating the repository, we realized that datasets that systematically vary a single chromatographic parameter were practically missing. As mentioned above, such datasets may be informative for a machine learning model that incorporates chromatographic conditions into its predictions. In order to fill this gap, we measured a set of small molecules, systematically varying chromatographic parameters. In detail, we varied the column model (six columns including Waters ACQUITY BEH C18), the organic solvent in the eluents (acetonitrile and methanol), and the temperature (30, 40, and 50°C). Flow rate and gradient remained the same for all datasets. See Methods for details. Overall, six column models, two eluent systems, and three temperatures would result in 36 conditions. Yet, three column models were measured only with a single eluent system (acetonitrile), so only 27 conditions were considered. In total, 1097 different small molecules were distributed into 41 mixes and measured in positive and negative ionization modes. This resulted in 54 datasets. Of the 1097 total compounds, 876 could be detected in at least one dataset.

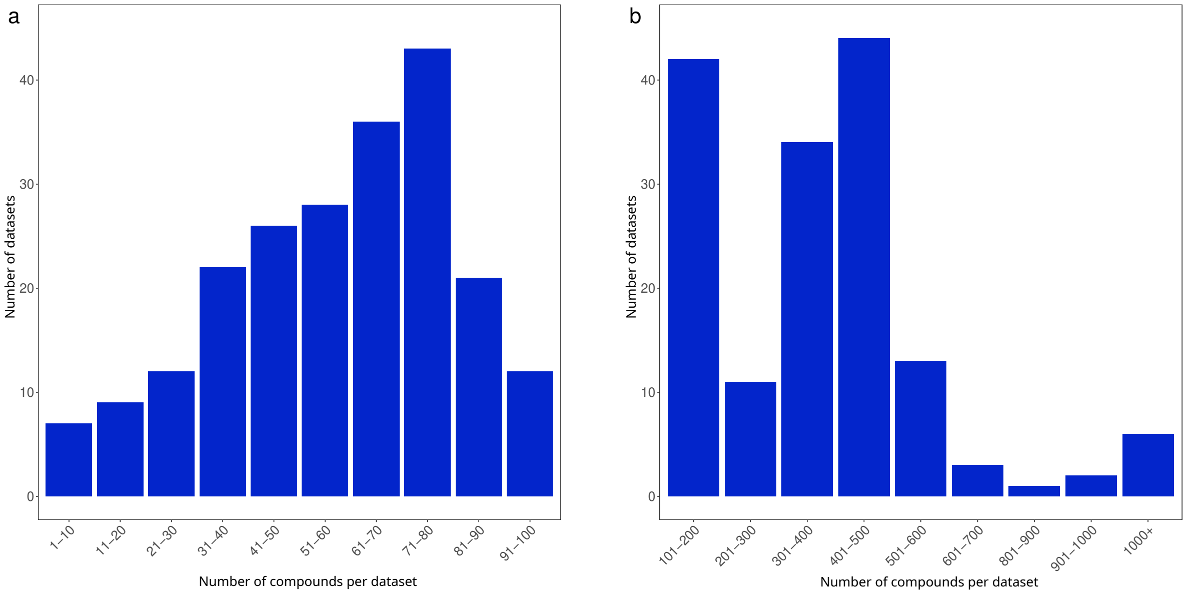

Figure 4. Distribution of the number of small molecule entries per dataset. (a) Datasets that contain 1 to 100 compounds and (b) datasets that contain 100 and more compounds. The range of entries per dataset is between 3 and 2285. The SMRT dataset with 80,038 compounds was excluded from this statistics.

For the 27 positive ionization mode datasets, 770 small molecules were detected. No compound was detected in all 27 conditions. We attribute this to the fact that three datasets contain less than 35 detected compounds. Yet, 228 small molecules were detected in at least 24 datasets. Notably, 47 compounds were only detected in a single dataset. In negative mode, 620 compounds were detected. Only nine small molecules are present in all 27 datasets. Again, one condition showed a low number of detected compounds (less than 90). We find that 124 compounds have been detected in at least 26 negative ion mode datasets, while 50 compounds were only detected in one dataset.

Current coverage of columns, conditions and compounds. To allow machine learning models to learn and predict retention time and order for both diverse molecular structures and diverse chromatographic conditions, we must provide this diversity in the training data, too. If a model is trained using data that does not cover the application space, then we cannot expect that the model will give practically useful predictions[49]. Hence, a particular emphasis of RepoRT is put on the diversity of compounds and conditions in the repository. Clearly, any dataset contains small molecules measured under identical chromatographic conditions. We conjecture that to train a broadly applicable model, one also needs the same compounds measured under different chromatographic conditions, as discussed above. Yet, beyond the variation of a single chromatographic parameter, we also need fundamentally different chromatographic setups differing in column model and mobile phase composition.

In total, 88,325 unique small molecules are covered across all 373 datasets of RepoRT, if we include the SMRT dataset. Here, we concentrate on the remaining 372 datasets covering 8809 unique small molecules. Sizes of the datasets range from 3 to 2285 small molecules, with an average of 200 and a median of 81 compounds (Fig. 4). Five datasets contain less than 10 compounds, while six datasets contain more than 1000 small molecules. Of all unique small molecules, 3775 are found only in a single dataset, 1651 in exactly two datasets, 1481 in three to five datasets, and 616 in six to ten datasets. Notably, certain small molecules appear in a huge number of datasets: 289 small

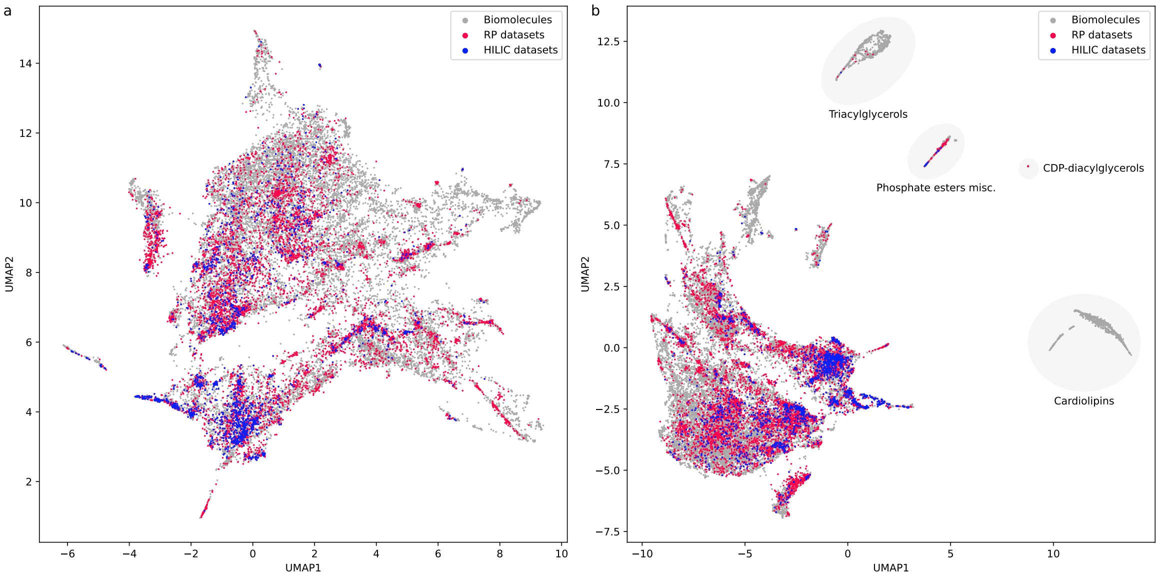

Figure 5. UMAP plot of the coverage of molecular structures of biological interest. As described in [49], molecular structures from the RepoRT repository are projected onto the space of known biomolecular structures. RepoRT structures are colored by column type (reverse phase vs. hydrophilic interaction chromatography). Structures from SMRT dataset excluded, see Supplementary Fig. 10. Projections are shown excluding (a) and including (b) outlier lipid clusters. Excluded lipid clusters are highlighted and labeled with the most characteristic ClassyFire compound class, except for the cluster of phosphate esters, mostly containing Glycerophosphoethanolamines, Phosphocholines, and Diacylglycerophosphates.

molecules are found in 11 to 20 datasets, 602 in 21 to 50 datasets, and 395 can be found in more than 50 datasets. These “ubiquitous” small molecules, such as kynurenic acid or adenine, may allow a model to transfer predictions between highly different chromatographic conditions.

Next, we examined how the small molecules from RepoRT cover the “universe of small biomolecules”[49]. For this, we prepared a Uniform Manifold Approximation and Projection (UMAP) plot of all molecular structures in RepoRT, embedded into 20,000 biomolecular structures. We use Maximum Common Edge Subgraph (MCES) distances for the layout of the plot. We observe that both for RP and HILIC column models, compounds in RepoRT already provide a very reasonable coverage of small biomolecules (Fig. 5). Recall that the SMRT dataset is excluded, see Supplementary Fig. 10. Examining the largest well-annotated datasets of RepoRT other than SMRT (Fig. 6), the coverage of biomolecules is relatively balanced and homogeneous across most areas of the UMAP visualization. Importantly, though, the difference in coverage to using the whole repository is evident, illustrating the advantage of a large collection of diverse datasets. Frequently occurring ("ubiquitous") compounds, likely crucial for transferring between chromatographic setups, are also spread across a large portion of the space of biomolecules (Fig. 7).

Finally, we examined the distribution of compounds in RepoRT using the two other methods from [49]. For *compound classes*, we observe that RepoRT almost perfectly covers the “universe of known biomolecules” (Table 1). Following Kretschmer *et al.*[49], we assume that a particular compound class can be learned from the data if sufficient training examples for this compound class are present; Kretschmer *et al.* suggested a threshold of 15 compounds. If we consider classes where at least 5 % biomolecular structures are part of the class, then fewest examples can be found for class “tricarboxylic acids and derivatives” with 116 training examples. Yet, even when lowering this threshold to 1 % biomolecular structures, only five of 311 compound classes have less than 15 training examples. In short, the coverage of compound classes in RepoRT appears to be excellent.

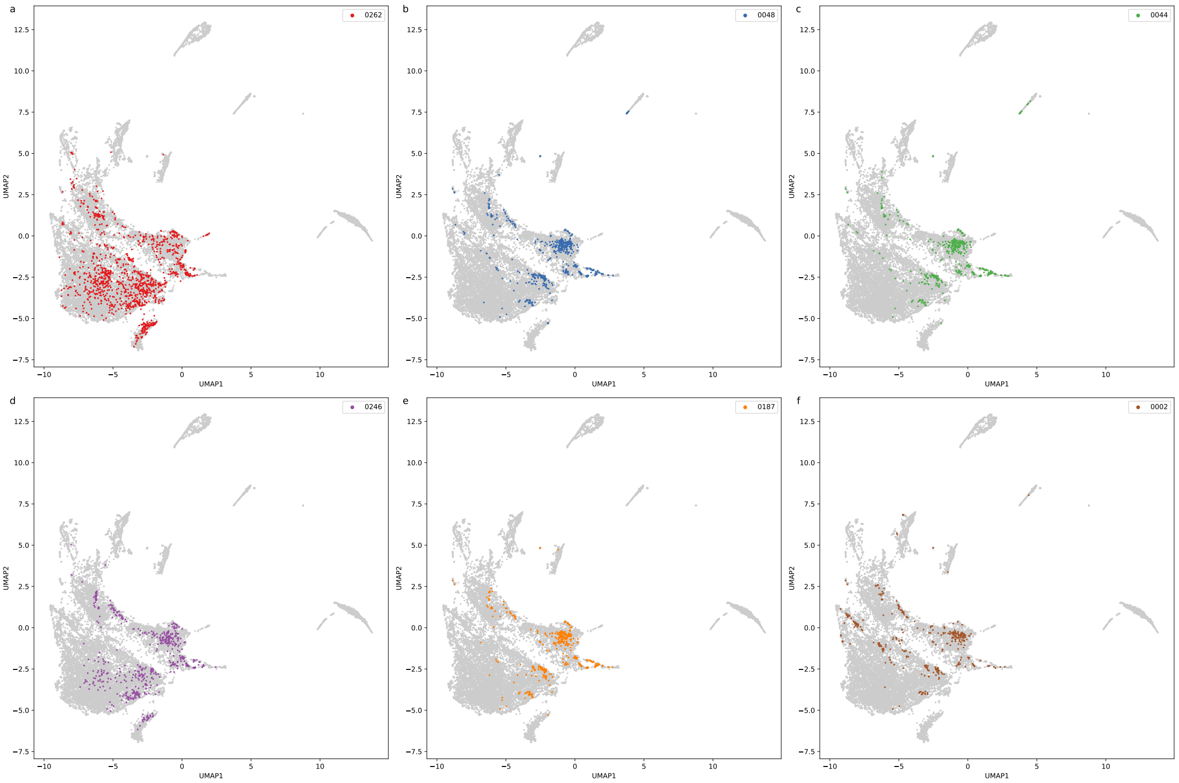

Figure 6. Coverage of biomolecules of the largest RepoRT datasets Visualization of RepoRT's coverage of molecular structures of biological interest based on Fig. 5 b, showing the six largest datasets individually (a–f in descending order of dataset size), excluding SMRT (see Supplementary Fig. 10). We consider only datasets that contain a minimum defined set of metadata (column name, temperature, and flow rate). For related datasets of systematic measurements, we chose the largest dataset. Shown are datasets 0262 from [28] (a), 0048 and 0044 from PredRet[14] (b,c), 0246 measured for this publication (d), 0187 from [24] (e), and 0002 from [43] (f).

Similarly, the *Natural Product-likeness score* does not reveal any lopsidedness of the data, either (Fig. 8). Notably, Kretschmer *et al.*[49] found that several datasets frequently used for small molecule machine learning show pronounced imbalances.

How do different datasets overlap in the small molecules they contain? Pairwise overlap was calculated as the Jaccard index, dividing the cardinality of the intersection by the cardinality of the union of small molecules. Fig. 9 shows the corresponding heat map. The highest overlap is observed between systematic datasets from the same submitters. However, a certain overlap is also observed between seemingly unrelated datasets.

We analyzed the most frequently used columns (not differentiating by particle size) in the repository (Supplementary Table 5). Clearly, the used column can have a massive impact on retention times and retention order. For RP, a total of 35 columns are found in RepoRT. The most frequently used column is Waters ACQUITY UPLC BEH C18 (63 datasets), followed by Waters CORTECS T3 (38 datasets) and Waters ACQUITY UPLC HSS T3 (24 datasets). Yet, 170 RP datasets use other column models. C18 columns are used in 88.8 % of the RP datasets; other columns use stationary phase chemistries such as octyl silane. For 29 of the 35 RP column models, HSM parameters are available, covering 232 (78.6 %) of the RP datasets. Tanaka parameters are available for 28 RP columns. Recall that Tanaka parameters can vary depending on the column's particle size. When

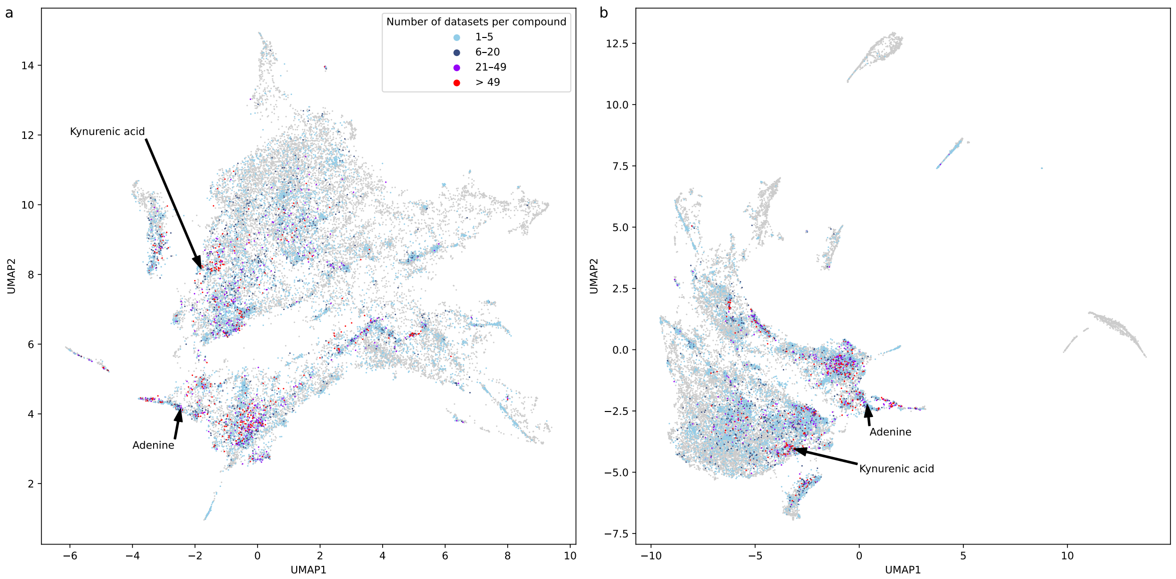

Figure 7. Frequency of RepoRT-compounds Visualization of RepoRT's coverage of molecular structures of biological interest based on Fig. 5, color-coding the number of datasets containing each compound. The positions of the compounds with the highest number of occurrences, adenine and kynurenic acid, are highlighted. Projections excluding (a) and including (b) outlier lipid clusters are shown.

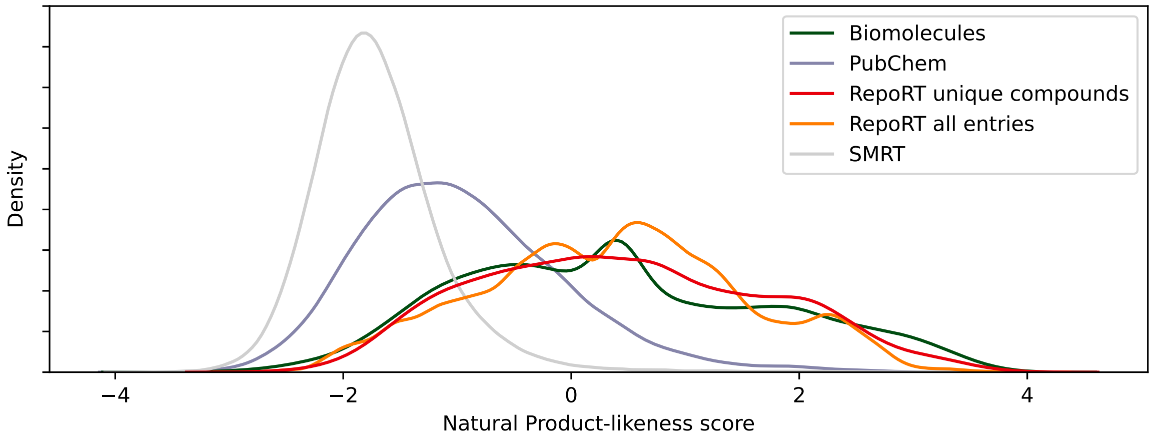

Figure 8. Distributions of Natural Product-likeness scores. We show kernel density estimates for Natural Product-likeness scores of compounds contained in RepoRT excluding SMRT, both for unique compounds and for all entries. For comparison, we show scores of biomolecular structures and PubChem structures, as well as the SMRT dataset. For biomolecules and PubChem, we subsample 20k molecular structures. In general, high scores indicate similarity to natural products.

taking particle sizes into account, Tanaka parameters can be assigned for 130 (44.1 %) of the datasets. If we consider the column name but ignore particle size, 227 (76.9 %) of datasets are covered.

For HILIC, 10 columns are present in RepoRT, and the most commonly used columns are Waters XBridge BEH Amide (14 datasets), Thermo Scientific Accucore HILIC (14) and Merck SeQuant ZIC-HILIC (12). These three columns represent the three different chemistries of the stationary phases in HILIC, which are alkyl amide-, silica- or sulfobetaine-based. Yet, 31 HILIC datasets use other

Table 1. Compound class coverage. The number of unique compounds in RepoRT ("Comp.") per compound class is shown, considering all ClassyFire compound classes occurring in at least 5 % of biomolecular structures ("Biomol."). The above analysis is not restricted to one class per rank, but rather takes into account the complete ClassyFire ontology. For details, see [49]. We also use the color-coding from there, and since all compound classes contain more than 15 examples, they are all colored green. Five of the 8809 unique compounds contained in RepoRT for which ClassyFire classes could not be computed, were discarded.

Class	Biomol.	Comp.
Chemical entities	100 %	8804
Organic compounds	99.89 %	8801
Hydrocarbon derivatives	99.28 %	8779
Organic oxygen compounds	94.02 %	8183
Organooxygen compounds	92.22 %	7940
Organic oxides	80.52 %	6775
Organic acids and derivatives	70.25 %	6015
Carbonyl compounds	63.73 %	4864
Organoheterocyclic compounds	63.62 %	4846
Carboxylic acids and derivatives	59.57 %	4520
Benzenoids	58.97 %	4747
Organic nitrogen compounds	56.95 %	5658
Organonitrogen compounds	56.84 %	5646
Carboxylic acid derivatives	51.03 %	3023
Organopnictogen compounds	45.98 %	5150
Benzene and substituted derivatives	43.10 %	3533
Azacyclic compounds	39.59 %	3344
Ethers	39.57 %	2520
Oxacyclic compounds	36.92 %	2269
Alcohols and polyols	35.58 %	3024
Heteroaromatic compounds	34.40 %	2770
Lipids and lipid-like molecules	34.03 %	2405
Carboxylic acid esters	30.51 %	1478
Secondary alcohols	27.14 %	2412
Amines	26.84 %	3262
Fatty Acyls	23.33 %	1659
Alkyl aryl ethers	22.52 %	1295
Monocarboxylic acids and derivatives	21.43 %	2584
Carboxylic acid amides	20.83 %	1484
Amino acids, peptides, and analogues	19.48 %	1927
Amino acids and derivatives	19.45 %	1927
Phenol ethers	19.22 %	1141
Ketones	16.84 %	1134
Organohalogen compounds	16.51 %	1419
Phenols	14.82 %	1301
Anisoles	14.76 %	749
Secondary carboxylic acid amides	14.73 %	1050
Fatty acid esters	14.68 %	532
Phenoxy compounds	14.54 %	1037
Phenylpropanoids and polyketides	13.52 %	1049
Primary alcohols	13.30 %	1298
1-hydroxy-2-unsubstituted benzenoids	12.92 %	1192
Carboxylic acids	12.29 %	2237
Polyols	12.01 %	1248
Aryl halides	11.98 %	1001
Tertiary amines	11.69 %	1110
Dialkyl ethers	11.61 %	606
Prenol lipids	11.29 %	504
Oxanes	10.84 %	925
Acetals	10.65 %	842
Organosulfur compounds	10.45 %	1032
Carbohydrates and carbohydrate conjugates	10.34 %	1185
Azoles	10.19 %	926
Alpha amino acids and derivatives	9.84 %	1127
Lactones	9.66 %	383
Pyrans	9.42 %	538
Methoxybenzenes	9.42 %	507
Lactams	9.19 %	584
Benzopyrans	9.16 %	591
Pyridines and derivatives	9.08 %	583
Trialkylamines	9.03 %	961
Aralkylamines	8.87 %	937
Halobenzenes	8.84 %	693
Cyclic alcohols and derivatives	8.83 %	684
1-benzopyrans	8.78 %	578
Benzoyl derivatives	8.63 %	642
Glycosyl compounds	8.38 %	938
Organochlorides	8.27 %	871
Secondary amines	8.06 %	732
Dicarboxylic acids and derivatives	7.99 %	721
Primary amines	7.94 %	1296
Tertiary carboxylic acid amides	7.62 %	427
Tertiary alcohols	7.48 %	497
Pyranones and derivatives	7.22 %	406
1-hydroxy-4-unsubstituted benzenoids	7.22 %	738
Aryl ketones	7.12 %	441
Organic 1,3-dipolar compounds	7.08 %	523
Monosaccharides	6.96 %	881
Organic phosphoric acids and derivatives	6.86 %	434
O-glycosyl compounds	6.79 %	639
Propargyl-type 1,3-dipolar organic compounds	6.78 %	502
Phosphate esters	6.75 %	417
Benzoic acids and derivatives	6.64 %	473
Aryl chlorides	6.62 %	718
Alkyl phosphates	6.61 %	399
Pyrroles	6.58 %	431
Cyclic ketones	6.29 %	453
Organofluorides	6.05 %	585
Vinylogous acids	6.03 %	531
Diazines	6.00 %	733
Piperidines	5.90 %	489
Indoles and derivatives	5.62 %	438
Glycerophospholipids	5.47 %	139
Dialkyl phosphates	5.45 %	173
Dialkylamines	5.42 %	546
Tricarboxylic acids and derivatives	5.33 %	116
Alpha,beta-unsaturated carboxylic esters	5.20 %	273
Enoate esters	5.20 %	273
Vinylogous amides	5.19 %	624
Pyrimidines and pyrimidine derivatives	5.13 %	657
Pyrrolidines	5.01 %	381

columns. Both RP and HILIC show the huge diversity and variability that exist in small molecule analysis in the life sciences. See again Supplementary Table 5 for details.

Finally, we analyzed the solvent composition of the separation modes. In RP, the most commonly used mobile phase is H2O + 0.1 % formic acid as weak eluent, and ACN + 0.1 % formic acid as strong eluent. The runner up is H2O + 0.1 % formic acid as weak, and MeOH + 0.1 % formic acid as strong eluent. Notably, these two eluent compositions were used in 211 (71.5 %) of the RP datasets. Also notably, RepoRT already contains 43 other solvent compositions used in 84 RP datasets. This indicates that it might be challenging to develop machine learning models that consider these "non-standard" solvent compositions. Here, estimated pH of the solvents is presumably an important proxy to allow generalization between different solvent compositions. HILIC has a wide variety in the composition of the mobile phase, so unlike RP, there is no preferred eluent system. The most commonly applied mobile phase is H2O + 20 mM ammonium carbonate + 5 µM medronic acid as strong eluent, and 10 % H2O / 90 % ACN + 2 mM ammonium carbonate + 0.5 µM medronic acid as weak eluent, which was reported in 12 datasets. On the other hand the number of different

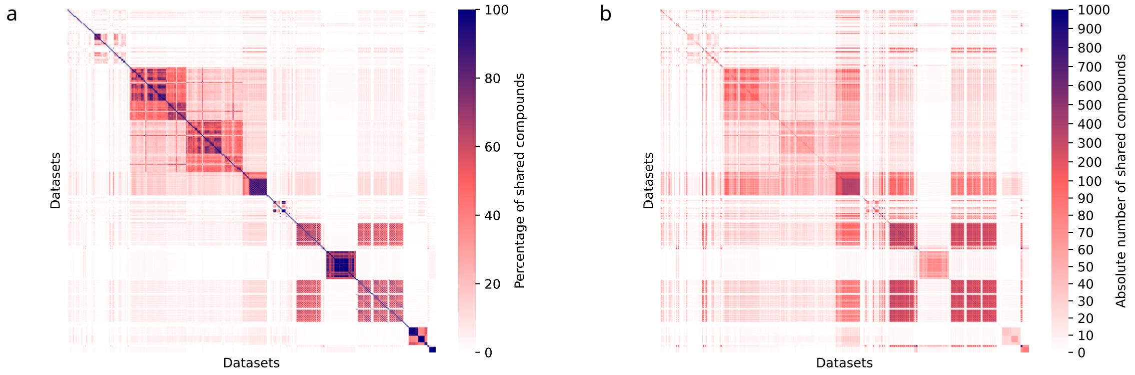

Figure 9. Number of shared compounds between all pairs of datasets. (a) Overlap percentage: Jaccard index, intersection divided by union. (b) Absolute number of shared compounds. Note the difference in scaling between values 0 to 100 and 100 to 1000. Values are clipped at 1000.

solvent compositions is lower compared to RP; RepoRT contains 23 solvent compositions for 71 HILIC datasets.

Discussion

Retention time represents a valuable piece of information for metabolite identification, but depends on the employed chromatographic system. Hence, transferable retention time prediction requires to represent not only properties of the compounds of interest, but also the used chromatographic system. RepoRT provides this information in a machine learning-ready format. It is open for submission from the community, and is not limited to specific chromatographic conditions.

To be of practical use, metadata describing the chromatographic system needs to be complete. In order to avoid differences in spelling of column names, we have compiled a large database of commercially available columns, and use it to standardize input data. Tanaka and HSM parameters describe the selectivity of a stationary phase, and allow a machine learning model to be generalized between different columns. Unfortunately, these parameters are not available for all column models, and certain gaps need to be filled. On the positive side, column models employed regularly in metabolomics, as recorded in MetaboLights and Metabolomics Workbench datasets[56], are also well-covered in RepoRT.

A large fraction of entries in RepoRT is currently missing stereochemistry information; this is particularly important for molecules forming diastereomeric pairs, which potentially can be separated by chromatography (Fig. 1). Missing information on isomeric structures is severely limiting the potential of machine learning models to predict retention time. We invite providers of reference datasets to include information on stereochemistry through the use of stereochemically pure standards and reporting of isomeric SMILES, whenever possible, so that in the future, this information may be incorporated into machine learning models. For future submissions, we ask contributors to provide multiple compound identifiers (Chemical Abstracts Service Registry number, Human Metabolome DataBase[57] or PubChem[58] identifiers) so that annotations can be checked for correctness, and stereochemistry information can be recovered if necessary.

For reversed-phase columns, we argue that RepoRT offers the data required to train machine learning models that predict retention time and order, both for compounds and for columns never seen by the model. In detail, Tanaka and HSM parameters allow the model to generalize between different columns. Unfortunately, the situation is worse for HILIC: Not only is the column chemistry substantially more diverse than for RP, we also do not have broadly available parameters that may allow a machine learning model to generalize between columns. Different approaches similar to Tanaka and HSM have been proposed[59,60]. However, since the chromatographic separation in HILIC is driven by multiple types of interaction (partition, van-der-Waals, electrostatic interaction) no clear separation mechanism can be established. Unless this can be established, transferable predictions for HILIC are unlikely.

Methods

GitHub Repository. Upon submission, data is stored in a folder called "raw_data" with a respective sub-folder with the corresponding 4-digit dataset ID for subsequent tracking. Input data is organized into four files. The first file contains the actual retention time data. If instead of SMILES, only InChIs or identifiers from PubChem, Human Metabolome DataBase (HMDB)[57], ChEBI, KEGG or LipidMaps are provided, the associated SMILES are retrieved using the respective Application Programming Interfaces (API)[61–63]. SMILES are standardized using the PubChem REST (REpresentational State Transfer) API[64]. All successfully standardized structures are stored in a file with the suffix "_success", while all failed structures are stored in a file with the suffix "_failed". This allows tracking of invalid and incorrect SMILES and potential correction. We use CDK[52] via the *rcdk* package (https://cran.r-project.org/web/packages/rcdk/) to compute InChIs and InChI-Keys. Compound classes of the ClassyFire taxonomy[51] with one class each per kingdom, superclass, class, subclass and two additional levels are computed using the ClassyFire web service at http://classyfire.wishartlab.com/ via the *classyfireR* package (https://github.com/aberHRML/classyfireR). Generally, not for all six levels classes can be assigned by ClassyFire: in June 2023, the mean number of assigned levels was 4.73 for the 8809 unique compounds in RepoRT; 46 unique compounds failed classification for all levels. Finally, we compute all molecular descriptors available via *rcdk* (454 descriptors with version 3.7) and ECFP6, MACCS and PubChem (CACTVS) molecular fingerprints likewise using *rcdk* with default parameters.

For each dataset, two files with and without isomeric structure information (named "isomeric" and "canonical", respectively) are provided. Importantly, computed molecular descriptors may vary between these two modes. All raw and processed dataset files describing retention times contain an additional column for flags (e.g., for doublets or potential duplicates) or comments related to individual entries.

Submitted chromatographic metadata is contained in a second file. Upon processing, chromatographic metadata is checked for completeness and consistency as described below. Furthermore, for each dataset, a time for void elution is estimated as $0.0005Ld^2/R$ for column length L, the inner diameter of the column d, and flow rate R. The programmed gradient used for the elution of the compounds is stored in a third separate file. For each dataset, RepoRT presently supports up to four eluents, but this limit can be increased if necessary. A fourth file contains further dataset-related information, most notably DOIs (digital object identifiers) and optionally PubMed IDs of associated publications when available, information on dataset authors or submitters, whether the chromatographic method can be classified into "RP", "HILIC", or "Other", a label for the dataset, and comments or tags hinting at dataset peculiarities of any sort.

To speed up processing, standardized SMILES, ClassyFire compound classes and molecular descriptors are cached; computations are only performed if no information was found in the cached data. The cache is automatically updated upon modifications. Computational times for the different datasets depend on the number of small molecules, and how many have been previously cached. All processing results are stored in a subfolder with the dataset ID in the folder “processed_data”; a description of all files contained in the subfolders can be found in Supplementary Table 6.

In addition to raw and processed datasets, we also store information on column models, column synonyms, as well as Tanaka and HSM parameters in the repository. An overview over all submitted datasets is also available. The workflow is implemented as R and Python scripts, executed automatically via GitHub actions running on GitHub-hosted "windows-2019" runners.

Manual curation of datasets. Numerous datasets from different sources were manually integrated into RepoRT, including data from publications and collections. We spent a particular effort to “clean” datasets and resolve discrepancies. For datasets from PredRet, information on column models, eluents, a system description, and a reference to the publication (DOI, website) was integrated. We found that some information differs between PredRet and the original publication, or additional information is specified in the publication. In such cases, we used information from the original publication to correct RepoRT entries. Furthermore, information about the small molecule structures was sometimes missing, incomplete or contradictory. We used compound names and/or other identifiers to correct and/or complete information on the compound’s structure and its identifiers. Resolving the isomeric structure requires the exact name of the small molecule. Unfortunately, not all compound names allowed us to do so, resulting in entries where no isomeric SMILES can be given. Examples are amino acids and monosaccharides, where the name is provided without stereochemistry information (D/L-conformation).

Eight datasets contained entries without associated retention times. These entries, comprising 1.01 % of all initial RepoRT entries, cannot be used for retention time or retention order prediction, and were discarded from the repository. We assume that such entries stem from compounds that were present in a reference sample mixture, but that could not be detected via mass spectrometry. Next, we removed 4187 (5.23 %) entries that are *duplicates* of other entries, that is, entries with identical structure (as defined by the SMILES identifier and the annotated compound name) and identical retention time in the same dataset. We assume that these duplicates are simply due to copying errors. We manually cleaned 868 “weak duplicates” that differ in compound name annotation: For entries with synonymous compound names or compound names that do not allow further specification, only one entry was kept. As a result, 102 “weak duplicates” were removed. Third, we flagged the remaining 633 “weak duplicates” that could not be resolved manually. Fourth, we flagged all *doublets* in RepoRT, that is, entries with identical structure but different retention times. Fifth, we manually resolved 143 entries where there was a mismatch between reported SMILES, InChI, compound name and/or PubChem identifier, confirming changes to structure annotation with dataset submitters if necessary. Finally, we found 392 entries in RepoRT where different SMILES but only one InChI are provided; here, we performed manual restandardization of the SMILES.

Automated data validation. For datasets measured under *nominally identical* chromatographic conditions, we must not assume that recorded retention times of a compound are identical between datasets. Numerous parameters of the chromatography, such as the “age” of the column, are outside of our control but nevertheless influence retention time. Hence, we resort to a weaker verification: We found that the *retention order* of any pair of compounds shared between two datasets, is usually identical for nominally identical chromatographic conditions. For a newly submitted dataset, we first find all datasets from the same laboratory with nominally identical chromatographic conditions. We then obtain all pairs of compounds shared between the datasets, plus their retention orders. For almost identical retention times, retention order is only of limited use, as a certain fluctuation of retention time has to be expected. We use a relatively large threshold of 30 seconds for the difference in retention time; pairs of compounds where the difference in retention time is smaller in one of the datasets, are excluded from validation. In addition, pairs of compounds with both compounds eluting in the void volume are excluded. We use twice the estimated value "column.t0" stored in the dataset's metadata as the threshold for the void time. Compound pairs eluting in different order are flagged for manual inspection. For the datasets presently contained in RepoRT, we found this validation method to be particularly useful if a certain set of compounds was measured both in positive and negative ionization mode, and some compounds were detected for both ionization modes.

To enable automated detection of outliers for individual datasets, we use QSPR models based on gradient boosting regression. Firstly, we obtain all structures and associated retention times for a dataset, excluding compounds eluting in the void volume. Again, twice "column.t0" is used as the threshold for exclusion. Computed molecular descriptors are used as the features in the QSPR model. Molecular descriptors that could not be computed for every compound are excluded. Descriptors calculated in isomeric mode (see above) take precedence over those calculated without isomeric information. We use Gradient boosting from scikit-learn[65] as the QSPR method, with the number of estimators set to 1000 and the maximum depth set to 2. We use ten-fold cross-validation, ensuring that for each compound, we have one model that has not seen this compound during training. For each compound, we then consider the *prediction error* between the true retention time and the predicted value as the absolute value of the difference. Potential outliers are determined based on the deviation from the prediction error distribution in a dataset. For robust statistics, we resort to quantiles: Let q_1 be the first quantile (25th percentile) and q_3 the third quantile (75th percentile) of prediction errors in the dataset. We define an error threshold as $T = q_3 + 3/2(q_3 - q_1)$. Now, any compound with prediction error larger than T is an outlier of the distribution, and is flagged as a potentially mislabeled compound in the dataset.

For measurements from a single publication or submitter, where one chromatographic parameter is systematically varied across datasets while keeping all other parameters constant, we apply a similar comparison of retention order to detect outliers. For each dataset in the series of systematic measurements, we obtain all pairs of compounds and their retention orders as described above. When there are multiple datasets for the same value of the varied chromatographic parameter (for example, datasets measured in both positive and negative ionization mode), we apply an additional filter to the considered pairs, excluding all pairs where the retention orders are already different for nominally identical conditions (see above). Applying these steps, we obtain a list of compound pairs for each value of the systematically varied chromatographic parameter, allowing comparison of retention orders between variations. Compound pairs that change retention order more than once when increasing the chromatographic parameter, are then flagged for manual inspection.

Measuring datasets with systematic variation of chromatographic parameters. We measured 54 datasets on an Agilent 1290 Infinity II UHPLC coupled to an Agilent 6560 IMS-Q-ToF (Agilent Technologies, Waldbronn, Germany), systematically altering individual chromatographic parameters. In total, we measured 1237 compounds from the Mass Spectrometry Metabolite Library of Standards (MSMLS, 634 compounds), Organic Acid Metabolite Library of Standards (OAMLS, 96 compounds), Fatty Acid Metabolite Library of Standards (FAMLS, 96 compounds), and Bile Acids/Carnitine/Sterol Metabolite Library of Standards (BACSMLS, 96 compounds) (Sigma-Aldrich, Taufkirchen, Germany). Stock solutions and mixes from these libraries were prepared according to the suppliers instructions, which included the solvent for solving substances and mixing schemes. The MSMLS standard plates 1 to 5 are resolved in 95 % water/ 5 % methanol, while plates 6 and 7 are solved in a 1:1 mixture of chloroform and methanol, the OAMLS standards were resolved in water, the FAMLS standards are either solved in chloroform (rows A-C) or in ethanol (rows D-H), and the

BACSMLS are resolved in methanol with the exceptions of A1, B8, B11, B12, C6 and C12, which are solved in chloroform, and B10, which is resolved in a 1:1 mixture of chloroform and methanol. Each standard had a final concentration of 5 µg/ 0.05 mL. For the measurements, all standards of a plate of MSMLS are pooled, while for OAMLS, FAMLS, and BACSMLS all standards of a row were pooled. The Agilent Pesticide Library (250 compounds) (Agilent Technologies, Waldbronn, Germany) and a home-made bile acid mix (45 compounds) have been used without further treatment. A mixture of N-Alkyl-Pyridinium sulfonates (NAPS, 20 standards) (National Research Council, Halifax, Canada)[24] were diluted 1:30 with methanol. In total, the libraries contain 1097 different small molecules (some where overlapping between the libraries). Eluents were always 100 % H2O + 0.1 % formic acid as eluent A and 100 % ACN + 0.1 % formic acid or 100 % MeOH + 0.1 % formic acid as eluent B. Flow rate was set to 400 µL/min and temperature was 30, 40 or 50 °C. Detection was carried out in positive and negative ionization mode and retention times of small molecules were picked manually in Agilent MassHunter Qualitative Analysis 10.0 (Agilent Technologies, Waldbronn, Germany). We used the following column models for separation of small molecules: Phenomenex Kinetex XB-C18 (100 mm x 2.1 mm, 1.7 µm) (Phenomenex, USA) , Waters ACQUITY UPLC BEH C18 (100 mm x 2.1 mm, 1.7 µm), Waters ACQUITY UPLC HSS T3 (100 mm x 2.1 mm, 1.7 µm), Waters ACQUITY UPLC HSS C18 SB (100 mm x 2.1 mm, 1.8 µm), Waters CORTECS UPLC C18 (100 mm x 2.1 mm, 1.6 µm) (Waters, Eschborn, Germany) and Restek Raptor Biphenyl (100 mm x 2.1 mm, 2.7 µm) (Restek, Bellefonte, USA). In total, six column models, three temperatures and two eluent systems were varied, while the gradient and the flow rate remained the same for all datasets. The gradient took 10.01 minutes with additional 2.5 minutes for column equilibrium (see Supplementary Table 7) at a flow rate of 0.4 mL/min.

UMAP plots, Natural Product-likeness scores, and Euclidean distances. We generated Uniform Manifold Approximation and Projection (UMAP) plots following [49]. In detail, 20k molecular structures were uniformly subsampled from a biomolecular structure database containing 718,097 unique molecular structures. These structures serve as a proxy of the universe of biomolecules. Based on an initial UMAP plot, certain lipid classes were optionally excluded, leaving us with 18,096 molecular structures. Note that we show both UMAP plots that include or exclude the lipid classes. The biomolecular structures define the layout of the UMAP plots. For the embedding, we use Maximum Common Edge Subgraph (MCES) distances, as these reflect the biochemical intuition of structural similarity well. Unfortunately, computing the exact MCES is NP-hard (Nondeterministic Polynomial time), so we instead use the myopic MCES distance from [49]. We computed myopic MCES distances ($T = 10$) from all molecular structures in RepoRT to the 18,096 biomolecular structures, and embedded RepoRT compounds into the UMAP plot based on resulting distance vectors.

As a measure of similarity of a compound to (known) natural products, we used the natural product-likeness score introduced by Ertl *et al.*[66] as implemented in RDKit (`https://www.rdkit.org/`, version 2023_3_1). We use the 20k subsampled biomolecular structures from above here, too. Similarly, PubChem was uniformly subsampled to 20k samples. Subsampling is done solely to speed up computations.

When plotting the distribution of Euclidean distances between column pairs, we use bounded kernel estimation using the "betakernel" estimator[67] implemented in the R package "bde". In contrast to a regular kernel density estimator, the bounded kernel estimation ensures that all distances in the kernel density are greater-or-equal to zero.

Statistics and reproducibility. Statistics on the content of RepoRT are based on data from June 2023. For entry-level statistics, retention time data of all datasets except for SMRT are combined. Statistics related to unique compounds are based on the compounds' SMILES representation; isomeric information on the compounds' structure is used when available. Statistics on availability of isomeric information are calculated on the level of entries. Statistics on doublets, entries with identical structures but multiple retention times in one dataset, are also calculated on the level of entries: each instance of a doublet is counted individually, meaning that one structure with two retention times in a set of ten entries results in a 20 % doublet rate. Duplicate entries, entries with identical structure and retention time in one dataset, are likewise counted on the level of entries; naturally, for each removed duplicate, the total number of duplicates normally decreases by more than one. The number of stereocenters for a given structure is computed using *rcdk*'s "get.stereocenters" function.

Data Availability

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

Code Availability

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

References

1. Héberger, K. Quantitative structure-(chromatographic) retention relationships. *J Chromatogr A* **1158,** 273–305 (2007).
2. Moruz, L. & Käll, L. Peptide retention time prediction. *Mass Spectrom Rev* **36,** 615–623 (2017).
3. Kohlbacher, O., Quinten, S., Sturm, M., Mayr, B. M. & Huber, C. G. Structure-activity relationships in chromatography: retention prediction of oligonucleotides with support vector regression. *Angew Chem* **45,** 7009–7012 (2006).
4. Kaliszan, R. *Quantitative Structure-Chromatographic Retention Relationships* (John Wiley & Sons, New York, USA, 1987).
5. Witting, M. & Böcker, S. Current status of retention time prediction in metabolite identification. *J Sep Sci* **43,** 1746–1754 (2020).
6. Dührkop, K., Shen, H., Meusel, M., Rousu, J. & Böcker, S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc Natl Acad Sci USA* **112,** 12580–12585 (2015).
7. Ruttkies, C., Schymanski, E. L., Wolf, S., Hollender, J. & Neumann, S. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J Cheminformatics* **8,** 3 (2016).
8. Sumner, L. W., Amberg, A., Barrett, D., *et al.* Proposed minimum reporting standards for chemical analysis. *Metabolomics* **3,** 211–221 (2007).
9. Schymanski, E. L., Jeon, J., Gulde, R., Fenner, K., Ruff, M., Singer, H. P. & Hollender, J. Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environ Sci Technol* **48,** 2097–2098 (2014).
10. Hu, M., Müller, E., Schymanski, E. L., Ruttkies, C., Schulze, T., Brack, W. & Krauss, M. Performance of combined fragmentation and retention prediction for the identification of organic micropollutants by LC-HRMS. *Anal Bioanal Chem* **410,** 1931–1941 (2018).

11. Bach, E., Szedmak, S., Brouard, C., Böcker, S. & Rousu, J. Liquid-Chromatography Retention Order Prediction for Metabolite Identification. *Bioinformatics* **34.** Proc. of *European Conference on Computational Biology* (ECCB 2018), i875–i883 (2018).
12. Samaraweera, M. A., Hall, L. M., Hill, D. W. & Grant, D. F. Evaluation of an Artificial Neural Network Retention Index Model for Chemical Structure Identification in Nontargeted Metabolomics. *Anal Chem* **90,** 12752–12760 (2018).
13. Bach, E., Schymanski, E. L. & Rousu, J. Joint structural annotation of small molecules using liquid chromatography retention order and tandem mass spectrometry data. *Nature Mach Intel* **4,** 1224–1237 (2022).
14. Stanstrup, J., Neumann, S. & Vrhovšek, U. PredRet: prediction of retention time by direct mapping between multiple chromatographic systems. *Anal Chem* **87,** 9421–9428 (2015).
15. Pasin, D., Mollerup, C. B., Rasmussen, B. S., Linnet, K. & Dalsgaard, P. W. Development of a single retention time prediction model integrating multiple liquid chromatography systems: Application to new psychoactive substances. *Anal Chim Acta* **1184,** 339035 (2021).
16. Low, D. Y., Micheau, P., Koistinen, V. M., *et al.* Data sharing in PredRet for accurate prediction of retention time: Application to plant food bioactive compounds. *Food Chem* **357,** 129757 (2021).
17. 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,** 462867 (2022).
18. Harrieder, E.-M., Kretschmer, F., Dunn, W., Böcker, S. & Witting, M. Critical assessment of chromatographic metadata in publicly available metabolomics data repositories. *Metabolomics* **18,** 97 (2022).
19. Kimata, K., Iwaguchi, K., Onishi, S., Jinno, K., Eksteen, R., Hosoya, K., Araki, M. & Tanaka, N. Chromatographic Characterization of Silica C18 Packing Materials. Correlation between a Preparation Method and Retention Behavior of Stationary Phase. *J Chromatogr Sci* **27,** 721–728 (1989).
20. Snyder, L. R., Dolan, J. W. & Carr, P. W. The hydrophobic-subtraction model of reversed-phase column selectivity. *J Chromatogr A* **1060,** 77–116 (2004).
21. Domingo-Almenara, X., Guijas, C., Billings, E., Montenegro-Burke, J. R., Uritboonthai, W., Aisporna, A. E., Chen, E., Benton, H. P. & Siuzdak, G. The METLIN small molecule dataset for machine learning-based retention time prediction. *Nat Commun* **10,** 5811 (2019).
22. Folberth, J., Begemann, K., Jöhren, O., Schwaninger, M. & Othman, A. MS^2 and LC libraries for untargeted metabolomics: Enhancing method development and identification confidence. *J Chromatogr B* **1145,** 122105 (2020).
23. Pezzatti, J., González-Ruiz, V., Codesido, S., Gagnebin, Y., Joshi, A., Guillarme, D., Schappler, J., Picard, D., Boccard, J. & Rudaz, S. A scoring approach for multi-platform acquisition in metabolomics. *J Chromatogr A* **1592,** 47–54 (2019).
24. Stoffel, R., Quilliam, M. A., Hardt, N., Fridstrom, A. & Witting, M. N-Alkylpyridinium sulfonates for retention time indexing in reversed-phase-liquid chromatography-mass spectrometry-based metabolomics. *Anal Bioanal Chem* **414,** 7387–7398 (2022).
25. Schönberger, K., Mitterer, M., Glaser, K., *et al.* LC-MS-Based Targeted Metabolomics for FACS-Purified Rare Cells. *Anal Chem* **95,** 4325–4334 (2023).
26. Aalizadeh, R., Nikolopoulou, V. & Thomaidis, N. S. Development of Liquid Chromatographic Retention Index Based on Cocamide Diethanolamine Homologous Series (C(n)-DEA). *Anal Chem* **94,** 15987–15996 (2022).
27. Lewis, M., Chekmeneva, E., Camuzeaux, S., *et al.* An Open Platform for Large Scale LC-MS-Based Metabolomics. *chemRxiv* (2022).

28. Huber, C., Müller, E., Schulze, T., Brack, W. & Krauss, M. Improving the screening analysis of pesticide metabolites in human biomonitoring by combining high-throughput *in vitro* incubation and automated LC-HRMS data processing. *Anal Chem* **93,** 9149–9157 (2021).

29. Della Corte, A., Chitarrini, G., Di Gangi, I. M., Masuero, D., Soini, E., Mattivi, F. & Vrhovsek, U. A rapid LC-MS/MS method for quantitative profiling of fatty acids, sterols, glycerolipids, glycerophospholipids and sphingolipids in grapes. *Talanta* **140,** 52–61 (2015).

30. Sawada, Y., Akiyama, K., Sakata, A., Kuwahara, A., Otsuki, H., Sakurai, T., Saito, K. & Hirai, M. Y. Widely targeted metabolomics based on large-scale MS/MS data for elucidating metabolite accumulation patterns in plants. *Plant Cell Physiol* **50,** 37–47 (2009).

31. Strehmel, N., Böttcher, C., Schmidt, S. & Scheel, D. Profiling of secondary metabolites in root exudates of Arabidopsis thaliana. *Phytochemistry* **108,** 35–46 (2014).

32. Arapitsas, P., Speri, G., Angeli, A., Perenzoni, D. & Mattivi, F. The influence of storage on the "chemical age" of red wines. *Metabolomics* **10,** 816–832 (2014).

33. Schymanski, E. L., Singer, H. P., Longrée, P., Loos, M., Ruff, M., Stravs, M. A., Ripollés Vidal, C. & Hollender, J. Strategies to characterize polar organic contamination in wastewater: exploring the capability of high resolution mass spectrometry. *Environ Sci Technol* **48,** 1811–1818 (2014).

34. Beisken, S., Earll, M., Portwood, D., Seymour, M. & Steinbeck, C. MassCascade: Visual Programming for LC-MS Data Processing in Metabolomics. *Mol Inf* **33,** 307–310 (2014).

35. Beisken, S., Earll, M., Baxter, C., *et al.* Metabolic differences in ripening of Solanum lycopersicum 'Ailsa Craig' and three monogenic mutants. *Sci Data* **1,** 140029 (2014).

36. Chaleckis, R., Ebe, M., Pluskal, T., Murakami, I., Kondoh, H. & Yanagida, M. Unexpected similarities between the Schizosaccharomyces and human blood metabolomes, and novel human metabolites. *Mol bioSystems* **10,** 2538–2551 (2014).

37. Bade, R., Bijlsma, L., Miller, T. H., Barron, L. P., Sancho, J. V. & Hernández, F. 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. *Sci Total Environ* **538,** 934–941 (2015).

38. Cao, M., Fraser, K., Huege, J., Featonby, T., Rasmussen, S. & Jones, C. Predicting retention time in hydrophilic interaction liquid chromatography mass spectrometry and its use for peak annotation in metabolomics. *Metabolomics* **11,** 696–706 (2015).

39. Naz, S., Gallart-Ayala, H., Reinke, S. N., Mathon, C., Blankley, R., Chaleckis, R. & Wheelock, C. E. Development of a Liquid Chromatography-High Resolution Mass Spectrometry Metabolomics Method with High Specificity for Metabolite Identification Using All Ion Fragmentation Acquisition. *Anal Chem* **89,** 7933–7942 (2017).

40. Ressom, H. W., Xiao, J. F., Tuli, L., *et al.* Utilization of metabolomics to identify serum biomarkers for hepatocellular carcinoma in patients with liver cirrhosis. *Anal Chim Acta* **743,** 90–100 (2012).

41. Koulman, A., Woffendin, G., Narayana, V. K., Welchman, H., Crone, C. & Volmer, D. A. High-resolution extracted ion chromatography, a new tool for metabolomics and lipidomics using a second-generation orbitrap mass spectrometer. *Rapid Commun Mass Spectrom* **23,** 1411–1418 (2009).

42. Rasche, F., Svatoš, A., Maddula, R. K., Böttcher, C. & Böcker, S. Computing fragmentation trees from tandem mass spectrometry data. *Anal Chem* **83,** 1243–1251 (2011).

43. Theodoridis, G., Gika, H., Franceschi, P., Caputi, L., Arapitsas, P., Scholz, M., Masuero, D., Wehrens, R., Vrhovsek, U. & Mattivi, F. LC-MS based global metabolite profiling of grapes: solvent extraction protocol optimisation. *Metabolomics* **8,** 175–185 (2012).

44. Xiao, J. F., Varghese, R. S., Zhou, B., *et al.* LC-MS based serum metabolomics for identification of hepatocellular carcinoma biomarkers in Egyptian cohort. *J Proteome Res* **11,** 5914–5923 (2012).

45. Roux, A., Xu, Y., Heilier, J.-F., Olivier, M.-F., Ezan, E., Tabet, J.-C. & Junot, C. Annotation of the human adult urinary metabolome and metabolite identification using ultra high performance liquid chromatography coupled to a linear quadrupole ion trap-Orbitrap mass spectrometer. *Anal Chem* **84,** 6429–6437 (2012).
46. Stravs, M. A., Schymanski, E. L., Singer, H. P. & Hollender, J. Automatic recalibration and processing of tandem mass spectra using formula annotation. *J Mass Spectrom* **48,** 89–99 (2013).
47. Stanstrup, J., Gerlich, M., Dragsted, L. O. & Neumann, S. Metabolite profiling and beyond: approaches for the rapid processing and annotation of human blood serum mass spectrometry data. *Anal Bioanal Chem* **405,** 5037–5048 (2013).
48. Dal Santo, S., Tornielli, G. B., Zenoni, S., Fasoli, M., Farina, L., Anesi, A., Guzzo, F., Delledonne, M. & Pezzotti, M. The plasticity of the grapevine berry transcriptome. *Genome Biol* **14,** r54 (2013).
49. Kretschmer, F., Seipp, J., Ludwig, M., Klau, G. W. & Böcker, S. Small molecule machine learning: All models are wrong, some may not even be useful. *bioRxiv,* 2023.03.27.534311 (2023).
50. Hähnke, V. D., Kim, S. & Bolton, E. E. PubChem chemical structure standardization. *J Cheminformatics* **10,** 36 (2018).
51. Djoumbou-Feunang, Y., Eisner, R., Knox, C., *et al.* ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J Cheminformatics* **8,** 61 (2016).
52. Willighagen, E. L., Mayfield, J. W., Alvarsson, J., *et al.* The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J Cheminformatics* **9,** 33 (2017).
53. Durant, J. L., Leland, B. A., Henry, D. R. & Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci* **42,** 1273–1280 (2002).
54. Bolton, E. E., Wang, Y., Thiessen, P. A. & Bryant, S. H. in (eds Wheeler, R. A. & Spellmeyer, D. C.) 217–241 (Elsevier, 2008).
55. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J Chem Inf Model* **50,** 742–754 (2010).
56. Harrieder, E.-M., Kretschmer, F., Böcker, S. & Witting, M. Current State-of-the-Art of Separation Methods Used in LC-MS Based Metabolomics and Lipidomics. *J Chromatogr B* **1188,** 123069 (2022).
57. Wishart, D. S., Guo, A., Oler, E., *et al.* HMDB 5.0: the Human Metabolome Database for 2022. *Nucleic Acids Res* **50,** D622–D631 (2022).
58. Kim, S., Chen, J., Cheng, T., *et al.* PubChem 2023 update. *Nucleic Acids Res* **51,** D1373–D1380 (2023).
59. Ibrahim, M. E. A., Liu, Y. & Lucy, C. A. A simple graphical representation of selectivity in hydrophilic interaction liquid chromatography. *J Chromatogr A* **1260,** 126–131 (2012).
60. Dinh, N. P., Jonsson, T. & Irgum, K. Probing the interaction mode in hydrophilic interaction chromatography. *J Chromatogr A* **1218,** 5880–5891 (2011).
61. Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., Turner, S., Swainston, N., Mendes, P. & Steinbeck, C. ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res* **44,** D1214–9 (2016).
62. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* **44,** D457–D462 (2016).
63. Sud, M., Fahy, E., Cotter, D., *et al.* LMSD: Lipid MAPS structure database. *Nucleic Acids Res* **35,** D527–D532 (2007).
64. Kim, S., Thiessen, P. A., Cheng, T., Yu, B. & Bolton, E. E. An update on PUG-REST: RESTful interface for programmatic access to PubChem. *Nucleic Acids Res* **46,** W563–W570 (2018).

65. Pedregosa, F., Varoquaux, G., Gramfort, A., *et al.* Scikit-learn: Machine Learning in Python. *J Mach Learn Res* **12,** 2825–2830 (2011).
66. Ertl, P., Roggo, S. & Schuffenhauer, A. Natural product-likeness score and its application for prioritization of compound libraries. *J Chem Inf Model* **48,** 68–74 (2008).
67. Chen, S. X. Beta kernel estimators for density functions. *Comput Stat Data Anal* **31,** 131–145 (1999).

Acknowledgments

F.K., M.A.H., and S.B. supported by Deutsche Forschungsgemeinschaft (BO 1910/23). F.K. and S.B. supported by the Ministry for Economics, Sciences and Digital Society of Thuringia (Framework ProDigital, DigLeben-5575/10-9). E.-M.H. and M.W. supported by Deutsche Forschungsgemeinschaft (MW 4382/10-1). We thank everybody who publicly shared their retention time data. We thank Jörg Büscher for providing and uploading presently unpublished data. The authors want to express their particular thanks to Jan Stanstrup, who generously 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 datasets systematically varying chromatographic parameters. 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

Correspondence and requests for materials should be addressed to Michael Witting or Sebastian Böcker.

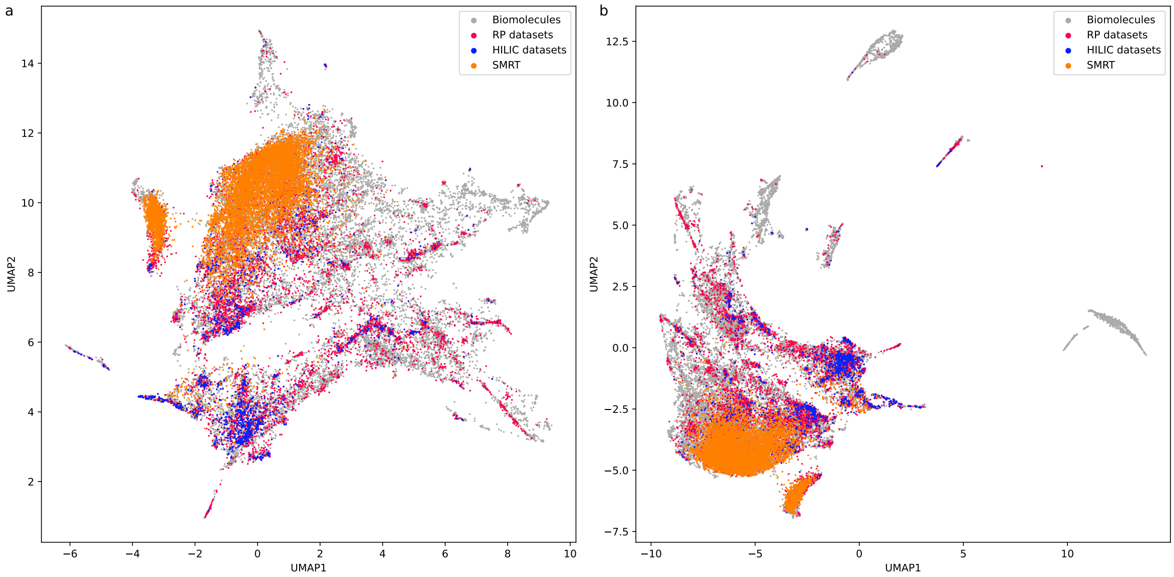

Figure 10. Supplement. Coverage of molecular structures of biological interest by the Small Molecule Retention Time (SMRT) dataset. For comparison, molecular structures from all other datasets in RepoRT are also shown. Due to the large size of the SMRT dataset, a uniform subsample of 10k structures was drawn. As in Fig. 5, projections excluding (a) and including (b) outlier lipid clusters are shown.

Table 2. Supplement. Standard column names used in RepoRT. A separate file will be provided containing the standard column names used within RepoRT, as well as known synonyms. The file can also be accessed from the RepoRT repository at https://github.com/michaelwitting/RepoRT/tree/7cd5539/resources/column_database/column_database.tsv. The current version of the file can be accessed at https://github.com/michaelwitting/RepoRT/tree/master/resources/column_database/column_database.tsv.

Table 3. Supplement. Tanaka parameters. A separate file will be provided containing the Tanaka parameters of columns, considering the pore size of the column. The file can also be accessed from the RepoRT repository at https://github.com/michaelwitting/RepoRT/tree/7cd5539/resources/tanaka_database/tanaka_database.tsv. The current version of the file can be accessed at https://github.com/michaelwitting/RepoRT/tree/master/resources/tanaka_database/tanaka_database.tsv.

Table 4. Supplement. Hydrophobic Selection Model parameters. A separate file will be provided containing the Hydrophobic Selection Model (HSM) parameters of columns. The file can also be accessed from the RepoRT repository at https://github.com/michaelwitting/RepoRT/tree/7cd5539/resources/hsm_database/hsm_database.tsv. The current version of the file can be accessed at https://github.com/michaelwitting/RepoRT/tree/master/resources/hsm_database/hsm_database.tsv.

Table 5. Supplement. Columns in RepoRT. Shown are columns with at least one dataset occurrence in the repository. Columns with unspecified or unknown names were excluded.

Column name	No. dataset using this column	Column type
Waters ACQUITY UPLC BEH C18	63	RP
Waters CORTECS T3	38	RP
Waters ACQUITY UPLC HSS T3	24	RP
Merck Supelco Ascentis Express C18	21	RP
Phenomenex Kinetex EVO C18	17	RP
Waters CORTECS UPLC C18	16	RP
Restek Raptor Biphenyl	16	RP
Waters ACQUITY UPLC HSS C18	16	RP
Phenomenex Kinetex PS C18	14	RP
Phenomenex Kinetex XB-C18	8	RP
Thermo Scientific Hypersil GOLD	6	RP
Phenomenex Kinetex C18	6	RP
Thermo Scientific Acclaim RSLC 120 C18	5	RP
Thermo Scientific Accucore C18	4	RP
Waters Atlantis T3	3	RP
Agilent ZORBAX RRHD Eclipse Plus C18	3	RP
Phenomenex Synergi Hydro-RP	2	RP
Waters ACQUITY UPLC BEH Shield RP18	2	RP
Phenomenex Luna Omega Polar C18	2	RP
Merck Supelco Ascentis Express ES-Cyano	2	RP
Merck Supelco Ascentis Express Phenyl-Hexyl	2	RP
Thermo Scientific Hypercarb	2	RP
Agilent ZORBAX Extend-C18	2	RP
Waters ACQUITY UPLC BEH C8	2	RP
Phenomenex Synergi Polar-RP	2	RP
Hichrom Alltima HP C18	1	RP
Waters XBridge C18	1	RP
Waters Symmetry C18	1	RP
Phenomenex Luna C18	1	RP
Merck Supelco SUPELCOSIL LC-C18	1	RP
Merck LiChrospher RP-18	1	RP
Agilent ZORBAX Eclipse XDB-C18	1	RP
Agilent InfinityLab Poroshell 120 EC-C18	1	RP
Agilent ZORBAX Eclipse Plus C18	1	RP
Advanced Chromatography Technologies ACE C18	1	RP
Phenomenex Kinetex PFP	2	Other
Merck Supelco Ascentis Express F5 (PFP)	2	Other
Thermo Scientific Hypersil GOLD PFP	1	Other
Thermo Scientific Accucore HILIC	14	HILIC
Waters XBridge BEH Amide	14	HILIC
Merck SeQuant ZIC-HILIC	12	HILIC
Waters Atlantis Premier BEH Z-HILIC	9	HILIC
HILICON iHILIC-(P) Classic, HILIC, PEEK	9	HILIC
Merck SeQuant ZIC-pHILIC	5	HILIC
Waters ACQUITY UPLC BEH Amide	4	HILIC
Phenomenex Kinetex HILIC	2	HILIC
Waters ACQUITY UPLC BEH HILIC	1	HILIC
Agilent InfinityLab Poroshell 120 HILIC-Z (Peek-lined)	1	HILIC

Table 6. Supplement. File name conventions. Description of the files produced for each processed dataset contained in RepoRT.

File name	Description
`xxxx_rtdata_canonical_success.txt`	Retention times with associated compounds in various structure representations (standardized SMILES, InChI, InChI-Key); ClassyFire compound classes (based on canonical SMILES); flags or comments for individual entries
`xxxx_rtdata_isomeric_success.txt`	Retention times with associated compounds in various structure representations (standardized SMILES, InChI, InChI-Key); ClassyFire compound classes (based on isomeric SMILES); flags or comments for individual entries
`xxxx_rtdata_canonical_failed.txt`	Dataset entries for which standardization of SMILES failed for canonical SMILES
`xxxx_rtdata_isomeric_failed.txt`	Dataset entries for which standardization of SMILES failed for isomeric SMILES
`xxxx_descriptors_canonical_success.txt`	Molecular descriptors computed by CDK for canonical SMILES
`xxxx_descriptors_isomeric_success.txt`	Molecular descriptors computed by CDK for isomeric SMILES
`xxxx_fingerprints_ecfp6_canonical_success.txt`	Extended-connectivity fingerprints computed for canonical SMILES
`xxxx_fingerprints_ecfp6_isomeric_success.txt`	Extended-connectivity fingerprints computed for isomeric SMILES
`xxxx_fingerprints_maccs_canonical_success.txt`	Molecular ACCess System fingerprints computed for canonical SMILES
`xxxx_fingerprints_maccs_isomeric_success.txt`	Molecular ACCess System fingerprints computed for isomeric SMILES
`xxxx_fingerprints_pubchem_canonical_success.txt`	PubChem (CACTVS) fingerprints computed for canonical SMILES
`xxxx_fingerprints_pubchem_isomeric_success.txt`	PubChem (CACTVS) fingerprints computed for isomeric SMILES
`xxxx_metadata.txt`	Specifications of the chromatographic setup (column, temperature, eluents etc.)
`xxxx_gradient.txt`	Information on the gradient applied
`xxxx_info.txt`	Information related to associated publications and other dataset descriptions (labels, comments, etc.)
`xxxx_report_canonical.pdf`	Visualization of the gradient, flow rate, retention time distribution, and compound classes (isomeric SMILES)
`xxxx_report_isomeric.pdf`	Visualization of the gradient, flow rate, retention time distribution, and compound classes (canonical SMILES)

Table 7. Supplement. Gradient of all measurements with systematic variations of the chromatographic system.

time (min)	%A	%B	flow rate (mL/min)
0.00	95	5	0.4
1.12	95	5	0.4
6.41	0.5	99.5	0.4
10.01	0.5	99.5	0.4
10.02	95	5	0.4
12.52	95	5	0.4