**Current status of retention time prediction in metabolite identification**

Michael Witting1,2,\* & Sebastian Böcker3

1 Research Unit Analytical BioGeoChemistry, Helmholtz Zentrum München, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany

2 Chair of Analytical Food Chemistry, TUM School of Life Sciences, Technische Universität München, Maximus-von-Imhof-Forum 2, 85354 Freising, Germany

3 Chair of Bioinformatics, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany

\* corresponding author:

Dr. Michael Witting

Research Unit Analytical BioGeoChemistry

Helmholtz Zentrum München

Ingolstädter Landstraße 1

85764 Neuherberg

Germany

Phone: 0049 (0)89 – 3187 4755

eMail: michael.witting@helmholtz-muenchen.de

**Abstract**

Metabolite identification is a crucial step in non-targeted metabolomics, but also represents the current bottleneck. Accurate identifications are required for correct biological interpretation. To data identification is mostly based on the use of tandem MS, but mostly neglects orthogonal information like retention times obtained by chromatographic separation. While several tools are available for the analysis of tandem MS 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 recent publications.

**Keyword**

Metabolite identification

Metabolomics

LC-MS

Retention time prediction

**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, under particular physiological conditions; it constitutes a snapshot of the cell’s or organism’s physiology. In contrast to transcripts and proteins, metabolites are (with few exceptions) not encoded in the organism’s genome. Metabolites are highly dependent on the surrounding environment, making it hard to estimate the exact number of metabolites present in an organism. Molecules that might be found in an organism include xenometabolites, food-derived metabolites, drugs and others. Metabolomics is utilized in many areas of life science, from fundamental research to translational and personalized medicine. Metabolites are measured on a routine basis from different origins such as body fluids, bacterial or cell cultures, tissues, microbiomes or even ocean water [2-4]. Urine and plasma are commonly employed body fluids in medical metabolomics, and are used for diabetes or nutrition research, among others. Toxicological studies, biomarker and target discovery, clinical trials and studies are further topics of interest.

Different analytical methods are used to measure the metabolome; mass spectrometry (MS) and nuclear magnetic resonance (NMR) are by far the two most employed. Mass spectrometry is usually combined with prior separation by chromatography, such as gas chromatography (GC), liquid chromatography (LC) or, less frequently, capillary electrophoresis (CE). Separation by GC is used in metabolomics since the 1970s; it is usually coupled to low resolution MS with Electron Ionization, and often used to investigate primary metabolites. Despite its widespread use, it is only applicable to volatile molecules and molecules that can be made volatile by derivatization. This is usually the case for molecules with a mass up to 500 Da. LC-MS offers different separation selectivity for the analysis of a wide range of compounds, including non-volatile compounds, secondary metabolites, drugs and drug degradation products, food compounds and others. Here, MS with different mass resolving power are in frequent use: This includes triple quadrupole (QqQ) MS for targeted, and Time of Flight (QqToF) and Orbitrap MS for non-targeted metabolomics investigations. The latter two techniques produce so-called “high-resolution” MS data; it is noteworthy that not the high resolution (that is, the possibility to differentiate between ions of almost identical mass) but rather the high mass accuracy of the instruments enables 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.

**Metabolite identification**

Identification of metabolites constitutes an important (arguably the currently most pressing) bottleneck of LC-MS-based metabolomics: Even for high-resolution mass spectrometric data, da Silva et al. [5] 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 [6, 7], it is undisputed that a large fraction of metabolites in this data remain unidentified and make up the “dark matter of metabolomics” [5]. Downstream bioinformatic and biochemical analysis requires accurately identified metabolites for correct interpretation of obtained metabolomics results.

While in targeted analysis accurate mass and retention times are established based on chemical reference standards, untargeted analysis uses MS2 data to establish (putative) identities of metabolites. It is understood that the accurate mass of a small molecule is insufficient for its structural elucidation: Searching ChemSpider [8] with mass 378.1678 Da and 10 ppm mass accuracy returns more than 9,500 structures; but even searching the exact molecular formula C20H26O7 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 numerous false identifications. To this end, tandem MS is usually employed for (partial) structural elucidation. Until recently, identifications were restricted to compounds where a reference spectrum from a chemical referenced standard was contained in some (commercial, open or in-house) spectral library, e.g. METLIN [9], Mass Bank of North America, MassBank [10] or Human Metabolome Database (HMDB) [11]. The use of spectral libraries for identification represents the current gold standard. Ideally, analytical conditions between the measurement and spectral libraries are comparable to achieve best results. However, the number of compounds in spectral libraries is small compared to detected metabolites. Recently, the coverage of different MS2 spectral libraries in different genome scale metabolic models has been evaluated. On average only 40% of metabolites in the models have one or several reference spectra from authentic chemical standards [12].

Recently, *in silico* approaches have been developed and are increasingly used by the metabolomics community, that allow to search in molecular structure databases such as PubChem [13] and ChemSpider [8]. Such structure databases are many orders of magnitude larger than any spectral library and, hence, have a much wider coverage of molecular structures. PubChem also contains many structures not of biological interest but can serve as a proxy of a very large molecular structure database. Databases of comparable size can be generated using, say, *in silico* metabolism prediction, such as Metabolite In Silico Network Extensions (MINEs) [14] and BioTransformerDB [15]. MetFrag [16, 17] is the oldest, best-known and also most widely used tool for this task, despite its subpar performance in recent evaluations [18]; other tools include MAGMa [19], CFM-ID [20] and, finally, CSI:FingerID [21] which is currently best-in-class.

Especially machine learning approaches such as CFM-ID or CSI:FingerID can help to identify substances, but often rely on the input of training data. Performing structural elucidation of novel metabolites by tandem 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.

Different levels of metabolite identification have been defined, based on the available evidence [22, 23]. But even identification by comparing tandem MS to reference data – constituting the second-highest identification level of the Metabolomics Standard Initiative will result in numerous spurious identifications: Different metabolites can produce similar or almost identical fragmentation patterns. To improve identification quality, independent parameters, e.g. mass, fragmentation pattern and retention time of a chemical reference standard have to be reported measured under identical analytical conditions. Here, retention time information can be used to correct false identifications made based on the MS data.

Figure 1 shows a typical example. The Extracted Ion Chromatogram (EIC) for m/z 287.0632 shows three distinct chromatographic peaks. Each of the peaks has tandem MS collected via data-dependent fragmentation. All three compounds fragment very similar: Only the spectrum from the peak at 3.7 minutes shows different abundances of fragment ions, whereas spectra from the peaks at 3.2 and 4.0 minutes are almost identical with a cosine score of 0.85. Here, analysis of tandem MS data by library matching and in silico tools lead to very similar results and the two peaks at 3.2 minutes and 4.0 minutes will, by all means, yield identical search results. However, chromatographic behavior clearly differentiates the three substances. Whereas the peaks at 3.7 and 4.0 minutes represent Luteolin and Kaempferol, the peak at 3.2 is no real metabolite signal, but an in-source fragment of Kaempferol-7-glucoside.

Ein Bild, das Text, Karte enthält.

Automatisch generierte Beschreibung

*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), 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).*

The example shows how the incorporation of the retention time dimension can be used for reduction of false positive identifications. However, experimental protocols employed in metabolomics are not standardized, and whereas the mass of a metabolite is a molecular property, retention times are a combination of the metabolite and the employed chromatographic system. Different column chemistries and solvents lead to different retention times of the same metabolites; unfortunately, this remains true if the chromatographic setup is nominally identical but realized on different instruments. Retention time 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 host standards of all possible compounds. On the other hand, it is not feasible to purchase and measure all possible standards for all putative annotations.

**Separation techniques in metabolomics**

Different separation techniques are used in metabolomics, e.g. CE, GC, LC, supercritical fluid chromatography (SFC) and ion mobility separation (IMS). While each technique has 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 two-dimensional separation prior to mass spectrometric detection. Likewise, GC and LC can be also used in two-dimensional 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 detection of more substances. Although methods like GCxGC [24] and LCxLC [25] are gaining more and more attention, they are mostly used by specialist laboratories. In this review we are focusing on LC-MS based non-targeted metabolomics and the use of retention time prediction for metabolite identification.

*Separation methods used in metabolomics*

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. Typically, two different separation methods are used. The first one uses a gradient from water to organic solvents like acetonitrile (ACN) or methanol (MeOH) for the chromatographic separation of mid-hydrophobic metabolites, while gradients from water/ACN to iPrOH are typically employed for the separation of lipids. Analysis of hydrophilic metabolites is performed using Hydrophilic Liquid Interaction Chromatography (HILIC).

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 towards solvents with higher elution strength (hydrophobicity) allows to also elute non-polar metabolites. Selectivity of separation can be fine-tuned by the addition of different functional groups or other ligands (e.g. phenyl-hexyl) to change selectivity. 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 play important roles. These include ionic interaction, hydrogen bonds and others. Therefore, the exact separation mechanism in HILIC is less defined and relies on the employed column and solvent. Since no “one-fits-all” protocol is available in metabolomics a diverse set of different separation conditions are used between the different laboratories. In order to get an overview on the employed separation methods we reviewed studies submitted to Metabolights [26], which were performed with HPLC-MS or UHPLC-MS and collected columns and solvents used (irrespective of the method being targeted or non-targeted). From 435 descriptions of chromatographic separations using LC-MS (date 23.12.2019) 330 were classified as RP, 92 as HILIC or 13 as other (e.g. pentafluoro phenyl, PFP) separations.



*Figure 2: Overview on chromatographic separation methods used in studies submitted to Metabolights [26]. 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.*

**Retention time prediction in LC-MS based metabolomics**

Prediction of retention times represents a promising approach of further filtering annotation results towards a reasonable number of candidates. Retention time prediction is performed by Quantitative Structure Retention Relationship (QSRR) modeling, which aims to relate physicochemical properties of metabolites with their retention times under specific chromatographic conditions. Typically, models are trained on a selection of several tens to hundreds of measured standards and allow the prediction of retention times for chemically closely related structures. Different QSRR models have been applied in metabolomics for the improved annotation and identification of metabolites.

Different publications have used different modeling algorithms for the prediction of retention times under different analytical conditions [27]. Similar to the general trend in metabolomics mostly using RP based separation, only few papers describe the use of QSRR for HILIC-based non-targeted metabolomics. In this review we discuss a few selected examples of retention time prediction approaches as summarized in Table 1, focusing on recent publications.

< Insert Table 1 here >

*Modeling of HILIC based separations*

Creek et al. [28] applied retention time prediction for a HILIC-based non-targeted metabolomics workflow. Based on 120 authentic standards and multilinear regression (MLR) they obtained a model which predicts retention times with a cross-validated R² 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 6 physicochemical properties, where the logD was the most predictive parameter. In a similar manner Cao et al. [29] performed modeling of 93 substances using MLR or random forest regression (RF). Comparable to Creek et al. also the logP was found to be one of the main descriptors driving the QSRR models. Furthermore, they found that RF outperforms the MLR. However, RF is prone to overfitting especially when small training sets are used.

*Modeling of RP based separations*

Bruderer et al. [30] used retention time prediction for support of Data Independent Acquisition (DIA) of MS2 data. They have measured 532 metabolites on two different C18 columns using either pH 3.0 or pH 8.0. In full, 12 molecular descriptors were predicted using software from ACD/Labs (Advanced Chemistry Development, Toronto, Canada). Out of these 12, 5 were selected for further modeling using multilinear regression. From the compounds a minimal set of 16 compounds they were able to predict retention times with 4 minutes root-mean-square error (RMSE). Riboflavin detected in urine was used as validation example. Two peak fitting to the 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.

Retention time prediction for a RP-based separation was also performed by Wolferer et al. [31]. Based on 442 standards and the Volsurf+ molecular descriptors using support vector regression (SVR), they were able to predict retention times for their experimental setup with errors of 13% of the retention time. Furthermore, an applicability domain approach was used to filter out molecules showing only low similarity to the training set and would have high prediction errors. Using their retention time prediction 95% of correct identifications in the validation were among the top 3 results.

Selection of the correct applicability domain is an important factor, especially if training datasets are small. Eugster et al. [32] 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) or artificial neural networks (ANN). They developed also sub-class specific models for different 8 compound classes, training each model individually. An additional model was trained on the complete dataset. In their validation experiment they showed that combining different prediction models improve the identification.

Aicheler et al. [33] developed a QSRR model based on SVR for lipid identifications. Predictions of retention times are very valuable for compound classes were many isomeric structures exists. Based on 201 lipids identified from mouse fat tissue they developed 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. [34, 35] focused on steroids and employed retention time prediction based on linear solvent strength (LSS) theory [36] and QSRR models based on 91 steroids. They also used the VolSurf+ descriptors. These descriptors take into account the three-dimensional structure of molecules. This is particularly important for steroids, since they often only differ in the stereochemistry of single groups. This can lead to different conformation of the rings, which has 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 retention times under two different gradients the LSS parameters Log kw and S were determined. QSRR modeling using the VolSurf+ descriptors and the GTWFs predicted the LSS parameters, which in turn are used to predict retention times. This approach was integrated into a dynamic retention time database and was used for the identification of steroids [37].

A common factor to all of the studies is the small size of available training data. Recently, the METLIN small molecule retention time data set was released. Covering 80038 small molecules measured with a single reversed phase method. Domingo-Almenara et al. [38] used this dataset for machine learning based retention time prediction. Deep learning was used to predict retention times. Validation showed that in 70% of the cases the correct molecule was among the top 3 candidates.

*Evaluation of different machine learning approaches*

In most cases of published QSRR approaches in metabolomics only one or two selected machine learning approaches are used for prediction. Bouwmeester et al. [39] evaluated 36 different metabolomics data sets against 7 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, either 151 or a minimal set of 11 descriptors, were used. Their analysis showed that no single approaches outperforms all other for all evaluated datasets, although GB performed best in most cases. Furthermore, they evaluated ensemble learning integrating different algorithms by simply averaging predicted retention times.

*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. [39]. 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 which nominally uses exactly the same experimental protocol, will be rather limited This may explain the huge number of paper which have been written on retention time prediction over the last decades [27]. So far, only few studies focus on the aspect of transferability.

Zisi et al. [40] performed QSRR for 94 metabolite standards and their results indicated that the inclusion of retention times from a different chromatographic column as an additional descriptor improves prediction accuracy.

Bach et al. [41] performed prediction of retention order instead of retention times. Using a ranking support vector machine (RankSVM) or SVR their evaluated 5 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 RankSVM trained on multiple systems outperforms SVR, with or without training on multiple systems.

Stanstrup et al. [42] developed a system called PredRet, which enable the projection of retention times 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 retention times of metabolites detected in one, but not the other system can be projected.

**Current limitations**

An interesting challenge for future work will be the differentiation between stereoisomers. As MS and tandem MS normally produces identical results, chromatographic separation delivers valuable information. To enable the differentiation three dimensional descriptors are required. Eugester et al. [32] briefly mention this problem. The metabolites isoquercitrin (quercetin-3-*O*-glucoside) and hyperoside (quercetin-3-*O*-galactoside) have different experimental retention times, while their predicted retention time is the same. Although only MS1 annotation was performed in their publication, even MS2 would not help. Cao et al. [29] described to have detected five peaks with the m/z 132.1023. Leucine and isoleucine have been recorded for their dataset. The prediction error of 0.68 min for their current model did not allow direct application, since the chromatographic resolution is beyond the model. Since direct annotation was not possible, they searched PubChem for compounds matching the calculated formula C6H13NO2. Out of 970 retrieved compounds only 18 compounds were predicted to elute in the range of an unknown peak with the m/z 132.1023. Furthermore, the authors used retention time prediction to filter putative false positive annotations, which arise from in-source fragments. First steps in this direction have been undertaken by the use of the VolSurf+ three dimensional descriptors [31, 34, 35].

Most of the retention time predictions have been used to filter annotations based on exact mass matching or formula look up in databases using only MS1 information. It remains to be elusive how far comparison of MS2 spectra for the substances with library spectra or the use of *in silico* methods for MS2 analysis have allowed to come to the same conclusion. Integration of retention time prediction with MS2 analysis tools are not common yet. Samaraweera et al. [43] 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 using PubChem as database. Using the smaller HMDB as 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 for PubChem through the integration of retention time predictions. These results were obtained using 78 compounds as “unknowns”. It remains elusive how exact retention time prediction can be integrated with MS2 annotation workflows.

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

Lastly, prediction of retention times 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 errors and hinder more detailed QSRR modeling. Furthermore, most species analyzed in metabolomics might be present as charged species, which is of great importance especially for HILIC based separations. Molecular fingerprints, such as (counting) fingerprints or fingerprints used by Domingo-Almenara et al. [38], may represent valuable alternatives once sufficient training data becomes available.

**Conclusion**

Retention time prediction is getting increasing 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 retention times using QSRR models based on different molecular descriptors and different machine learning approaches. Utilized approaches are similar variable as the different separation methods used in LC-MS based metabolomics.

So far, most approaches have been focusing on RP- based data. Although, HILIC was one of the first datasets used for retention time prediction in metabolomics [28], it is currently neglected. Retention time prediction in HILIC will be more complicated, since several kinds 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 for the separation having a strong effect on the analyte retention. However, HILIC is getting more and more attention as a method for 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, with correct appropriate experimental conditions, the performance of HILIC columns is very reproducible.

With more and more data becoming publicly available through metabolomic data repositories such as Metabolights [26], Metabolomics Workbench [44] or Global Natural Products Social Molecular Networking (GNPS) [45] the amount of available training data for retention time prediction also increases. Therefore, the future of retention time prediction and integration into metabolite identification workflows is bright and new developments and approaches will hopefully be developed in the near future.

**Acknowledgment**

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

**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 DOI: 10.1038/nrm3314.

[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. *Analytical Chemistry* 2012, *84*, 5992-5999 DOI: 10.1021/ac300716u.

[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. *Environmental Microbiology* 2012, 398-408 DOI: 10.1111/j.1462-2920.2012.02840.x.

[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, 2380-2396 DOI: 10.1038/ismej.2014.79.

[5] da Silva, R. R., Dorrestein, P. C., Quinn, R. A., Illuminating the dark matter in metabolomics. *Proceedings of the National Academy of Sciences* 2015, *112*, 12549-12550 DOI: 10.1073/pnas.1516878112.

[6] Mahieu, N. G., Patti, G. J., Systems-Level Annotation of a Metabolomics Data Set Reduces 25 000 Features to Fewer than 1000 Unique Metabolites. *Analytical Chemistry* 2017, *89*, 10397-10406 DOI: 10.1021/acs.analchem.7b02380.

[7] Baran, R., Untargeted metabolomics suffers from incomplete raw data processing. *Metabolomics* 2017, *13*, 107 DOI: 10.1007/s11306-017-1246-3.

[8] Pence, H. E., Williams, A., ChemSpider: An Online Chemical Information Resource. *Journal of Chemical Education* 2010, *87*, 1123-1124 DOI: 10.1021/ed100697w.

[9] 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. *Analytical Chemistry* 2018, *90*, 3156-3164 DOI: 10.1021/acs.analchem.7b04424.

[10] 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* DOI: 10.1002/jms.1777.

[11] 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 Research* 2018, *46*, D608-D617 DOI: 10.1093/nar/gkx1089.

[12] 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.

[13] 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 Research* 2016, *44*, D1202-D1213 DOI: 10.1093/nar/gkv951.

[14] 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. *Journal of Cheminformatics* 2015, *7*, 44 DOI: 10.1186/s13321-015-0087-1.

[15] 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. *Journal of Cheminformatics* 2019, *11*, 2 DOI: 10.1186/s13321-018-0324-5.

[16] Wolf, S., Schmidt, S., Muller-Hannemann, M., Neumann, S., In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics* 2010, *11*, 148.

[17] Ruttkies, C., Schymanski, E. L., Wolf, S., Hollender, J., Neumann, S., MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *Journal of Cheminformatics* 2016, *8*, 3 DOI: 10.1186/s13321-016-0115-9.

[18] Schymanski, E. L., Ruttkies, C., Krauss, M., Brouard, C., Kind, T., Dührkop, K., Allen, F., Vaniya, A., Verdegem, D., Böcker, S., Rousu, J., Shen, H., Tsugawa, H., Sajed, T., Fiehn, O., Ghesquière, B., Neumann, S., Critical Assessment of Small Molecule Identification 2016: automated methods. *Journal of Cheminformatics* 2017, *9*, 22 DOI: 10.1186/s13321-017-0207-1.

[19] 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* DOI: 10.1002/rcm.6364.

[20] Allen, F., Greiner, R., Wishart, D., Competitive fragmentation modeling of ESI–MS/MS spectra for putative metabolite identification. *Metabolomics* 2015, *11* DOI: 10.1007/s11306-014-0676-4.

[21] 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* 2015, *112* DOI: 10.1073/pnas.1509788112.

[22] 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 DOI: 10.1007/s11306-007-0082-2.

[23] 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. *Environmental Science & Technology* 2014, *48*, 2097-2098 DOI: 10.1021/es5002105.

[24] 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 DOI: <https://doi.org/10.1016/j.talanta.2017.01.003>.

[25] 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. *Journal of chromatography. A* 2018, *1536*, 50-57 DOI: 10.1016/j.chroma.2017.07.013.

[26] 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 meta-data. *Nucleic Acids Research* 2013, *41*, D781-D786 DOI: 10.1093/nar/gks1004.

[27] Héberger, K., Quantitative structure–(chromatographic) retention relationships. *Journal of Chromatography A* 2007, *1158*, 273-305 DOI: <https://doi.org/10.1016/j.chroma.2007.03.108>.

[28] 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. *Analytical Chemistry* 2011, *83*, 8703-8710 DOI: 10.1021/ac2021823.

[29] 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 DOI: 10.1007/s11306-014-0727-x.

[30] Bruderer, T., Varesio, E., Hopfgartner, G., The use of LC predicted retention times to extend metabolites identification with SWATH data acquisition. *Journal of Chromatography B* 2017, *1071*, 3-10 DOI: <https://doi.org/10.1016/j.jchromb.2017.07.016>.

[31] 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 DOI: 10.1007/s11306-015-0888-2.

[32] 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 DOI: <http://dx.doi.org/10.1016/j.phytochem.2014.10.005>.

[33] Aicheler, F., Li, J., Hoene, M., Lehmann, R., Xu, G., Kohlbacher, O., Retention Time Prediction Improves Identification in Nontargeted Lipidomics Approaches. *Analytical Chemistry* 2015, *87*, 7698-7704 DOI: 10.1021/acs.analchem.5b01139.

[34] 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. *Analytica Chimica Acta* 2016, *916*, 8-16 DOI: <http://dx.doi.org/10.1016/j.aca.2016.02.014>.

[35] 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. *Journal of Chromatography B* 2017, *1071*, 11-18 DOI: <https://doi.org/10.1016/j.jchromb.2017.04.032>.

[36] Snyder, L. R., Dolan, J. W., Carr, P. W., The hydrophobic-subtraction model of reversed-phase column selectivity. *Journal of Chromatography A* 2004, *1060*, 77-116 DOI: <http://dx.doi.org/10.1016/j.chroma.2004.08.121>.

[37] 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.

[38] 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. *Nature Communications* 2019, *10*, 5811 DOI: 10.1038/s41467-019-13680-7.

[39] Bouwmeester, R., Martens, L., Degroeve, S., Comprehensive and Empirical Evaluation of Machine Learning Algorithms for Small Molecule LC Retention Time Prediction. *Analytical Chemistry* 2019, *91*, 3694-3703 DOI: 10.1021/acs.analchem.8b05820.

[40] 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.

[41] Bach, E., Szedmak, S., Brouard, C., Böcker, S., Rousu, J., Liquid-chromatography retention order prediction for metabolite identification. *Bioinformatics* 2018, *34*, i875-i883 DOI: 10.1093/bioinformatics/bty590.

[42] Stanstrup, J., Neumann, S., Vrhovšek, U., PredRet: Prediction of Retention Time by Direct Mapping between Multiple Chromatographic Systems. *Analytical Chemistry* 2015, *87*, 9421-9428 DOI: 10.1021/acs.analchem.5b02287.

[43] 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. *Analytical Chemistry* 2018, *90*, 12752-12760 DOI: 10.1021/acs.analchem.8b03118.

[44] 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 Research* 2016, *44*, D463-D470 DOI: 10.1093/nar/gkv1042.

[45] Wang, M., Carver, J. J., Phelan, V. V., Sanchez, L. M., Garg, N., Peng, Y., Nguyen, D. D., Watrous, J., Kapono, C. A., Luzzatto-Knaan, T., Porto, C., Bouslimani, A., Melnik, A. V., Meehan, M. J., Liu, W.-T., Crüsemann, M., Boudreau, P. D., Esquenazi, E., Sandoval-Calderón, M., Kersten, R. D., Pace, L. A., Quinn, R. A., Duncan, K. R., Hsu, C.-C., Floros, D. J., Gavilan, R. G., Kleigrewe, K., Northen, T., Dutton, R. J., Parrot, D., Carlson, E. E., Aigle, B., Michelsen, C. F., Jelsbak, L., Sohlenkamp, C., Pevzner, P., Edlund, A., McLean, J., Piel, J., Murphy, B. T., Gerwick, L., Liaw, C.-C., Yang, Y.-L., Humpf, H.-U., Maansson, M., Keyzers, R. A., Sims, A. C., Johnson, A. R., Sidebottom, A. M., Sedio, B. E., Klitgaard, A., Larson, C. B., Boya P, C. A., Torres-Mendoza, D., Gonzalez, D. J., Silva, D. B., Marques, L. M., Demarque, D. P., Pociute, E., O'Neill, E. C., Briand, E., Helfrich, E. J. N., Granatosky, E. A., Glukhov, E., Ryffel, F., Houson, H., Mohimani, H., Kharbush, J. J., Zeng, Y., Vorholt, J. A., Kurita, K. L., Charusanti, P., McPhail, K. L., Nielsen, K. F., Vuong, L., Elfeki, M., Traxler, M. F., Engene, N., Koyama, N., Vining, O. B., Baric, R., Silva, R. R., Mascuch, S. J., Tomasi, S., Jenkins, S., Macherla, V., Hoffman, T., Agarwal, V., Williams, P. G., Dai, J., Neupane, R., Gurr, J., Rodríguez, A. M. C., Lamsa, A., Zhang, C., Dorrestein, K., Duggan, B. M., Almaliti, J., Allard, P.-M., Phapale, P., Nothias, L.-F., Alexandrov, T., Litaudon, M., Wolfender, J.-L., Kyle, J. E., Metz, T. O., Peryea, T., Nguyen, D.-T., VanLeer, D., Shinn, P., Jadhav, A., Müller, R., Waters, K. M., Shi, W., Liu, X., Zhang, L., Knight, R., Jensen, P. R., Palsson, B. Ø., Pogliano, K., Linington, R. G., Gutiérrez, M., Lopes, N. P., Gerwick, W. H., Moore, B. S., Dorrestein, P. C., Bandeira, N., Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology* 2016, *34*, 828-837 DOI: 10.1038/nbt.3597.