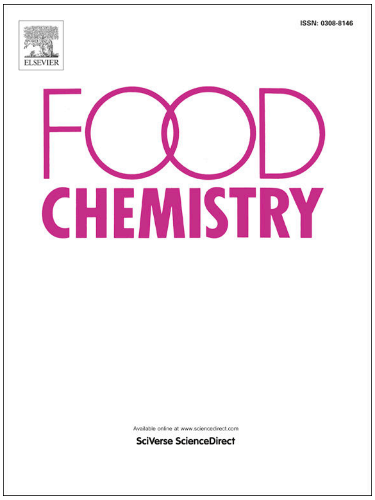

Data sharing in PredRet for accurate prediction of retention time: application to plant food bioactive compounds

Dorrain Yanwen Low, Pierre Micheau, Ville Mikael Koistinen, Kati Hanhineva, László Abrankó, Ana Rodriguez-Mateos, Andreia Bento da Silva, Christof van Poucke, Conceiç ão Almeida, Cristina Andres-Lacueva, Dilip K. Rai, Esra Capanoglu, Francisco A. Tomás Barberán, Fulvio Mattivi, Gesine Schmidt, Gözde Gürdeniz, Kateřina Valentová, Letizia Bresciani, Lucie Petrásková, Lars Ove Dragsted, Mark Philo, Marynka Ulaszewska, Pedro Mena, Raúl González-Domínguez, Rocío Garcia-Villalba, Senem Kamiloglu, Sonia de Pascual-Teresa, Stéphanie Durand, Wieslaw Wiczkowski, Maria Rosário Bronze, Jan Stanstrup, Claudine Manach

PII: S0308-8146(21)00763-9
DOI: https://doi.org/10.1016/j.foodchem.2021.129757
Reference: FOCH 129757

To appear in: *Food Chemistry*

Received Date: 7 January 2021
Revised Date: 29 March 2021
Accepted Date: 6 April 2021

Data sharing in PredRet for accurate prediction of retention time: application to plant food bioactive compounds

Dorrain Yanwen Low[1,29+], Pierre Micheau[1], Ville Mikael Koistinen[2,3], Kati Hanhineva[2,3], László Abrankó[4], Ana Rodriguez-Mateos[5], Andreia Bento da Silva[6,7], Christof van Poucke[8], Conceição Almeida[6], Cristina Andres-Lacueva[9,10], Dilip K. Rai[11], Esra Capanoglu[12], Francisco A. Tomás Barberán[13,14], Fulvio Mattivi[15,16], Gesine Schmidt[17], Gözde Gürdeniz[18], Kateřina Valentová[19], Letizia Bresciani[20], Lucie Petrásková[19], Lars Ove Dragsted[18], Mark Philo[21], Marynka Ulaszewska[15,30], Pedro Mena[22], Raúl González-Domínguez[9,10], Rocío Garcia-Villalba[13], Senem Kamiloglu[12,23], Sonia de Pascual-Teresa[24], Stéphanie Durand[25], Wieslaw Wiczkowski[26], Maria Rosário Bronze[6,27,28], Jan Stanstrup[18]*, Claudine Manach[1]*

[1]Université Clermont Auvergne, INRAE, UNH, F-63000 Clermont Ferrand, France.

[2]Institute of Public Health and Clinical Nutrition, University of Eastern Finland, FI-70211 Kuopio, Finland.

[3]Department of Biochemistry, University of Turku, FI-20014 Turun yliopisto, Finland.

[4]Faculty of Food Science, Department of Applied Chemistry, Szent István Egyetem, 29-43 Villanyi street, 1118 Budapest, Hungary.

[5]Department of Nutritional Sciences, School of Life Course Sciences, King's College London, SE1 9NH London, United Kingdom.

[6]Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa, Av. da República, 2780-157 Oeiras, Portugal.

[7]Faculty of Pharmacy, University of Lisbon, Avenida Professor Gama Pinto, 1649-003 Lisbon, Portugal.

[8]Technology and Food Science Department, Flanders research institute for Agriculture Fisheries and Food (ILVO), Brusselsesteenweg 370, B-9090 Melle, Belgium.

[9]Nutrition, Food Science and Gastronomy Department, Pharmacy faculty, University of Barcelona, Av Joan XXlll, 08028 Barcelona, Spain.

[10]CIBER Fragilidad y Envejecimiento Saludable (CIBERfes), Instituto de Salud Carlos III, 28029 Madrid, Spain.

[11]Department of Food BioSciences, Teagasc Food Research Centre Ashtown, Dublin, D15 KN3K, Ireland.

[12]Faculty of Chemical and Metallurgical Engineering, Department of Food Engineering, Istanbul Technical University, 34469, Maslak, Istanbul, Turkey.

[13]Department of Food Science and Technology, CEBAS-CSIC, Campus Universitario de Espinardo, edf 25, 30100 Murcia, Spain.

[14]Department of Biotechnology, College of Science, Taif University, Taif 26571, Saudi Arabia.

[15]Department of Food Quality and Nutrition, Metabolomics Unit, Research and Innovation Centre, Fondazione Edmund Mach, 38010 San Michele all'Adige, Italy.

[16]Department of Cellular, Computational and Integrative Biology, CIBIO, University of Trento, 38123 Trento, Italy.

[17]Department of Food and Health, Nofima, Norwegian Institute of Food, Fisheries and Aquaculture Research, PB 210, NO-1433 Ås, Norway.

[18]Department of Nutrition, Exercise and Sports, University of Copenhagen, Frederiksberg, DK-1985, Denmark.

[19]Laboratory of Biotransformation, Institute of Microbiology of the CAS, Vídeňská 1083, CZ-142 20 Prague, Czechia.

[20]Human Nutrition Unit, Department of Veterinary Science, University of Parma, Via Volturno, 39, 43125 Parma PR, Italy.

[21]Quadram Institute Biosciences, Norwich Research Park, NR4 7 UQ, United Kingdom.

[22]Human Nutrition Unit, Department of Food and Drug, University of Pharma, Via Volturno, 39, 43125 Parma PR, Italy.

[23]Science and Technology Application and Research Center (BITUAM), Bursa Uludag University, 16059 Gorukle, Bursa, Turkey.

[24]Department of Metabolism and Nutrition, Institute of Food Science, Technology and Nutrition (ICTAN-CSIC), Jose Antonio Novais 10, 28040 Madrid, Spain.

[25]Université Clermont Auvergne, INRAE, UNH, Plateforme d'Exploration du Métabolisme, MetaboHUB Clermont, F-63000 Clermont Ferrand, France.

[26]Institute of Animal Reproduction and Food Research of the Polish Academy of Sciences, Tuwima 10, 10-748 Olsztyn, Poland.

[27]Instituto de Biologia Experimental Tecnológica, Av. da República, Quinta do Marquês, Edificio iBET/ITQB, 2780-157 Oeiras, Portugal.

[28]Research Institute for Medicines, Faculty of Pharmacy, University of Lisbon, Avenida Professor Gama Pinto, 1649-003 Lisbon, Portugal.

[29]Present address: Lee Kong Chian School of Medicine, Nanyang Techological University, 59 Nanyang Drive, Singapore 636921.

[30]Present address: Proteomics and Metabolomics Facility, Center for Omics Sciences, IRCCS San Raffaele Scientific Institute, Via Olgettina n.60 20132 Milan, Italy.

*Claudine Manach and Jan Stanstrup share last co-authorship.

[+]Corresponding author: Dorrain Low; Lee Kong Chian School of Medicine, Nanyang Technological University, 59 Nanyang Drive, Singapore 636921; email: dorrain.low@ntu.edu.sg

Abstract

Prediction of retention times (RTs) is increasingly considered in untargeted metabolomics to complement MS/MS matching for annotation of unidentified peaks. We tested the performance of PredRet (http://predret.org/) to predict RTs for plant food bioactive metabolites in a data sharing initiative containing entry sets of 29–103 compounds (totalling 467 compounds, >30 families) across 24 chromatographic systems (CSs). Between 27 and 667 predictions were obtained with a median prediction error of 0.03–0.76 min and interval width of 0.33–8.78 min. An external validation test of eight CSs showed high prediction accuracy. RT prediction was dependent on shape and type of LC gradient, and number of commonly measured compounds. Our study highlights PredRet's accuracy and ability to transpose RT data acquired from one CS to another CS. We recommend extensive RT data sharing in PredRet by the community interested in plant food bioactive metabolites to achieve a powerful community-driven open-access tool for metabolomics annotation.

Keywords

Predicted retention time, metabolomics, plant food bioactive compounds, metabolites, data sharing, PredRet

1. Introduction

The dark matter in metabolomics refers to the large fraction of molecular signals that are detected with untargeted analyses but remain unidentified. Part of this dark matter corresponds to the food metabolome. Currently, >26,000 compounds have been described in foods (https://foodb.ca), and upon ingestion and digestion, these food components are further transformed into various metabolites (Scalbert et al., 2014), many of which are not identified or inventoried yet in databases (Barabási, Menichetti, & Loscalzo, 2020). Plant food bioactive compounds (also referred as dietary phytochemicals, e.g., (poly)phenols, carotenoids, glucosinolates, alkaloids) and their phase I, -II and gut microbial metabolites represent an important class of the food metabolome that receive widespread interest for their protective health effects and more recently, for their usefulness as food intake biomarkers. They cover a large chemical space ranging from highly polar to lipophilic compounds, and their identification in untargeted methods remains a challenging feat.

Identification of unknowns in untargeted metabolomics combines multiple types of information and tools, such as matching of exact mass in compound databases, comparison of experimental to reference MS^n spectral data and chromatographic retention time (RT) of authentic standards for Metabolomics Standards Initiative (MSI) level I identification, or to publicly available spectral databases for MSI level II (Sumner et al., 2007). But searches in databases often return an excessive number of structurally similar hypotheses (Hall, Hill, Cawley, Hall, Chen, & Grant, 2018), and purchasing all corresponding standards is not feasible due to limited availability and high cost. In the case of plant food bioactive compounds and their metabolites, identification is further challenged by the lack of commercial standards and the high structural similarity between many isomeric compounds, which makes their MS/MS spectra indistinguishable.

Leveraging orthogonal data such as RT becomes valuable for assisting the certainty of identification to MSI levels I and II, by narrowing the number of plausible hypotheses within an observed RT window. Recent years have seen several approaches to adopt RT prediction models for integration into untargeted analysis workflows with varying degrees of success (McEachran, Mansouri, Newton, Beverly, Sobus, & Wiliiams, 2018; Witting & Bocker, 2020). Existing types of RT prediction models include i) simple algorithms based on log P or gradient back-calculation (Abate-Pella et al., 2015; Boswell, Schellenberg, Carr, Cohen, & Hegeman, 2011), ii) monotonically constrained generalised additive model (GAM) (Stanstrup, Neumann, & Vrhovšek, 2015) of retention times and iii) complex *in silico* quantitative structure-retention relationship (QSSR) models based on combinations of molecular descriptors. QSSRs can be built using different machine learning approaches, such as artificial neural network, random forest and support vector regression models (Aalizadeh, Nika, & Thomaidis, 2019; Bade, Bijlsma, Miller, Barron, Sancho, & Hernández, 2015; Bouwmeester, Martens, & Degroeve, 2020; Domingo-Almenara et al., 2019; Hall et al., 2018; McEachran et al., 2018; Naylor, Catrow, Maschek, & Cox, 2019; Tada et al., 2019; Wolfer, Lozano, Umbdenstock, Croixmarie, Arrault, & Vayer, 2015). However, these prediction models are limited in their application, as RT data are specific to one chromatographic system (CS) and the models do not provide accurate predictions outside the trained conditions.

As analytical methods are not harmonised and most laboratories tend to have their own routine semi-targeted or untargeted LC methods for covering plant food bioactive compounds in various types of matrices (serum, plasma, urine, digestive fluids, food materials), it is ideal that RT prediction models be customisable across CSs. PredRet (Stanstrup et al., 2015) represents an original approach that enables users of the scientific community to benefit from RT data sharing through its open access RT database, and obtain predictions in their own CS if the RT of a compound has been experimentally determined by another user or laboratory. In this aspect, PredRet is relevantly applicable for transposing RTs between CSs differing in mobile phase composition, gradient, flow rate and column dimensions. In the framework of the COST Action POSITIVe (https://www6.inra.fr/cost-positive, FA1403), we evaluated the performance of PredRet to predict the RTs of plant food bioactive compounds and their metabolites in a multi-laboratory test involving 19 laboratories across Europe, using 24 gradient-based reversed-phase CSs. We also expanded PredRet database with experimental RTs of 467 plant food compounds.

2. Experimental Section

2.1. Chemical compounds

All participating laboratories purchased their own chemicals, differing from one laboratory to another, except that 10 laboratories previously involved in a multiplatform coverage test organised by the COST Action POSITIVe, received two common standard mixtures comprising of 56 plant food bioactive compounds (Koistinen et al., 2018). Synthesised standards (*n* = 49) were accepted in addition to commercial standards, provided that the structure was unambiguously elucidated by NMR and MS/MS spectra and that the compounds are entered in the online platform for food compound exchange, FoodComEx (https://foodcomex.org/). Depending on laboratories, chemicals were analysed in solvent or spiked in biological matrices (urine or plasma). A full list of the 467 analysed compounds is provided in Table S1, with their common name, InChI, IDs in HMDB, FooDB and PhytoHub, taxonomy, chemical structure, formula, monoisotopic mass, predicted logP and the number of CSs where they were analysed.

Experimental RT datasets containing compound name, InChI and/or chemical structure were provided by the involved laboratories. InChIs were used as unambiguous identifiers for recognition of identical compounds between CSs and compound names were harmonised across laboratories. For polyphenol metabolites, we applied the new reference KCC nomenclature (Kay et al., 2020). InChIs were either extracted from databases such as PhytoHub (http://phytohub.eu), PubChem (https://pubchem.ncbi.nlm.nih.gov) (Kim et al., 2019), HMDB v4.0 (www.hmdb.ca) (Wishart et al., 2018) or computed from chemical structures using Marvin v19.7, 2019, ChemAxon (https://www.chemaxon.com). LogP values were computed using ALOGPS v2.1 (http://www.vcclab.org/lab/alogps/) (Tetko et al., 2005; VCCLAB, 2005) after conversion of InChIs to SMILES via InChIToSMILES (http://www.chemspider.com/inchi.asmx) (Pence & Williams, 2010). In PredRet database, the main InChI layer containing chemical formula, atom connections and hydrogen atom sublayers is considered when matching compounds, and information after the main layer (e.g., charge, stereochemical and isotopic layers) is ignored.

2.2 Chromatographic systems

Experimental RT data were collected from 24 CSs across 19 laboratories. These CSs were not intentionally optimised for the RT prediction test but rather represent the routine semi-targeted or untargeted metabolomic methods of the various laboratories. A full description of instrument, column and analytical conditions used in the 24 CSs is provided in Table 1. Overall, 15 C18 reverse-phase (RP) columns from various manufacturers were used with dimensions ranging from 0.5 to 4.6 mm (internal diameter), 50 to 250 mm (length) and 1.6 to 5 μm (particle size). HPLC or UHPLC methods were used in acidic conditions. Water and acetonitrile acidified with formic acid (0.1-0.9%) or trifluoroacetic acid (0.1%) were most commonly used as mobile phases A and B, while three CSs used methanol or acetone as mobile phase B. The gradients utilised in 13 UHPLC methods consisted of linear and multiphasic slopes with flow rates of 0.4 to 0.6 mL/min and total run times ranging from 6 to 26 min. There were four HPLC methods with multiphasic slopes with flow rates of 0.015 to 1.5 mL/min and longer run times of 20 to 135 min. Figure S1 shows the diversity of gradient slopes in the 24 CS.

2.3 Prediction of retention times

Experimentally measured RTs (Table S3) were entered in PredRet for the 467 compounds listed in Table S1. The number of measured RTs by CS varied from 29 in CS9 to 103 in CS14. For each CS, the compound names, InChI and experimentally measured RTs were entered into PredRet web interface (http://predret.org) along with a description of the respective CS method. PredRet is then able to predict RTs for compounds that have not been previously experimentally measured in one CS but have been determined in some other CS. The prediction is achieved by constructing GAMs between all pairs of CSs in the PredRet database using the compounds that were measured in both CSs. Empirical prediction intervals (PI) were established via bootstrapping of GAMs, as described in more details by Stanstrup et al (2015). The model providing the prediction with narrowest PI was then used. Predictions were flagged as suspicious by the program if the RT is considered potentially incorrect, when the difference between experimental and predicted RTs was ≥ twice the distance from the predicted RT to outer limits of the PI. Predictions were automatically discarded if their PI widths were ≥ 2 min or ≥ 20% of the predicted RT. The total number of RT predictions between CSs, as well as accuracy and coverage of PI relative to the total chromatographic run time, were compared.

2.4. Validation of predicted retention times

A validation test was conducted on CSs 1, 2, 4, 5, 14, 18, 19 and 22, which had the highest number of experimental RT values. These eight CSs comprise of UHPLC and UPLC methods varying in LC instrument and gradient, column, mobile phases, flow rate and run time. The experimental RT datasets of these CSs were split into training sets (80% data, *n* = 79, 71, 73, 67, 82, 63, 63 and 78 compounds respectively) and test sets (20% data, *n* = 20, 18, 18, 17, 21, 16, 16 and 19 compounds respectively). For selection of compounds in the test sets, the datasets were split into three equal sections covering the beginning, middle and end of the chromatographic run, and then 20% of the compounds were randomly selected from the three sections to ensure a uniform distribution of RT along the entire chromatographic run. Another criterion was to select, in the test set, the same proportion of unique compounds as in the whole dataset of the selected CSs. Validation of RT predictions for each of the eight selected CSs was performed in conditions where the complete datasets of the remaining 23 CSs were entered into the PredRet database.

3. Results and Discussion

3.1. Large diversity of plant food metabolites analysed

A total of 1583 experimental RT values were collected for 467 plant food compounds or related human metabolites in one or several of the 24 CSs used by the 19 participating platforms. The 467 compounds belong to >30 families including flavonoids (anthocyanins, flavonols, flavones, flavanols, flavanones, isoflavones), phenolic acids, lignans, ellagitannins, coumarins and furanocoumarins, nitrogen-containing compounds (i.e., alkaloids, amines, indoles), glucosinolates, alkylresorcinols, thiosulfinates, tocopherols, phytosterols, carotenoids and mono, di-, sesqui- and triterpenoids, and their human metabolites, e.g. glucuronidated and sulfated conjugates, as well as gut microbial metabolites. They cover a large chemical space from highly polar to lipophilic with predicted logP values from –3.48 to 10.40 and with monoisotopic masses from 95.0371 to 934.0712 daltons (Figure 1). The PredRet database is growing continuously with addition of new compounds and associated RT data by registered users. At the time of our experiment, a limited number of plant food compounds was present in PredRet, and our datasets represented a major update for this category of compounds.

The number of CSs in which each compound was analysed is provided in Table S1. Of the 467 entered compounds, 212 were analysed in one CS only, while 4'-hydroxy-3'-methoxycinnamic (ferulic), 4-hydroxy-3-methoxybenzoic (vanillic), 3,4-dihydroxybenzoic (protocatechuic), 5-*O*-caffeoylquinic and 4'-hydroxycinnamic (*p*-coumaric) acids were most commonly measured in 20 of the 24 CSs (Figure S2). The size of the datasets varied from 29 to 103 experimental RTs. CSs 1, 2, 4, 5, 7, 14, 17, 18, 19, 22 and 23 contained ≥ 75 RTs, as illustrated by their large node size in Figure 2, in contrast to CS9 and CS16, which contained the least RT data (29 and 35 RTs, respectively). Across the platforms, CSs 2, 6, 11, 13 and 15 shared the highest compound overlap as evidenced by their highly connected nodes (Figure 2) while still showing relatively good overlap with CSs 1, 3, 7, 14, 16, 22 and 23. Pairwise clusters of CSs 18-19 and 4-5 were observed as they shared > 90% compounds similarity, corresponding to two analytical methods from the same platform.

3.2. Retention time prediction coverage and rate

A total of 6382 new RT predictions were obtained for the 24 CSs, with up to 667 predictions for one CS (Table 2 and Figure S3). Compounds that were entered in PredRet prior to this study (1783 unique compounds, ~10% were plant food bioactive compounds) contributed to prediction of additional compounds beyond the 467 compounds entered in this study. We observed a general trend that as more experimental RTs are entered in PredRet, more RT predictions are generated for compounds not previously analysed. This is demonstrated in CSs 1, 2, 22 and 23 where 559, 539, 667 and 572 new RT predictions were generated from 98, 89, 97 and 75 compounds entered into PredRet respectively (Table 2). However, RT prediction was also dependent on shape (Figure S1) and type (i.e., UHPLC or HPLC) of the LC gradient as well as number of common compounds shared with other CSs. For example, infrequently used mobile phases may limit the predictability of a CS. The entry of 29 compounds for CS9 was not sufficient to obtain RT predictions. However, despite relatively small RT datasets (35 to 46 compounds) were entered for CSs 11, 15 and 16, they had a high prediction rate, explained by a versatile CS and/or good combination of compounds.

3.3. Retention time prediction accuracy

PredRet provided RT predictions for compounds never analysed in the CSs but also for compounds in the entry dataset. We used the latter to compare prediction accuracy between CSs. RT predictions were highly accurate across the 24 CSs, with median prediction errors between 0.03 and 0.76 min (Table 2). As run times vary greatly across CSs (5 to 135 min), median prediction errors were also expressed in percentage relative to the total runtime, ranging from 0.3% to 1.8% (CS9 excluded).

A graph comparing experimental and predicted RTs for compounds of CS1 entry dataset is given in Figure 3 as an example. Equivalent graphs for all other CSs are provided in Figure S4. In CS1, accurate predictions with narrow PI were obtained for most compounds with RT ranging between 6.6 and 14.2 min. Predictions for eight compounds (myo-inositol, proline betaine, dopamine, 3,4,5-trihydroxybenzoic acid (gallic acid), 1,3-dimethyluric acid, α-tocopherol, ursolic acid and alkylresorcinol C17:0) were discarded by PredRet algorithm as their PI widths were $\geq$ 2 min or > 20% of the predicted RT. PI width is an important indicator of prediction accuracy as it represents how accurate the projection models are, based on the number of experimentally known RTs in the RT range of compounds that are being projected in the pairwise CS models (Stanstrup et al., 2015). We observed that predictions were usually missing at the beginning and end of the runs, where there tends to be a low density of known RTs, and conditions are approaching the analytical limits of the CSs (Figure S5). In CS1, predictions were not generated before the first 1.5 min and after 14 min. For 15 compounds (1-methylpiperidine, arbutin, 1-methylxanthine, 1*H*-pyrrole-2-carboxaldehyde, cyclo(Leu-Pro), 5-(3’,4’-dihydroxyphenyl)valeric acid, homoeriodictyol, tomatidine, formononetin, bergapten, nobiletin, isosakuranetin, kaempferide, biochanin A and bergamottin), RT prediction was not expected, as they were not present in any other CS. Globally, PredRet performed well for CS1 with a median prediction error of 0.07 min (0.27% of runtime) and median PI width of 0.83 min. For 77 non-unique compounds entered into PredRet, 559 new predictions for compounds never analysed in this system were obtained, in the range of 1.01 to 14.18 min.

Amongst CS7, CS8, CS9, CS14, CS17, CS20 and CS24, a common trait is the high proportion of rare plant compounds unique to their CSs, which indirectly resulted in a low number of common compounds shared with other CSs. For example, CS7 contained sesquiterpenoids not represented in other CSs, likewise for anthocyanin glycosides in CS8, urolithins and conjugated isoflavone metabolites in CS14, glucosinolates and rare flavonoids in CS17, flavonolignans (e.g., dehydrosilydianin), rare flavonoids and sulfated conjugates in CS20, and urolithins and conjugated flavonoids in CS24. Adding the RTs of these rare plant food compounds and metabolites contributes to the richness of PredRet database; however, a caveat is that these CSs themselves may not receive the benefit of good prediction coverage. In such circumstances, the user is encouraged to include common plant compounds that are also frequently represented in the PredRet database. As an example, we propose a list of 14 compounds frequently analysed in our study (≥ 67% of 24 CSs), which covers a wide RT range: 4’-hydroxy-3’-methoxycinnamic acid (ferulic acid), 4-hydroxy-3-methoxybenzoic acid (vanillic acid), 5-*O*-caffeoylquinic acid, 4'-hydroxycinnamic acid (*p*-coumaric acid), 3,4-dihydroxybenzoic acid (protocatechuic acid), 3',4'-dihydroxycinnamic acid (caffeic acid), 3,4,5-trihydroxybenzoic acid (gallic acid), 3',5'-dimethoxy-4'-hydroxycinnamic acid (sinapic acid), (–)-epicatechin, kaempferol, hippuric acid, luteolin, phloretin and hesperetin (Table S4).

To further validate the predictive performance of GAM in PredRet, we performed an external validation test on a subset of eight CSs, splitting the experimental datasets into 80% for training sets and 20% for test sets. The training sets were used to build GAMs between CSs in PredRet database to obtain predictions with PIs for the compounds in the test sets. Predictions were compared to experimental data to obtain the prediction error for each compound (Table S5) and the prediction statistics for each CS are provided in Table S6. Accurate predictions were achieved, with the median prediction error in the test sets ranging from 0.04 to 0.41 min across the eight CSs. The maximum absolute prediction error was 3.55 min for α-tocopherol (CS2), followed by catechol (2.45 min, CS5). It is difficult to compare the performance of PredRet with other RT prediction tools as those only allow predictions within the same CS, while PredRet predicts RTs from one CS to another CS differing in mobile phase composition, gradient and flow rate.

Despite accurate models being built for the CSs, we observed that early- and late-eluting compounds were generally omitted from predictions, likely due to their extreme polarity. Compounds unique to respective CSs (e.g., nobiletin in CS1, 2',5'-dihydroxyphenylacetic acid in CS2 and 9-hydroxy-urolithin-3-glucuronide (isourolithin A glucuronide) in CS14) did not obtain RT predictions as well as compounds that did not have sufficient RT data density in the RT area (e.g., *N*-(3-hydroxybenzoyl)glycine (2-furoylglycine) in CS19 and pinoresinol in CS22). Between 14 and 39% of the compounds in the CSs of the validation (test) set had experimental RTs that fall outside the estimated PI, showing that the PIs should be interpreted with caution, as previously noted in the original paper describing PredRet. The practical implication is that a proposed annotation for an experimental RT cannot be completely discarded even if the RT falls outside the proposed annotation’s PI. A few limitations of PredRet were identified in our study that may be corrected in the future. Firstly, users have no information about the standards that have been considered for providing predictions in their CS: e.g., commercial or synthesised standard, analysed in solvent or spiked in a biological matrix. Secondly, PredRet algorithm recognises the entered compounds based on the main InChI layer only and therefore stereochemical information is ignored during RT prediction.

3.4. Application of PredRet predictions for identification of plant food compounds in metabolomic studies

The effectiveness of RT prediction using PredRet allowed the distinction of isomeric compounds. In Figure 4A, 3-(3',4'-dihydroxyphenyl)propanoic (dihydrocaffeic) acid with a predicted RT of 8.3 min (PI: 8.1 to 8.5 min) could be distinguished from its isomers, 4'-hydroxy-3'-methoxyphenylacetic acid (homovanillic) acid (PI: 8.5 to 9.0 min) and 3,4-dimethoxybenzoic acid (veratric) acid (PI: 9.3 to 9.6 min). In Figure 4B, the predicted RTs of fisetin (PI: 9.8 to 10.7 min), kaempferol (PI: 11.4 to 12 min) and luteolin (PI: 10.6 to 11.3 min) were also clearly distinguished, except for the narrow overlap in the PIs of luteolin and fisetin (10.6 to 10.7 min). This is particularly useful as an orthogonal parameter to eliminate hypotheses when identifying unknown features with the same *m/z* in untargeted metabolomics studies. In addition, as RTs of flavonoid conjugates (glycosides, glucuronides) differ from that of their aglycones, prediction of RT may help to distinguish between aglycones truly present in the samples and detected aglycones that are generated during the analysis as in-source fragments of glycosides or glucuronides.

Another useful application of PredRet is aiding in annotation of rare plant food compounds in untargeted metabolomics studies, when the standards are not commercially available or difficult to synthesise. As soon as a user enters experimental data for a rare plant food compound in a

CSs, PredRet provides RT prediction with PI for this compound in CSs where it has not been experimentally measured. For example, the contribution of tomatidine's experimental RT (11.8 min) from CS1 enabled the prediction of RTs in 15 other CSs, while formononetin (CS1), 8-hydroxy-urolithin-3-sulfate (CS14) and 8-deoxylactucin (CS7) enabled the prediction of RTs in 13 other CSs. To optimise this process, it is crucial that users who entered experimental RTs for rare compounds also enter experimental RTs for common compounds such as those suggested above.

4. Conclusion

PredRet, based on pairwise GAMs, was demonstrated to be a useful tool for obtaining a good number and highly accurate RT predictions for plant food bioactive compounds and their metabolites. Its use in untargeted metabolomics studies can definitely help for tentative identification, by eliminating hypotheses that do not fall within the predicted RT range, or when commercial standards are not readily available. PredRet predictions are precise enough to distinguish structural isomers. Our data sharing initiative and multi-laboratory study contributed to the expansion of the PredRet database with > 1500 experimental RTs in 24 CSs for > 467 plant food bioactive compounds and their metabolites (> 30 families). Importantly, as more experimentally known RTs are entered, more RT predictions are generated and accuracy of the predictions increases. The PredRet database has grown considerably since its introduction and now contains 15,000 RT entries across 68 CSs. Overall, the database covers 4,000 unique compounds, beyond plant food bioactive compounds. In comparison, spectral libraries such as the MassBank of North America (MoNa), contain mass spectra for >200,000 compounds, so there remains a large potential for RT sharing. If sufficiently developed to allow accurate RT prediction in any CSs, PredRet would facilitate comparisons between-studies and minimise the need to develop a consensus LC–MS method for plant food compounds. We thus invite the scientific community to contribute to the community-driven open access PredRet database as part of the global effort for annotation of the dark matter of metabolomes. We suggest that sharing of RT as well as collisional cross section data should be as commonplace in the future as sharing of MS/MS data to provide enough orthogonal data for unambiguous identification in metabolomics.

Acknowledgements: The authors acknowledge networking support by the COST Action FA 1403 POSITIVe (Interindividual Variation in Response to Consumption of Plant Food Bioactives and Determinants Involved), supported by the European Cooperation in Science and Technology (COST). Dorrain Low has received the support of the EU in the framework of the Marie-Curie FP7 COFUND People Programme, through the award of the AgreenSkills+ Fellowship (grant number 609398) and Nanyang Technological University, Singapore, through the award of the Presidential Postdoctoral Fellowship (grant number 001991-00001). Kateřina Valentová and Lucie Petrásková acknowledge financial support from the Czech Science Foundation (grant number 19-00043S). The Faculty of Pharmacy of Lisbon University acknowledges FUNDAÇÃO PARA A CIÊNCIA E TECNOLOGIA and PORTUGAL 2020 for financial support of the Portuguese Mass Spectrometry Network (grant number LISBOA-01-0145-FEDER-402-022125). Kati Hanhineva and Ville Koistinen have received funding from the Academy of Finland (grant numbers 277986 and 312550) and Lantmännen Foundation. Biocenter Finland has financially supported the LC-MS metabolomics unit of University of Eastern Finland. Cristina Andres-Lacueva and Raúl González-Domínguez thank CIBERFES and ISCIII projects AC19/00111 and AC19/00096 (co-funded by FEDER Program from EU, "A way to make Europe"), Generalitat de Catalunya's Agency AGAUR (grant number 2017SGR1546), "Juan de la Cierva" program from MINECO (grant number IJC2019-041867-I) and ICREA Academia award 2018. Francisco A. Tomás-Barberán has received financial support from the Spanish National Research program (grant numbers AGL-2015-73107-EXP/AEI, CSIC 201870E014) and Fundación Seneca Región de Murcia (grant number 19900/GERM/15). Gesine Schmidt acknowledges support through the Norwegian Agriculture and Food Industry Research Funds (grant number 262300). Lars Dragsted and Jan Stanstrup thank the Carlsberg Foundation for a Semper Ardens grant to support this work. László Abrankó acknowledges the Hungarian Academy of Sciences for the János Bolyai Scholarship and support of the EU and ESF co-financed project of SZIU (grant number EFOP-3.6.3-VEKOP-16-2017-00005). Sonia de Pascual-Teresa thanks the Spanish MINECO for financial support (grant number AGL2016-76832-R). The MetaboHUB French infrastructure has supported the INRAE platform (PFEM, MetaboHUB-Clermont) involved in this work (grant number ANR-INBS-0010). Dilip K. Rai acknowledges Teagasc for the financial support through the Walsh Fellowship (grant number ANR-INBS-0010) 2016038).

Author contributions: Experimental data were provided by all authors, and data analysis was conducted by D.Y.L, P.M, J.S and C.M. Manuscript was drafted by D.Y.L and C.M, and reviewed by all authors.

Conflict of interest: Kati Hanhineva and Ville Koistinen are affiliated with Afekta Technologies, Ltd. All other authors declare no conflict of interest.

Supporting information: Further methods and analysis details; Table S1: 467 analysed compounds; Table S2: 24CS methods; Table S3: 24CS experimental RTs; Table S4: frequently measured compounds; Table S5: validation test results; Table S6: validation test summary; Figure S1: %B gradient; Figure S2: RT data density in 24CS; Figure S3: predictions per CS; Figure S4: CS1-24 RT graphs; Figure S5: pairwise GAM graphs.

References

Aalizadeh, R., Nika, M.-C., & Thomaidis, N. S. (2019). Development and application of retention time prediction models in the suspect and non-target screening of emerging contaminants. *Journal of Hazardous Materials, 363*, 277-285. https://doi.org/10.1016/j.jhazmat.2018.09.047.

Abate-Pella, D., Freund, D. M., Ma, Y., Simón-Manso, Y., Hollender, J., Broeckling, C. D., . . . Boswell, P. G. (2015). Retention projection enables accurate calculation of liquid chromatographic retention times across labs and methods. *Journal of Chromatography A, 1412*, 43-51. https://doi.org/10.1016/j.chroma.2015.07.108.

Bade, R., Bijlsma, L., Miller, T. H., Barron, L. P., Sancho, J. V., & Hernández, F. (2015). Suspect screening of large numbers of emerging contaminants in environmental waters using artificial neural networks for chromatographic retention time prediction and high resolution mass spectrometry data analysis. *Science of the Total Environment, 538*, 934-041.

Barabási, A.-L., Menichetti, G., & Loscalzo, J. (2020). The unmapped chemical complexity of our diet. *Nature Food, 1*, 33-37. https://doi.org/10.1038/s43016-019-0005-1.

Boswell, P. G., Schellenberg, J. R., Carr, P. W., Cohen, J. D., & Hegeman, A. D. (2011). Easy and accurate high-performance liquid chromatography retention prediction with different gradients, flow rates, and instruments by back-calculation of gradient and flow rate profiles. *Journal of Chromatography A, 1218*(38), 6742-6749. https://doi.org/10.1016/j.chroma.2011.07.070.

Bouwmeester, R., Martens, L., & Degroeve, S. (2020). Generalized calibration across liquid chromatography setups for generic prediction of small-molecule retention times. *Analytical Chemistry, 92*, 6571-6578. https://doi.org/10.1021/acs.analchem.0c00233.

Domingo-Almenara, X., Guijas, C., Billings, E., Montenegro-Burke, R., Uritboonthai, W., Aisporna, A., . . . Siuzdak, G. (2019). The METLIN small molecule dataset for machine learning-based retention time prediction. *Nature Communications, 10*(5811), 1-9. https://doi.org/10.1038/s41467-019-13680-7.

Hall, M. L., Hill, D. W., Cawley, S., Hall, L. H., Chen, M. H., & Grant, D. F. (2018). Development of a reverse phase HPLC retention index model for nontargeted metabolomics using synthetic compounds. *Journal of Chemical Information and Modeling, 58*, 591-604.

Kay, C. D., Clifford, M. N., Mena, P., McDougall, G. J., Andres-Lacueva, C., Cassidy, A., . . . Crozier, A. (2020). Recommendations for standardizing nomenclature for dietary (poly)phenol catabolites. *The American Journal of Clinical Nutrition, 112*(4), 1051-1068. https://doi.org/10.1093/ajcn/nqaa204.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., . . . Bolton, E. E. (2019). PubChem 2019 update: improved access to chemical data. *Nucleic Acids Research, 47*, 1102-1109. https://doi.org/10.1093/nar/gky1033.

Koistinen, V. M., da Silva, A. B., Abrankó, L., Low, D., Villalba, R. G., Tomás-Barberán, F., . . . Bronze, M. R. (2018). Interlaboratory coverage test on plant food bioactive compounds and their metabolites by mass spectrometry-based untargeted metabolomics. *Metabolites, 8*(46), 1-17. https://doi.org/10.3390/metabo8030046.

McEachran, A. D., Mansouri, K., Newton, S. R., Beverly, B. E. J., Sobus, J. R., & Wiliiams, A. J. (2018). A comparison of three liquid chromatography (LC) retention time prediction models. *Talanta, 182*, 371-379. https://doi.org/10.1016/j.talanta.2018.01.022.

Naylor, B., Catrow, L., Maschek, A., & Cox, J. (2019). QSRR automator: A tool for automating retention time prediction in lipidomics and metabolomics. *Metabolites, 10*(237), 1-15. https://doi.org/10.3390/metabo10060237.

Pence, H. E., & Williams, A. (2010). ChemSpider: An online chemical information resource. *Journal of Chemical Education, 87*(11), 1123-1124. https://doi.org/doi.org/10.1021/ed100697w.

Scalbert, A., Brennan, L., Manach, C., Andres-Lacueva, C., Dragsted, L. O., Draper, J., . . . Wishart, D. S. (2014). The food metabolome: a window over dietary exposure. *American Journal of Clinical Nutrition, 99*(6), 1286-1308.

Stanstrup, J., Neumann, S., & Vrhovšek, U. k. (2015). PredRet: Prediction of retention time by direct mapping between multiple chromatographic systems. *Analytical Chemistry, 87*(18), 9421-9428. https://doi.org/10.1021/acs.analchem.5b02287.

Sumner, L., Amberg, A., Barrett, D., Beale, M. H., beger, R., Daykein, C. A., . . . Viant, M. R. (2007). Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics, 3*(3), 211-221. https://doi.org/10.1007/s11306-007-0082-2.

Tada, I., Tsugawa, H., Meister, I., Zhang, P., Shu, R., Katsumi, R., . . . Chaleckis, R. (2019). Creating a reliable mass spectral-retention time library for all ion fragmentation-based metabolomics. *Metabolites, 9*(251), 1-15. https://doi.org/10.3390/metabo9110251.

Tetko, I. V., Gasteiger, J., Todeschini, R., Mauri, A., Livingstone, D., Ertl, P., . . . Prokopenko, V. V. (2005). Virtual computational chemistry laboratory - design and description. *Journal of Computer-Aided Molecular Design, 19*, 453-363. https://doi.org/10.1007/s10822-005-8694-y.

VCCLAB. (2005). Virtual Computational Chemistry Laboratory. Retrieved from: http://www.vcclab.org Accessed.

Wishart, D. S., Feunang, Y. D., Marcu, A., Guo, A. C., Liang, K., Vázquez-Fresno, R., . . . Scalbert, A. (2018). HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Research, 46*, 608-617. https://doi.org/10.1093/nar/gkx1089.

Witting, M., & Bocker, S. (2020). Current status of retention time prediction in metabolite identification. *Journal of Separation Science, 43*, 1746-1754. https://doi.org/10.1002/jssc.202000060.

Wolfer, A. M., Lozano, S., Umbdenstock, T., Croixmarie, V., Arrault, A., & Vayer, P. (2015). UPLC-MS retention time prediction: a machine learning approach to metabolite identification in untargeted profiling. *Metabolites, 12*(8), 1-13.

Figure captions

Figure 1. Chemical space covered by the 467 plant food metabolites entered in PredRet.

Figure 2. Network map illustrating compound coverage overlap. Size of node represents the number of compounds present in the dataset while the thickness and colour of edges represent the number of common compounds between the paired datasets. The thicker the edge, the larger number of common compounds. Edge colours and denote low (< 10) and high (> 60) similarity of compounds, respectively. E: number of RT data entered into PredRet; P: number of new RT predictions made.

Figure 3. Retention time (RT) prediction accuracy and coverage of 98 compounds with experimentally known RTs in Chromatographic System (CS) 1. Refer to Supporting Information Tables 3 and 4 for more details of individual compounds.

Figure 4. Retention time (RT) prediction for isomers A) 3,4-dimethoxybenzoic (veratric acid), 4'-hydroxy-3'-methoxyphenylacetic (homovanillic) acid and 3-(3',4'-dihydroxyphenyl)propanoic (dihydrocaffeic) acid, and B) kaempferol, luteolin and fisetin in CS1. Coloured areas represent the prediction interval width.

Table captions

Table 1. Instrument and conditions of chromatographic systems used by participating platforms.

Table 2. Statistics of PredRet retention time predictions for 24 liquid chromatographic systems (CSs) with an entry dataset of 467 plant compounds.

Highlights

- Identifying food bioactive compounds in untargeted metabolomics is challenging.
- Predicted retention time is valuable towards effort in metabolite identification.
- 24 Chromatographic systems obtained predicted retention times from PredRet database.
- High accuracy and coverage of retention time predictions for new compounds obtained.
- We recommend extensive retention time data sharing in open access PredRet database.

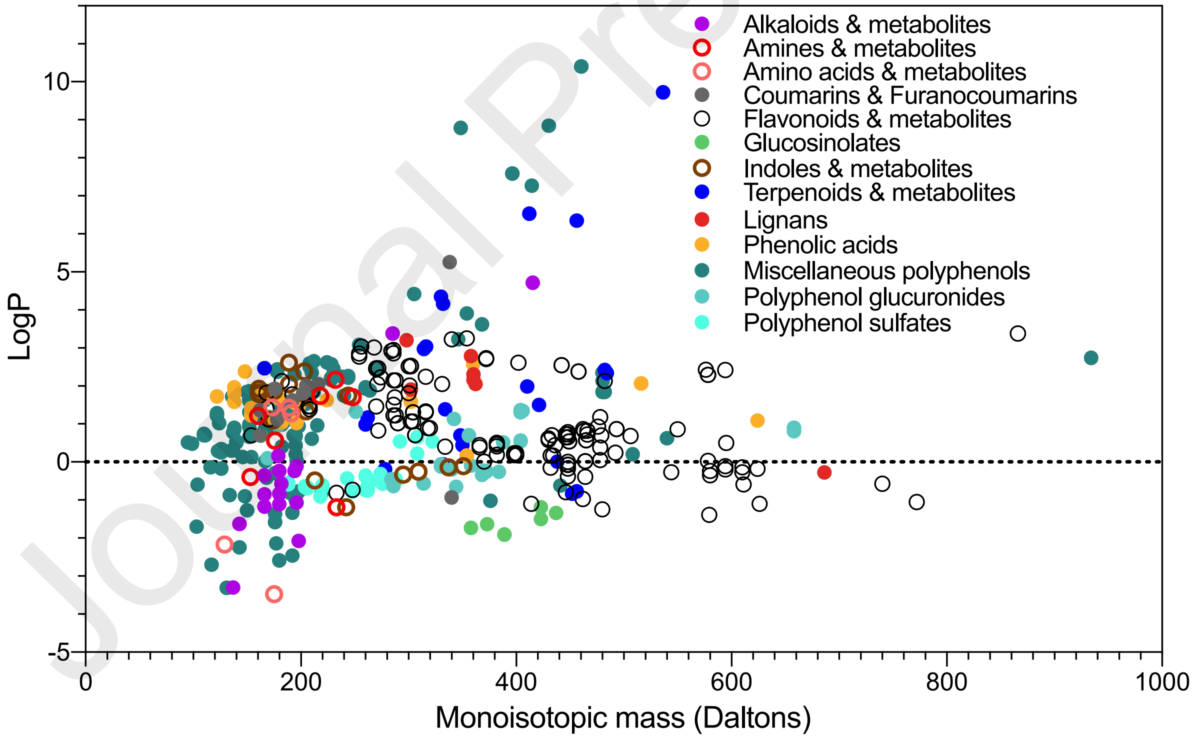

Figure 1. Chemical space covered by the 467 plant food metabolites entered in PredRet.

(Color should be used for figure in print)

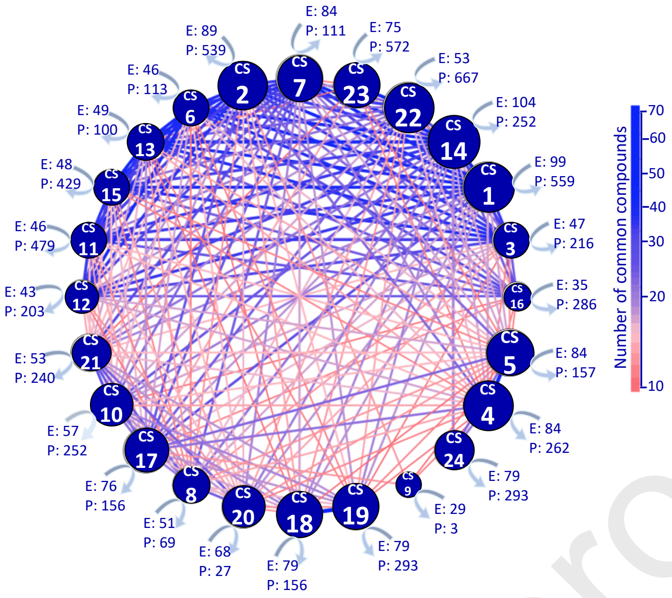

Figure 2. Network map illustrating compound coverage overlap. Size of node represents the number of compounds present in the dataset while the thickness and colour of edges represent the number of common compounds between the paired datasets. The thicker the edge, the larger number of common compounds. Edge colours and denote low (<10) and high (>60) similarity of compounds respectively. E: number of RT data entered into PredRet; P: number of new RT predictions made.

(Color should be used for figure in print)

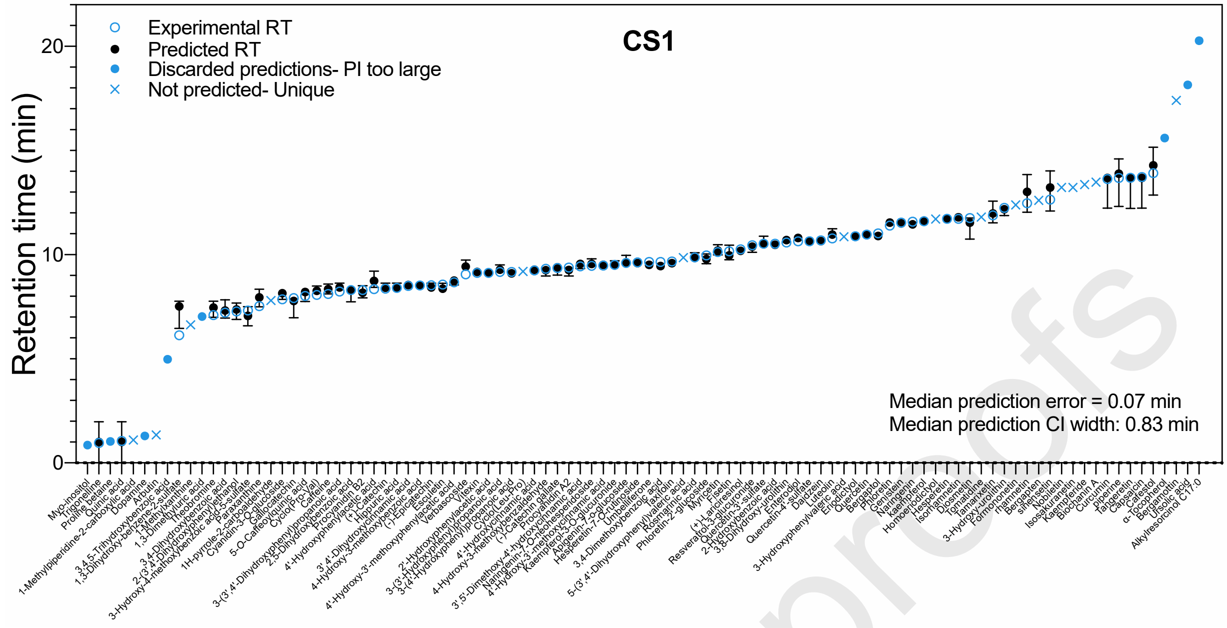

Figure 3. Retention time (RT) prediction accuracy and coverage of 98 compounds with experimentally known RTs in Chromatographic System (CS) 1. Refer to supporting information Tables 3 and 4 for more details of individual compounds.

(Color should be used for figure in print)

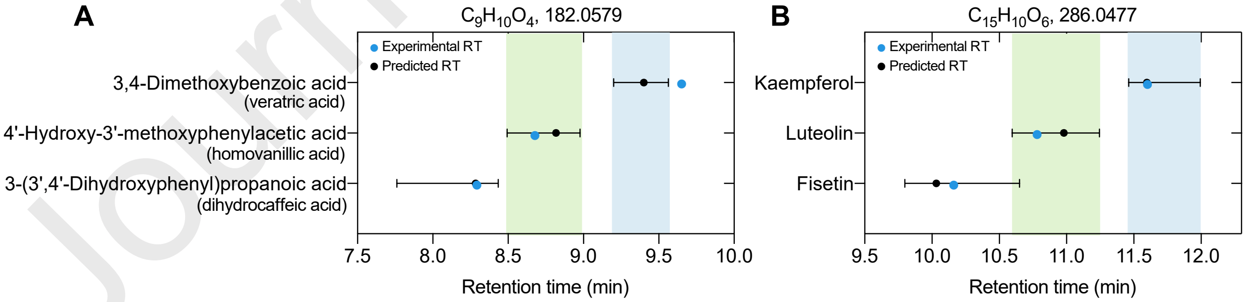

Figure 4. Retention time (RT) prediction for isomers A) 3,4-dimethoxybenzoic (veratric acid), 4'-hydroxy-3'-methoxyphenylacetic (homovanillic) acid and 3-(3',4'-dihydroxyphenyl)propanoic

(dihydrocaffeic) acid, and B) kaempferol, luteolin and fisetin in CS1. Coloured areas represent the prediction interval width.

(Color should be used for figure in print)