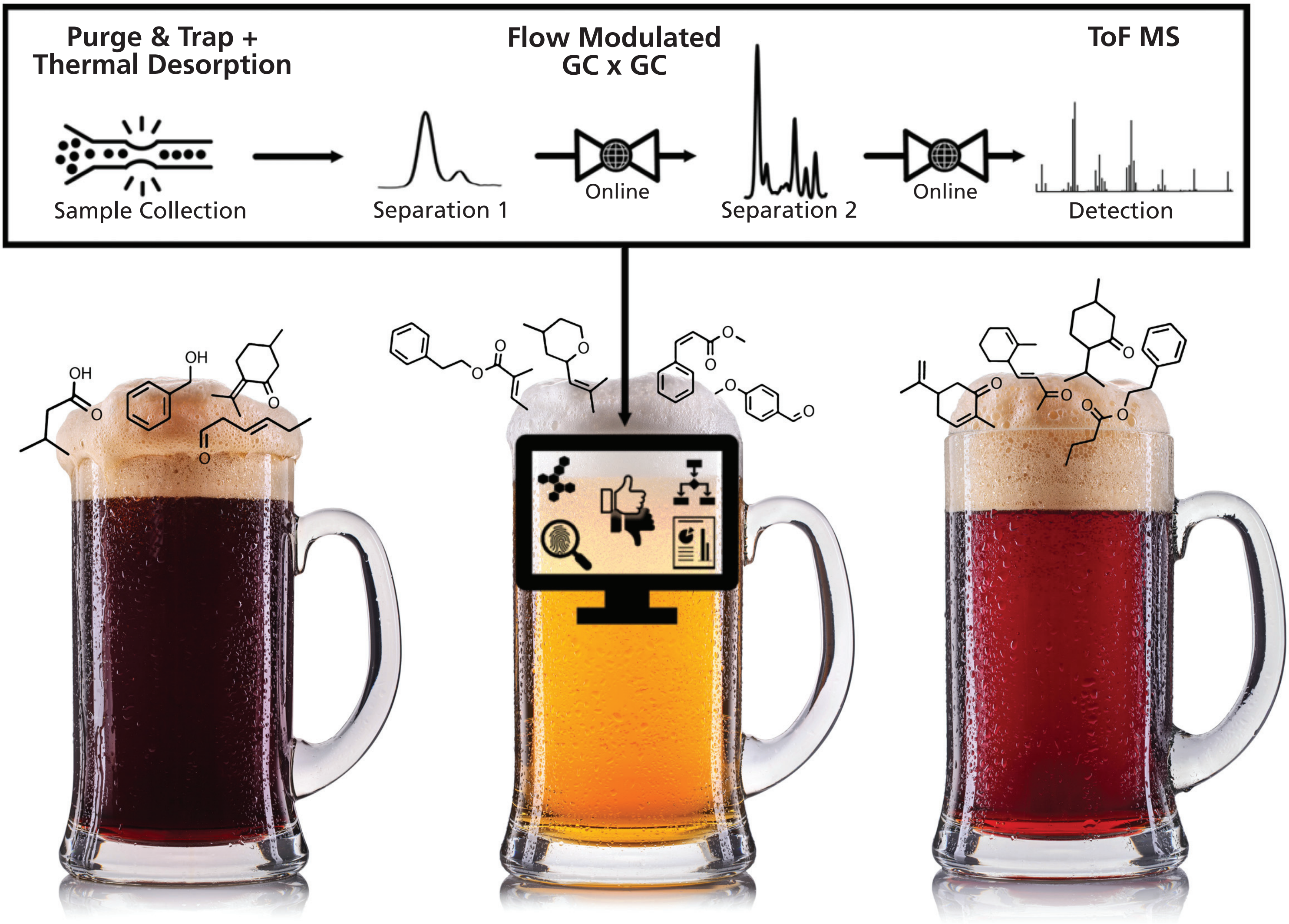
ISSN 1615-9306 · JSSCCJ 43 (9-10) 1615–2012 (2020) · Vol. 43 · No. 9-10 · May 2020 · D 10609

JOURNAL OF SEPARATION SCIENCE

9-10|20



Special Issue
Emerging Thought Leaders in Separation Science

Guest Editor
Elia Psillakis

www.jss-journal.com

WILEY-VCH

Received: 19 January 2020 | Revised: 4 March 2020 | Accepted: 4 March 2020

DOI: 10.1002/jssc.202000060

Journal of SEPARATION SCIENCE

REVIEW ARTICLE

Current status of retention time prediction in metabolite identification

Michael Witting[1,2] | Sebastian Böcker[3]

[1]Research Unit Analytical BioGeoChemistry, Helmholtz Zentrum München, Neuherberg, Germany

[2]Chair of Analytical Food Chemistry, TUM School of Life Sciences, Technische Universität München, Freising, Germany

[3]Chair of Bioinformatics, Friedrich-Schiller-Universität Jena, Jena, Germany

Correspondence
Dr. Michael Witting, Research Unit Analytical BioGeoChemistry, Helmholtz Zentrum München, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany.
Email: michael.witting@helmholtz-muenchen.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: 425789784

Metabolite identification is a crucial step in nontargeted metabolomics, but also represents one of its current bottlenecks. Accurate identifications are required for correct biological interpretation. To date, annotation and identification are usually based on the use of accurate mass search or tandem mass spectrometry analysis, but neglect orthogonal information such as retention times obtained by chromatographic separation. While several tools are available for the analysis and prediction of tandem mass spectrometry data, prediction of retention times for metabolite identification are not widespread. Here, we review the current state of retention time prediction in liquid chromatography–mass spectrometry-based metabolomics, with a focus on publications published after 2010.

KEYWORDS
liquid chromatography, mass spectrometry, metabolite identification, metabolomics, retention time prediction

1 | INTRODUCTION

Metabolomics, the systematic study of metabolites in a biological system, has been called “the apogee of the omics trilogy” [1]. The metabolome is the entirety of all metabolites in a biological system; it constitutes a snapshot of the cell’s or organism’s physiology under particular physiological conditions. In contrast to transcripts and proteins, metabolites are (with few exceptions) not directly encoded in an organism’s genome but are rather determined by the metabolic potential of the encoded enzymes. Furthermore, metabolites and metabolite levels are highly dependent on the surrounding environment, making it hard to estimate the exact number of metabolites present in an organism. In addition to metabolites produced by an organism itself, we may find xenometabolites, food-derived metabolites, drugs, and others. Metabolomics is used in many areas of life science, spanning from fundamental research to translational and personalized medicine.

Article Related Abbreviations: AB, adaptive boosting; ACN, acetonitrile; ANN, artificial neural networks; BRR, Bayesian ridge regression; DDA, data-dependent acquistion; GB, gradient boosting; GNPS, Global Natural Products Social Molecular Networking; GSM, genome scale metabolic model; HMDB, Human Metabolome Database; iPrOH, 2-propanol; LASSO, Least Absolute Shrinkage and Selection Operator Regression; LSVR, linear support vector regression; MINEs, Metabolite In Silico Network Extensions; PLS, partial least square; QqQ, triple quadropole; QSRR, quantitative structure retention relationship; RBF, radial basis function kernel; RF, random forest; RMSE, root-mean-square error; RT, retention time; SVM, support vector machine.

Metabolites from different origins such as body fluids, bacterial or cell cultures, tissues, microbiomes, or even ocean water [2–4] are measured on a routine basis. In clinical metabolomics, urine and plasma are commonly employed body fluids and, among others, are used for diabetes or nutrition research. Further topics of interest include toxicological studies, biomarker and target discovery, as well as clinical trials and studies. It is understood that the above list is incomplete, and that many more application areas of metabolomics exist.

Different analytical methods can be used to measure the metabolome, of which MS and NMR spectroscopy are by far the two most widely used. MS is usually combined with prior separation by chromatography, such as GC, LC or, less frequently, CE. Separation by GC has been used in metabolomics since the 1970s; it is usually coupled to low resolution MS with electron ionization, and often used to investigate primary metabolites. However, combination of GC with high resolution MS such as Orbitrap or Time of Flight (ToF) instruments is recently gaining interest in metabolomics [5,6]. Despite its widespread use, GC–MS is only applicable to volatile molecules and molecules that can be made volatile by derivatization. LC–MS allows to separate a diverse set of compounds including non-volatile compounds, secondary metabolites, drugs, drug metabolism products, food compounds, and others. Here, MS with different mass resolving power is in frequent use: this includes triple quadrupole (QqQ) MS for targeted, and ToF and Orbitrap MS for non-targeted metabolomic investigations. The latter two techniques produce so-called "high-resolution" MS data: high mass resolution allows us to differentiate between ions of almost identical mass; this and the high mass accuracy of the instruments enable non-targeted investigations without prior selection of subsets of metabolites. This facilitates the collection of a comprehensive snapshot of the metabolic state of an organism, cell, or ecosystem and includes the detection of known and unknown metabolites.

2 | METABOLITE IDENTIFICATION

Identification of metabolites constitutes an important, and arguably the currently most pressing bottleneck of LC–MS-based metabolomics: Even for high-resolution mass spectrometric data, da Silva et al. [7] reported that "only 1.8% of spectra in an untargeted metabolomics experiment can be annotated." Despite ongoing discussion on how many features detected in an LC–MS run actually correspond to metabolites [8,9] and how many features are detected for a single metabolite (e.g., different adducts and in-source fragments), it is undisputed that a large fraction of metabolites in the data remain unidentified and make up the "dark matter of metabolomics" [7]. Yet, downstream bioinformatic and biochemical analysis requires accurately identified metabolites for the correct interpretation of results.

Whereas in targeted analysis, information on metabolites of interest is established based on chemical reference standards, untargeted analysis uses MS^2 data to establish (putative) identities of metabolites. Here, metabolite annotation and identification can be performed at different levels, e.g., accurate mass search, formula calculation from accurate mass and isotope pattern, and analysis of tandem MS data [10,11]. It is understood that the accurate mass of a small molecule is insufficient for its structural elucidation: Searching ChemSpider [12] with mass 378.1678 Da and 10 ppm mass accuracy returns more than 9500 structures; but even searching the exact molecular formula $C_{20}H_{26}O_7$ results in 300 structures. Hence, the accurate mass of a small molecule cannot provide information beyond its molecular formula, and trying to identify a small molecule based on its mass will result in a long list containing the putatively correct identity along with numerous false identifications. To this end, tandem MS is usually employed for (partial) structural elucidation. Until recently, identifications were restricted to compounds for which a reference spectrum from a chemical referenced standard was contained in some (commercial, open, or in-house) spectral library, such as METLIN [13], Mass Bank of North America [14], MassBank [15], mzCloud [16], or the Human Metabolome Database (HMDB) [17]. Ideally, analytical conditions between the measurement and spectral libraries are highly similar; this will substantially improve results. The use of spectral libraries for identification represents the current practice, whereas the gold standard for MS-based identification is the comparison of tandem MS and RT data for a chemical standard and biological sample under identical experimental conditions. However, the number of compounds in spectral libraries and for which chemical standards are available is small, compared to the number of detected metabolites. Recently, the coverage of different MS^2 spectral libraries in different genome scale metabolic models (GSMs) has been evaluated: on average, <40% of metabolites in the models have one or several reference spectra from authentic chemical standards [18]. This is despite the fact that GSMs only contain reactions and metabolites belonging to the endogenous metabolism; coverage is clearly even worse if we go beyond that.

Recently, in silico approaches have been developed that allow to search in molecular structure databases such as PubChem [19] and ChemSpider [12]; these in silico approaches are increasingly used by the metabolomics community. MetFrag [22,23] is the oldest, best known, and also most widely used tool for this task; other tools include MAGMa [24], CFM-ID [25], MassFrontier, and, finally, CSI:FingerID [26] which is currently best-in-class. Employed structure databases are many orders of magnitude larger than any

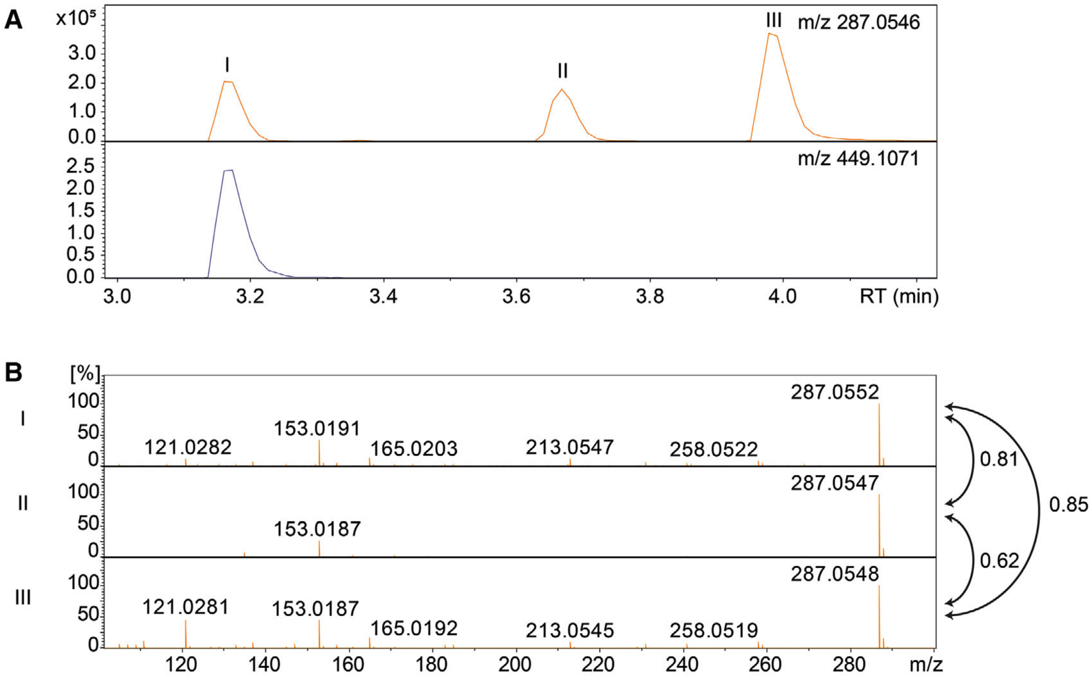


FIGURE 1 Retention time information is augmenting MS data. (A) Extracted ion chromatograms of *m/z* 287.0546 corresponding to an in-source fragment of Kaempferol-7-glucoside (I), Kaempferol (II), and Luteolin (III), and *m/z* 449.1071 corresponding to Kaempferol-7-glucoside. The three isomeric structures show distinct chromatographic retention times. (B) DDA MS2 spectra of the respective chromatographic peaks (I-III). Numbers on the right indicate the cosine score (dot product) between the different spectra based on *m/z* range 100–280 (ignoring the intense precursor *m/z*). The in-source fragment of Kaemperol-7-glucoside and Kaempferol show the highest similarity (0.85)

spectral library and, hence, have a much wider coverage of molecular structures. PubChem contains numerous structures not of biological interest but can serve as a proxy of a very large molecular structure database with more than 100 million entries. Databases of comparable size can be generated using, say, in silico metabolism prediction, such as Metabolite In Silico Network Extensions (MINEs) [20] and BioTransformerDB [21].

Machine learning approaches such as CFM-ID or CSI:FingerID are of particular help to identify substances, but often rely on the input of training data. Performing structural elucidation of novel metabolites by MS/MS with structures not similar to known substances or training data remains an extremely challenging if not impossible task, and typically requires purification of sufficient amount of the substance and 2D NMR measurements.

We noted above that different levels of metabolite identification have been defined, based on the available evidence [10,11]. But even identification by comparing MS/MS to reference data, constituting the second-highest identification level of the Metabolomics Standard Initiative, will result in numerous spurious identifications: Different metabolites can show similar or almost identical fragmentation patterns or RTs. To improve identification quality, combination of independent parameters such as mass, fragmentation pattern, and RT of a chemical reference standard have to be measured under identical analytical conditions and compared to those of the query molecule.

Figure 1 shows a typical example. The extracted ion chromatogram (EIC) for *m/z* 287.0632 shows three distinct chromatographic peaks. For each of the peaks, MS/MS data were collected via data-dependent analysis (DDA). All three compounds fragment very similarly, only the spectrum of the peak at 3.7 min shows different abundances of fragment ions, whereas spectra of the peaks at 3.2 and 4.0 min are almost identical with a cosine score of 0.85. Hence, analyzing the MS/MS data by library matching and/or in silico tools may yield identical search results for peaks at 3.2 and 4.0 min. However, chromatographic behavior clearly differentiates the three substances, whereas the peaks at 3.7 and 4.0 min represent Luteolin and Kaempferol, respectively, the peak at 3.2 min is not a real metabolite signal but an in-source fragment of kaempferol-7-glucoside.

The example shows how the incorporation of the RT dimension can be used to reduce false positive identifications. However, experimental protocols employed in metabolomics are not standardized, and whereas the mass of a metabolite is a molecular property and is consistent across different experiments and laboratories, RTs arise from the combination of the metabolite and the employed chromatographic system. Different column chemistries and solvents lead to different RTs of the same metabolites; unfortunately, this remains true if

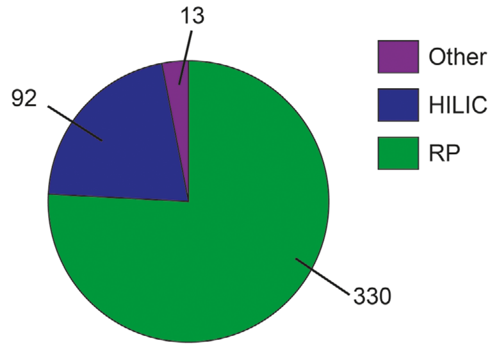


FIGURE 2 Overview on chromatographic separation methods used in studies submitted to Metabolights [29]. Description of chromatographic methods were searched for columns, which were classified into RP, HILIC, or other methods (e.g. pentafluoro phenyl, PFP). In total 435 method descriptions were used

the chromatographic setup is nominally identical but realized on different instruments. RT is often employed at a late stage of metabolite identification, typically when comparing with a chemical reference standard. However, it is not possible for a single laboratory to purchase and host standards of all possible standards for all putative annotations.

3 | SEPARATION TECHNIQUES IN METABOLOMICS

Different separation techniques are used in metabolomics, including CE, GC, LC, SFC, and ion mobility separation (IMS). While each technique has its unique strengths and areas of application, LC is the most widely employed method in non-targeted metabolomics. It can be combined with IMS for a 2D separation prior to mass spectrometric detection. Likewise, both GC and LC can be also used in 2D approaches in which the eluent from the first dimension is transferred to a second dimension with an orthogonal separation chemistry. This increases the peak capacity dramatically and allows for the detection of more substances. Although methods like GC×GC [27] and LC×LC [28] are gaining more attention, they are mostly used by specialist laboratories. While the use of RT in GC has been standardized, e.g., by the use of the Kovats index; in contrast, LC–MS shows much higher variation. In this review, we focus on LC–MS-based nontargeted metabolomics and the use of RT prediction for metabolite identification.

Since metabolites cover a wide range of polarity, no single analytical method can cover the entire metabolome of a given sample or organism. Reversed-phase separation (RP) is used for the separation of hydrophobic substances. Two separation methods are commonly used, the first one uses a gradient from water to organic solvents such as acetonitrile (ACN) or methanol (MeOH) for the chromatographic separation of mid-hydrophobic metabolites, whereas the second uses gradients from water/ACN to 2-propanol (iPrOH) and is typically employed for the separation of lipids.

The main driver for metabolite separation in RP is the partitioning between the hydrophobic stationary phase, e.g., octadecyl modified silica particles, and the hydrophilic mobile phase. Gradient elution toward solvents with higher elution strength (hydrophobicity, e.g., MeOH, ACN, or iPrOH) allows to also elute nonpolar metabolites. Selectivity of separation can be fine-tuned by the addition of different functional groups or other ligands (e.g., phenyl-hexyl).

Analysis of hydrophilic metabolites can be performed using HILIC. In contrast to RP, the separation mechanism of HILIC is not completely understood. While the main driver is also the partitioning between two phases, the water-enriched hydrophilic stationary phase and the hydrophobic mobile phase, several secondary interactions also play important roles. These include ionic interaction, hydrogen bonds, and others. Therefore, the exact separation mechanism in HILIC is less well-defined and relies on the employed column and solvent. Metabolomics does not allow for a "one-size-fits-all" experimental protocol; hence, a diverse set of separation conditions are used in different laboratories. In order to get an overview on the employed separation methods, we reviewed studies submitted to Metabolights [29] that were performed with HPLC–MS or UHPLC–MS, and collected columns and solvents used, irrespective of the method being targeted or nontargeted. From 435 descriptions of chromatographic separations using LC–MS, 330 were classified as RP, 92 as HILIC, and 13 as other (e.g. pentafluoro phenyl, PFP) separations (Figure 2).

4 | RETENTION TIME PREDICTION IN LC–MS-BASED METABOLOMICS

Prediction of RTs can be a promising venue to further filter annotation results toward a reasonable number of candidates. RT prediction is performed by Quantitative Structure Retention Relationship (QSRR) modeling, which aims to relate physicochemical properties of metabolites with their RTs under specific chromatographic conditions. Typically, models are trained on a selection of several tens to hundreds of measured standards and allow the prediction of RTs for chemically closely related structures.

Numerous publications have used different modeling methods for the prediction of RTs under different analytical conditions [30]. Similar to the general trend in metabolomics mostly using RP-based separation, only few papers describe the use of QSRR for HILIC-based nontargeted metabolomics. In this review, we discuss a few selected examples of RT prediction approaches as summarized in Table 1, focusing on publications published after

TABLE 1 Overview on the selected papers employing retention time prediction for metabolomics reviewed here

Publication	Chromatography	No. of metabolites	Data available?
Creek et al. (2011)	HILIC	120	Yes
Domingo-Almenara et al. (2019)	RP	80038	Yes
Eugster et al. (2014)	RP	260	Yes
Cao et al. (2014)	HILIC	93	Yes
Randazzo et al. (2016)	RP	91	Yes
Broeckling et al. (2016)	RP	904	Yes
Bruderer et al. (2016)	RP	532	Yes
Bach et al. (2019)	RP	5 datasets	Yes
Wolfer et al. (2015)	RP	442	No
Aicheler et al. (2015)	RP	201	Yes
Samaraweera et al. (2018)	RP	—	No

2010. A direct comparison of the individual approaches and performance of the used RT predictions is impossible due to different reporting of errors and performance metrics.

4.1 | Modeling of HILIC-based separations

Creek et al. [31] applied RT prediction for a HILIC-based non-targeted metabolomics workflow. Based on 120 authentic standards and multilinear regression (MLR), they obtained a model that predicts RTs with a cross-validated R^2 of 0.82 and a mean squared error (MSE) of 0.14. Modeling was based on selection of optimal descriptors from a set of 11 physicochemical properties. The final model included six physicochemical properties, where logD was the most predictive parameter. Using a similar setup, Cao et al. [32] performed modeling of 93 substances using MLR and random forest (RF) regression. Similar to Creek et al., logP was found to be one of the main features driving the QSRR models. Furthermore, the authors found that RF outperforms the MLR. However, it is known that RF is prone to overfitting, especially when small training sets are used.

4.2 | Modeling of reversed-phase-based separations

Bruderer et al. [33] used RT prediction for support of data independent acquisition (DIA) of MS^2 data. They measured 532 metabolites on two different C18 columns using either pH 3.0 or 8.0. In full, 12 molecular descriptors were predicted using software from ACD/Labs (Advanced Chemistry Development, Toronto, Canada). Out of these 12 features, five were selected for further modeling using multilinear regression. Using a minimal set of 16 compounds, the authors were able to predict retention times with 4 min root-mean-square error (RMSE). Riboflavin detected in urine was used as a validation example. Two chromatographic peaks fitting to the theoretical *m*/*z* of riboflavin in positive ionization mode were detected, but only one was fitting the predicted retention time using the minimal model with 16 compounds.

RT prediction for an RP-based separation was also performed by Wolferer et al. [34]. Based on 442 standards, using the Volsurf+ molecular descriptors as features and support vector regression (SVR) as the machine learning model, the authors were able to predict RTs for their experimental setup with errors of 13% of the RT. Furthermore, an applicability domain approach was used to filter out molecules showing only low similarity to the training set, as these would have high prediction errors. Using their RT prediction, 95% of correct identifications in the validation were among the top three results.

Selection of the correct applicability domain is an important factor, especially if training datasets are small. Eugster et al. [35] restrained their model to only CHO-containing natural products, using a training set of 260 compounds and different models based on partial least square regression (PLS) and artificial neural networks (ANN). They trained sub-class specific models for eight individual compound classes. An additional model was trained on the complete dataset. In their validation experiment, they showed that combining different prediction models improves the identification power.

Aicheler et al. [36] trained a QSRR model using SVR for lipid identifications. Predictions of RTs are particularly valuable for compound classes such as lipids where many isomeric structures exist. Based on 201 lipids identified from mouse fat tissue, they trained a model that was able to remove more than 50% of potential identifications using accurate mass search, which retaining 95% of the correct identifications.

Randazzo et al. [37,38] focused on steroids and employed RT prediction based on linear solvent strength (LSS) theory [39] and QSRR models based on 91 steroids. They also used the VolSurf+ descriptors as descriptors. These descriptors take into account the 3D structure of molecules; this is particularly important for steroids, since these molecules often differ only in the stereochemistry of single groups. This can lead to different conformation of the rings, which can have a huge influence on the chromatographic behavior. Additionally, gonane topological weighted fingerprints (GTWF) specific for steroids and their gonane-based structure were used. Based on experimental RTs under two different gradients, the LSS parameters Log k_w and S were determined. QSRR modeling using the VolSurf+ descriptors and the GTWFs predicted the LSS parameters, which in turn are used to predict RTs. This approach was integrated into a dynamic RT database and was used for the identification of steroids [40].

A common factor to all of the studies is the small size of available training data. Recently, the METLIN small molecule RT dataset was released. This dataset covers 80 038 small molecules, all measured with a single reversed phase method. The dataset is particularly noteworthy because it is one to two orders of magnitude larger than any dataset previously used for RT prediction. Domingo-Almenara et al. [41] used this dataset for RT prediction using Deep Neural Networks (DNN). Validation showed that in 70% of the cases, the correct molecule was among the top three candidates.

4.3 | Evaluation of different machine learning approaches

In most cases of published QSRR approaches for metabolomics, only one or two machine learning approaches are used for prediction. In contrast, Bouwmeester et al. [42] evaluated 36 different metabolomics datasets against seven different machine learning approaches (Bayesian Ridge Regression [BRR]; Least Absolute Shrinkage and Selection Operator Regression [LASSO]; Artificial Neural Networks [ANN]; Adaptive Boosting [AB]; Gradient Boosting [GB]; Random Forest [RF]; linear Support Vector Regression [LSVR], and nonlinear SVR using a Radial Basis Function (RBF) kernel). Two sets of molecular descriptors were used, constituting either 151 features or a minimal set of 11 features. Their analysis showed that no single approach outperforms all other for all evaluated datasets, although GB performed best in most cases. Furthermore, the authors evaluated ensemble learning integrating different algorithms by simply averaging predicted RTs.

4.4 | Integration of multiple separation systems

Common to all modeling approaches is that they only investigate a single chromatographic setup at a time; this is true even for Bouwmeester et al. [42]. Unfortunately, this means that predictions are of no use for anybody using a different experimental setup; and in view of our above remarks, even the transferability to a setup that nominally uses exactly the same experimental protocol, will be rather limited. This may explain the huge number of papers that have been written on RT prediction, not only for use in metabolomics, over the last decades; see the review by Héberger [30]. So far, only few studies focused on the aspect of transferability, despite its obvious importance.

Zisi et al. [43] performed QSRR for 94 metabolite standards and their results indicated that the inclusion of RTs from a different chromatographic column as an additional descriptor improves prediction accuracy.

Bach et al. [44] performed prediction of retention order instead of RTs. This is based on the observation that the retention order of two molecules (which molecule eludes first) is more similar for different chromatographic setups than RT of the two molecules itself. Using a ranking support vector machine (RankSVM), they evaluated five different chromatographic systems, all based on RP separation. For training of the respective models, either single chromatographic systems or multiple systems were used. Their results show that the RankSVM trained on multiple systems outperforms direct RT prediction SVR, with or without training on multiple systems.

Stanstrup et al. [45] developed a system called PredRet, which enable the projection of RTs of measured substances between different chromatographic systems of similar separation chemistry. Commonly detected metabolites are used to define a function for mapping between the different systems and RTs of metabolites detected in one, but not the other system can be projected.

5 | CURRENT LIMITATIONS

An interesting challenge for future work will be the differentiation between stereoisomers. MS and MS/MS data of two stereoisomers are often highly similar or even indistinguishable; therefore, chromatographic separation can deliver valuable additional information. However, to allow us to distinguish between stereoisomers, 3D descriptors are required; but such descriptors are rarely used in current RT prediction approaches. First steps in this direction have been undertaken by using the VolSurf+ 3D descriptors [34,37,38].

Eugester et al. [35] briefly mention this problem: the metabolites isoquercitrin (quercetin-3-O-glucoside) and hyperoside (quercetin-3-*O*-galactoside) have different experimental RTs, while their predicted RT is identical.

Most of the RT predictions have been evaluated by filtering annotations based on exact mass matching or formula look up in databases, using only MS^1 information. However, MS^1 plus RT is not considered sufficient for compound identification or annotation; if a compound is to be annotated, we may assume that MS^2 spectra have been measured for this compound. The important question in this context is: Can RT prediction methods facilitate the annotation of substances beyond the information already available through comparison of MS^2 spectra with library spectra or the use of in silico methods for MS^2 analysis? Integration of RT prediction with MS^2 analysis tools is not common yet. Samaraweera et al. [46] evaluated the prediction of an ANN-based retention index model with different in silico tools: CFM-ID, CSI:FingerID, Mass Frontier, and MetFrag. In case of CFM-ID, MetFrag and Mass Frontier, a significant improvement could be achieved when searching in PubChem. Using the smaller HMDB as a proxy of a biological database, no substantial improvements were observed. CSI:FingerID consistently reached the best annotation results, but showed only modest and non-significant improvements through the integration of RT predictions even when searching PubChem. These results were obtained using 78 compounds as “unknowns”. To this end, it remains elusive how exact RT prediction can be integrated with MS^2 annotation workflows, and whether it can improve annotations of best-in-class methods.

At present, a new model based on ideally several hundred reference standards must be trained individually for each new separation system. Integrative approaches such as the one of Bach et al. are of huge importance for untargeted metabolomics, since they potentially allow us to transfer the trained models to new separation systems.

Lastly, prediction of RTs often relies on molecular descriptors such as logP, logD, and others, which in turn are also predicted using chemoinformatic models. But these predictions already introduce errors. Furthermore, most of these descriptors are on fully aqueous media, whereas chromatographic separation is performed with hydro-organic solvent mixtures. This might introduce further errors and hinder more precise QSRR modeling. Also, most compounds analyzed in metabolomics are charged under the employed analytical conditions, which are of great importance particularly for HILIC-based separations: molecular fingerprints such as (counting) fingerprints or the fingerprints used by Domingo-Almenara et al. [41] may be valuable alternatives once sufficient training data becomes available.

6 | CONCLUSION

RT prediction is increasingly getting attention in non-targeted metabolomics, since it supplies information orthogonal to (tandem) MS data for metabolite identification. Different approaches have been used for the prediction of RTs, using QSRR models based on different molecular descriptors and different machine learning models. Utilized approaches show as much variability as the separation methods used in LC–MS-based metabolomics.

RT prediction in HILIC is assumedly more complicated, since several types of HILIC columns with different surface chemistries are available. Different RP columns are not exactly identical, but they share more common characteristics than different HILIC columns. Furthermore, different pH values are used in HILIC for the separation, causing a strong effect on analyte retention. However, HILIC is getting more attention as a method for the analysis of the polar metabolome. Although initially believed to be not reproducible, an increasing number of studies focusing on the use of HILIC show that there is a major interest: if used properly and with appropriate experimental conditions, the performance of HILIC columns appears to be reproducible.

With more and more data becoming publicly available through metabolomic data repositories such as Metabolights [29], Metabolomics Workbench [47], or Global Natural Products Social Molecular Networking (GNPS) [48], the amount of available training data for RT prediction also increases. Therefore, we argue that the future of RT prediction and integration into metabolite identification workflows is bright; we believe that new approaches will be developed in the near future that will make RT and RT prediction a highly valuable asset in untargeted metabolomics.

ACKNOWLEDGEMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 425789784. We thank Eva-Maria Harrieder for critical reading and corrections.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

ORCID

Michael Witting https://orcid.org/0000-0002-1462-4426
Sebastian Böcker https://orcid.org/0000-0002-9304-8091

REFERENCES

1. Patti, G. J., Yanes, O., Siuzdak, G., Innovation: metabolomics: the apogee of the omics trilogy. *Nat. Rev. Mol. Cell Biol.* 2012, *13*, 263–269.
2. Whiley, L., Godzien, J., Ruperez, F. J., Legido-Quigley, C., Barbas, C., In-vial dual extraction for direct LC-MS analysis of plasma for comprehensive and highly reproducible metabolic fingerprinting. *Anal. Chem.* 2012, *84*, 5992–5999.
3. Behrends, V., Ryall, B., Zlosnik, J. E. A., Speert, D. P., Bundy, J. G., Williams, H. D., Metabolic adaptations of Pseudomonas aeruginosa during cystic fibrosis chronic lung infections. *Environ. Microbiol.* 2012, *15*, 398–408.
4. Walker, A., Pfitzner, B., Neschen, S., Kahle, M., Harir, M., Lucio, M., Moritz, F., Tziotis, D., Witting, M., Rothballer, M., Engel, M., Schmid, M., Endesfelder, D., Klingenspor, M., Rattei, T., Castell, W. Z., de Angelis, M. H., Hartmann, A., Schmitt-Kopplin, P., Distinct signatures of host-microbial meta-metabolome and gut microbiome in two C57BL/6 strains under high-fat diet. *ISME J.* 2014, *8*, 2380–2396.

5. Weidt, S., Haggarty, J., Kean, R., Cojocariu, C. I., Silcock, P. J., Rajendran, R., Ramage, G., Burgess, K. E. V., A novel targeted/untargeted GC-Orbitrap metabolomics methodology applied to *Candida albican*s and *Staphylococcus aureus* biofilms. *Metabolomics* 2016, *12*, 189.
6. Ji, J., Sun, J., Pi, F., Zhang, S., Sun, C., Wang, X., Zhang, Y., Sun, X., GC-TOF/MS-based metabolomics approach to study the cellular immunotoxicity of deoxynivalenol on murine macrophage ANA-1 cells. *Chem.-Biol. Interact.* 2016, *256*, 94–101.
7. da Silva, R. R., Dorrestein, P. C., Quinn, R. A., Illuminating the dark matter in metabolomics. *Proc. Natl. Acad. Sci. U. S. A.* 2015, *112*, 12549–12550.
8. Mahieu, N. G., Patti, G. J., Systems-level annotation of a metabolomics data set reduces 25 000 features to fewer than 1000 unique metabolites. *Anal. Chem.* 2017, *89*, 10397–10406.
9. Baran, R., Untargeted metabolomics suffers from incomplete raw data processing. *Metabolomics* 2017, *13*, 107.
10. Sumner, L., Amberg, A., Barrett, D., Beale, M., Beger, R., Daykin, C., Fan, T. M., Fiehn, O., Goodacre, R., Griffin, J., Hankemeier, T., Hardy, N., Harnly, J., Higashi, R., Kopka, J., Lane, A., Lindon, J., Marriott, P., Nicholls, A., Reily, M., Thaden, J., Viant, M., Proposed minimum reporting standards for chemical analysis. *Metabolomics* 2007, *3*, 211–221.
11. 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.* 2014, *48*, 2097–2098.
12. Pence, H. E., Williams, A., ChemSpider: an online chemical information resource. *J. Chem. Educ.* 2010, *87*, 1123–1124.
13. Guijas, C., Montenegro-Burke, J. R., Domingo-Almenara, X., Palermo, A., Warth, B., Hermann, G., Koellensperger, G., Huan, T., Uritboonthai, W., Aisporna, A. E., Wolan, D. W., Spilker, M. E., Benton, H. P., Siuzdak, G., METLIN: a technology platform for identifying knowns and unknowns. *Anal. Chem.* 2018, *90*, 3156–3164.
14. http://mona.fiehnlab.ucdavis.edu/ (last time accessed February 27, 2020).
15. Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass. Spectrom.* 2010, *45*, 703–714.
16. https://www.mzcloud.org/ (last time accessed February 27, 2020).
17. Wishart, D. S., Feunang, Y. D., Marcu, A., Guo, A. C., Liang, K., Vázquez-Fresno, R., Sajed, T., Johnson, D., Li, C., Karu, N., Sayeeda, Z., Lo, E., Assempour, N., Berjanskii, M., Singhal, S., Arndt, D., Liang, Y., Badran, H., Grant, J., Serra-Cayuela, A., Liu, Y., Mandal, R., Neveu, V., Pon, A., Knox, C., Wilson, M., Manach, C., Scalbert, A., HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res*. 2018, *46*, D608-D617.
18. Frainay, C., Schymanski, E. L., Neumann, S., Merlet, B., Salek, R. M., Jourdan, F., Yanes, O., Mind the gap: mapping mass spectral databases in genome-scale metabolic networks reveals poorly covered areas. *Metabolites* 2018, *8*, 51.
19. Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B. A., Wang, J., Yu, B., Zhang, J., Bryant, S. H., PubChem substance and compound databases. *Nucleic Acids Res*. 2016, *44*, D1202-D1213.
20. Jeffryes, J. G., Colastani, R. L., Elbadawi-Sidhu, M., Kind, T., Niehaus, T. D., Broadbelt, L. J., Hanson, A. D., Fiehn, O., Tyo, K. E. J., Henry, C. S., MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. *J. Cheminf.* 2015, *7*, 44.
21. Djoumbou-Feunang, Y., Fiamoncini, J., Gil-de-la-Fuente, A., Greiner, R., Manach, C., Wishart, D. S., BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *J. Cheminf.* 2019, *11*, 2.
22. Wolf, S., Schmidt, S., Muller-Hannemann, M., Neumann, S., In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinf.* 2010, *11*, 148.
23. Ruttkies, C., Schymanski, E. L., Wolf, S., Hollender, J., Neumann, S., MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J. Cheminf.* 2016, *8*, 3.
24. Ridder, L., Hooft, J. J. J., Verhoeven, S., Vos, R. C. H., Schaik, R., Vervoort, J., Substructure-based annotation of high-resolution multistage MSn spectral trees. *Rapid. Commun. Mass. Spectrom.* 2012, *26*, 2461–2471.
25. Allen, F., Greiner, R., Wishart, D., Competitive fragmentation modeling of ESI–MS/MS spectra for putative metabolite identification. *Metabolomics* 2015, *11*, 98–110.
26. 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. U. S. A.* 2015, *112*, 12580-12585.
27. Yu, Z., Huang, H., Reim, A., Charles, P. D., Northage, A., Jackson, D., Parry, I., Kessler, B. M., Optimizing 2D gas chromatography mass spectrometry for robust tissue, serum and urine metabolite profiling. *Talanta* 2017, *165*, 685–691.
28. Lipok, C., Hippler, J., Schmitz, O. J., A four dimensional separation method based on continuous heart-cutting gas chromatography with ion mobility and high resolution mass spectrometry. *J. Chromatogr. A* 2018, *1536*, 50–57.
29. Haug, K., Salek, R. M., Conesa, P., Hastings, J., de Matos, P., Rijnbeek, M., Mahendraker, T., Williams, M., Neumann, S., Rocca-Serra, P., Maguire, E., González-Beltrán, A., Sansone, S.-A., Griffin, J. L., Steinbeck, C., MetaboLights—an open-access general-purpose repository for metabolomics studies and associated metadata. *Nucleic Acids Res*. 2013, *41*, D781-D786.
30. Héberger, K., Quantitative structure–(chromatographic) retention relationships. *J. Chromatogr. A* 2007, *1158*, 273–305.
31. Creek, D. J., Jankevics, A., Breitling, R., Watson, D. G., Barrett, M. P., Burgess, K. E. V., Toward global metabolomics analysis with hydrophilic interaction liquid chromatography–mass spectrometry: improved metabolite identification by retention time prediction. *Anal. Chem.* 2011, *83*, 8703–8710.
32. 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* 2014, 1–11.
33. Bruderer, T., Varesio, E., Hopfgartner, G., The use of LC predicted retention times to extend metabolites identification with SWATH data acquisition. *J. Chromatogr. B* 2017, *1071*, 3–10.
34. Wolfer, A. M., Lozano, S., Umbdenstock, T., Croixmarie, V., Arrault, A., Vayer, P., UPLC–MS retention time prediction: a machine learning approach to metabolite identification in untargeted profiling. *Metabolomics* 2015, *12*, 8.
35. Eugster, P. J., Boccard, J., Debrus, B., Bréant, L., Wolfender, J.-L., Martel, S., Carrupt, P.-A., Retention time prediction for dereplication of natural products (CxHyOz) in LC–MS metabolite profiling. *Phytochemistry* 2014, *108*, 196–207.

36. Aicheler, F., Li, J., Hoene, M., Lehmann, R., Xu, G., Kohlbacher, O., Retention time prediction improves identification in nontargeted lipidomics approaches. *Anal. Chem.* 2015, *87*, 7698–7704.
37. Randazzo, G. M., Tonoli, D., Hambye, S., Guillarme, D., Jeanneret, F., Nurisso, A., Goracci, L., Boccard, J., Rudaz, S., Prediction of retention time in reversed-phase liquid chromatography as a tool for steroid identification. *Anal. Chim. Acta* 2016, *916*, 8–16.
38. Randazzo, G. M., Tonoli, D., Strajhar, P., Xenarios, I., Odermatt, A., Boccard, J., Rudaz, S., Enhanced metabolite annotation via dynamic retention time prediction: steroidogenesis alterations as a case study. *J. Chromatogr. B* 2017, *1071*, 11–18.
39. Snyder, L. R., Dolan, J. W., Carr, P. W., The hydrophobic-subtraction model of reversed-phase column selectivity. *J. Chromatogr. A* 2004, *1060*, 77–116.
40. Codesido, S., Randazzo, G. M., Lehmann, F., González-Ruiz, V., García, A., Xenarios, I., Liechti, R., Bridge, A., Boccard, J., Rudaz, S., DynaStI: a dynamic retention time database for steroidomics. *Metabolites* 2019, *9*, 85.
41. 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.* 2019, *10*, 5811.
42. Bouwmeester, R., Martens, L., Degroeve, S., Comprehensive and empirical evaluation of machine learning algorithms for small molecule LC retention time prediction. *Anal. Chem.* 2019, *91*, 3694–3703.
43. Zisi, C., Sampsonidis, I., Fasoula, S., Papachristos, K., Witting, M., Gika, H., Nikitas, P., Pappa-Louisi, A., QSRR modeling for metabolite standards analyzed by two different chromatographic columns using multiple linear regression. *Metabolites* 2017, *7*, 7.
44. Bach, E., Szedmak, S., Brouard, C., Böcker, S., Rousu, J., Liquid-chromatography retention order prediction for metabolite identification. *Bioinformatics* 2018, *34*, i875-i883.
45. Stanstrup, J., Neumann, S., Vrhovšek, U., PredRet:prediction of retention time by direct mapping between multiple chromatographic systems. *Anal. Chem.* 2015, *87*, 9421–9428.
46. 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.* 2018, *90*, 12752–12760.
47. Sud, M., Fahy, E., Cotter, D., Azam, K., Vadivelu, I., Burant, C., Edison, A., Fiehn, O., Higashi, R., Nair, K. S., Sumner, S., Subramaniam, S., Metabolomics Workbench: an international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res.* 2016, *44*, D463-D470.
48. Wang, M., Carver, J. J., Phelan, V. V., et al., Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* 2016, *34*, 828–837.

AUTHOR BIOGRAPHY

Dr. Michael Witting studied Applied Chemistry with the functional direction Biochemistry at the Georg-Simon-Ohm University of Applied Sciences in Nuremberg and received his Ph.D. in 2013 from the Technical University of Munich. He works as a scientist at the Helmholtz Zentrum München and his research focuses on the development of new methods for the analysis and annotation of the *Caenorhabditis elegans* metabolome and lipidome as well as new methods in metabolite identification.

How to cite this article: Witting M, Böcker S. Current status of retention time prediction in metabolite identification. *J Sep Sci.* 2020;43:1746–1754. https://doi.org/10.1002/jssc.202000060