Tutor/s

Dra. Anna de Juan Capdevila
Departament d'Enginyeria Química i Química Analítica,
Universitat de Barcelona

Dr. Joaquim Jaumot Soler
Departament de Química Ambiental, IDAEA-CSIC

Grau de Química

Treball Final de Grau

Assessment of chromatographic separations and retention time predictions of polar metabolites.

Avaluació de separacions cromatogràfiques i prediccions de temps de retenció de metabòlits polars.

Míriam Condeminas Rodríguez

July 2020

UNIVERSITAT DE BARCELONA

B:KC Barcelona Knowledge Campus
Campus d'Excel·lència Internacional

Agraïments

Aquest treball no hauria estat possible sense els meus tutors, l’Anna i el Joaquim. Gràcies per la vostra supervisió, i totes les indicacions, crítiques, suggeriments i comentaris.

També vull agrair l’acollida que he tingut en el grup de Quimiometria. Encara que ens hem vist menys del que m’hagués agradat, estic encantada d’haver pogut tenir llargues xerrades davant del cromatògraf amb la Míriam, d’haver compartit esmorzars amb ella i l’Albert, i d’haver treballat al costat de la Carme. Gràcies, també he après molt amb vosaltres.

Finalment, vull agrair als meus pares i familiars el seu suport incondicional.

REPORT

CONTENTS

1. SUMMARY

Metabolomics is an interdisciplinary scientific branch dedicated to the study of the metabolome, the set of low-molecular-weight molecules involved in the chemical reactions of all living creatures. The absence or presence of some metabolites, and their concentrations, enables the assessment of the state in which an organism, tissue or cell finds itself at a particular time. The obtention of this information, which can be extremely relevant from a pharmacological perspective and/or in research contexts, requires the development of analysis methods that detect and unambiguously identify as many metabolites as possible.

In this work, six liquid chromatography coupled to mass spectrometry (LC-MS) methods have been applied to the separation of mixtures containing amino acids, nucleosides and other metabolites, such as carbohydrates and short-chain organic acids. In order to ensure a proper resolution of the mixtures, which contained metabolites of varying polarity, the separations were performed by hydrophilic interaction liquid chromatography (HILIC). The chromatographer was coupled to an electrospray ionization (ESI) source and analytes were detected through MS using a time-of-flight (ToF) analyzer.

The obtained chromatograms were examined and compared through the calculation of two chromatographic response functions (CRFs) to determine which were the best elution conditions. Aiming to increase peak detection and assignment, models relating the annotated metabolites' physicochemical properties to their experimental retention times were generated. Then, the retention times of the unassigned metabolites were predicted. A careful inspection of the chromatograms, taking into account the predictions of the models, allowed for an increase in the detected and annotated metabolites, resulting into more of a 10 % augmentation in the total number of unequivocally assigned peaks.

Keywords: metabolites, LC-MS, HILIC-ESI/ToF, CRF, retention time prediction.

2. RESUM

La metabolòmica és una branca interdisciplinària de la ciència que s'encarrega de l'estudi del metaboloma, el conjunt de molècules de baix pes molecular que intervenen en les reaccions químiques de tots els éssers vius. L'absència o presència de diferents metabòlits, així com les seves concentracions, permeten avaluar l'estat en què un organisme, teixit o cèl·lula es troba en un instant concret. Per obtenir aquesta informació, que pot ser extremadament rellevant a nivell farmacològic i/o en contextos de recerca, és necessari desenvolupar mètodes d'anàlisi que detectin i identifiquin de manera inequívoca tants metabòlits com sigui possible.

En aquest treball s'ha dut a terme la separació de mescles d'aminoàcids, nucleòsids i altres metabòlits, com glúcids i àcids orgànics de cadena curta, amb sis mètodes de cromatografia de líquids acoblada a espectrometria de masses. Per tal de garantir la correcta resolució de les mescles, que contenien metabòlits de polaritat diversa, les separacions es van realitzar mitjançant cromatografia d'interacció hidròfila (HILIC). El cromatògraf estava acoblat a una font d'ionització per electrospray (ESI) i els analits injectats es van detectar per MS utilitzant un analitzador de temps de vol (ToF).

Els cromatogrames obtinguts van ser examinats i comparats mitjançant el càlcul de dues funcions de resposta cromatogràfica (CRFs), que van ser utilitzades per determinar les millors condicions d'elució. Amb l'objectiu d'incrementar la detecció de pics i la seva assignació, es van generar models que relacionessin els temps de retenció experimentals dels metabòlits identificats amb les seves propietats fisicoquímiques. A continuació, es van predir els temps de retenció de les molècules no assignades. Una inspecció més detallada dels cromatogrames, tenint en compte les prediccions dels models, va permetre un augment en la detecció i la identificació de metabòlits, que es va traduir en un increment superior al 10 % en el nombre total de pics assignats inequívocament.

Paraules clau: metabòlits, HILIC-ESI/ToF, CRF, predicció de tems de retenció.

3. INTRODUCTION

Understanding living organisms to the molecular level is of paramount importance in both health and disease. Over the last decades, technological, statistical and informatic improvements have provided tools for the analysis of huge biological datasets, giving rise to the **"-omic" sciences**[1,2]. These fields, named after the particular entities or processes they focus on, are known as **genomics**, **transcriptomics**, **proteomics** and **metabolomics**[1,2], and provide invaluable information that helps unveil the complexity of life.

3.1. METABOLOMICS AND METABOLITES

Metabolomics is the discipline in charge of comprehensively analyzing the metabolome, that is, the complete set of metabolites in a given sample, by means of various techniques[3,4].

Metabolites are low-molecular-weight (MW < 1500 g·mol^{-1}) intermediate or final products formed in metabolic reactions. As such, their absence or presence, and their concentrations, give insight into the state at which a cell, tissue or organism finds itself at a specific moment in time[4]. The human metabolome is comprised of more than 100,000 molecules, including amino acids, small peptides, nucleosides, organic acids, lipids and carbohydrates. Keeping track of so many molecules, their functions and locations, as well as their structural and spectral features is a titanic task. To simplify metabolomics analyses, several databanks, regularly updated, have been created. The Human Metabolome DataBase (HMDB)[5], the Metabolite and Chemical Entity database (METLIN)[6] and MassBank[7] are some important examples.

There are two main strategies to follow when facing a metabolomics problem[3,4,8,9]. On one hand, **targeted metabolomics** aims to quantify a small number of molecules of interest, typically because they are related to a specific metabolic pathway[3,4,8]. The beforehand definition of the analytes to be determined enables the development and optimization of analytical methods using commercial standards[8], when available. On the other hand, **untargeted metabolomics** is more global in scope, its goal being to measure as many known or unknown metabolites as possible[3,4,9]. While quantitation is more precise in targeted experiments[9], untargeted metabolomics covers a wider range of molecular families and yields bigger and more complex datasets[4,9] than targeted metabolomics, which enable the detection of new pathways, for example.

In either approach, it is important to consider sample extraction, treatment and analysis. Among the various currently available methodologies, some, namely nuclear magnetic resonance, require little preparation, whereas others first separate the metabolites based on their physicochemical characteristics (size, polarity, acid/base equilibria…) to later detect them. These include gas chromatography, liquid chromatography and capillary electrophoresis coupled to detection through mass spectrometry[4,10].

3.2. LIQUID CHROMATOGRAPHY COUPLED TO MASS SPECTROMETRY

Technological advancements have made **liquid chromatography coupled to mass spectrometry** (LC-MS) a very robust and widely used methodology in metabolomics[10]. It presents many advantages, including the need for very little quantities of sample, the capacity to separate and detect large numbers of analytes and the possibility to unequivocally identify them based on their mass to charge ratio (m/z)[4,10]. However, its main drawbacks are the difficulty to detect small molecules (MW < 100 $g \cdot mol^{-1}$) and the fact that not all metabolites ionize efficiently under the same conditions.

3.2.1. Hydrophilic Interaction Liquid Chromatography

Even though chromatography dates back to the late 1850's[11], the fundaments of **liquid chromatography** (LC) theory were established in 1941[12]. In LC, a mixture containing the substances to be separated is injected into a column filled with solid particles, the surface of which is commonly covered with a fixed liquid that constitutes the **stationary phase**. The mixture is eluted through the column with a flow of a solution immiscible to the stationary phase referred to as the **mobile phase**. The analytes' separation is based on each compound's different distribution equilibrium between the mobile and the stationary phases. The introduction of packed columns and the reduction of the particles' diameter led to improved chromatographic separations by means of high-performance[13,14] and ultra-high-performance liquid chromatography[15] (HPLC and UPLC, respectively).

The metabolome is formed by a wide variety of molecular families (i.e. carbohydrates, lipids, acids, amino acids, nucleosides…) with different polarities. Reversed phase liquid chromatography (RP-LC), in which the stationary phase is less polar than the mobile phase[16], is ideal for the separation of non-polar compounds such as lipids. Normal phase liquid chromatography (NP-LC), where the most polar phase is the mobile phase[16], successfully resolves polar analytes. The separation of non-polar to very polar molecules was significantly improved in 1990 with the introduction of **hydrophilic interaction liquid chromatography**

(HILIC)[17], which is sometimes classified as a subtype of NP-LC, although its separation mechanism is different and more complex[17,18].

The two main types of **HILIC stationary phases** are based on either silica or polymeric particles[18–20]. Surface-modified silica stationary phases are prepared by chemically binding polar alkoxysilanes to the silanol groups present in the unmodified silica surface[19,21]. These alkoxysilanes may include neutral (e.g. amide, cyano, diol), charged (e.g. amine) or zwitterionic (e.g. sulphobetaine, phosphorylcholine) functionalities. Amide-silica columns, such as TSKgel Amide-80, have been widely applied to separations of highly polar molecules[19], including metabolomics studies[22–25]. This has also been the case for ethylene bridged hybrid (BEH) columns[26], made up from a type of polymeric stationary phase with improved chemical resistance when compared to silica-based particles[27]. As a result, they can be operated under UPLC conditions[28], resulting in shorter analysis times. Figure 1, below, shows the functional groups present in the columns used in the herein described experimental work.

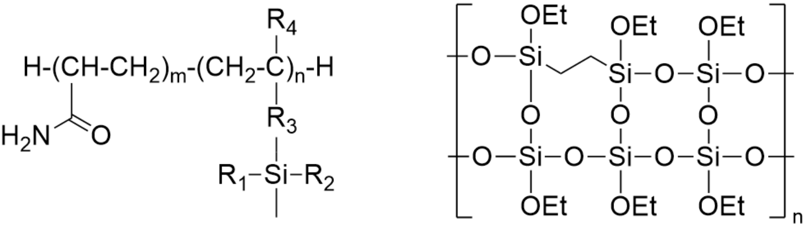


Figure 1: Amide-functionalized silica (left) and BEH (right) HILIC stationary phases.

Typical **HILIC mobile phases**, which ionize very well in MS[29], consist of a hydro-organic mixture with a precisely stablished pH that defines the analytes' ionization state and their polarity[18]. A minimum of 3-5 % v/v of water[17] and 60 % v/v of organic solvent[21] are essential to establish the interactions regulating the HILIC retention process. The most commonly used organic solvent is acetonitrile because it is water-miscible, aprotic and presents a relatively high elution strength[21] when compared to other solvents, such as ethanol or propanol. HILIC separations may be performed in isocratic or gradient elution mode, the latter starting with a small percentage of water that is increased over time[21]. To maintain a stable pH throughout the chromatographic experiment, typically in the 2-8 range, buffer salts soluble in the selected solvents are added to the eluent. Ammonium formate or acetate, highly volatile and soluble[21], are preferred over phosphate salts, as they prevent the obturation of the MS analyzer's entry.

In HILIC, a fixed water-enriched layer is formed in the vicinity of the stationary phase as a result of its high polarity. The analytes' **retention process** is due to partitioning equilibria occurring between this aqueous layer and the mobile phase[17]. Many physical and chemical phenomena (electrostatic forces, dipoles, Van der Waals interactions, hydrogen bonds and hydrophobic interactions) are believed to be involved in this process[18,20]. In gradient elution separations, the first eluting compounds are those with lower polarities. Augmenting the water concentration in the mobile phase also increases hydrophilic interactions, thereby incrementing the affinity of polar analytes for the mobile phase and shortening their retention times[18].

3.2.2. Electrospray ionization/Time-of-Flight Mass Spectrometry

Mass spectrometry appeared at the end of the XIX century as an unexpected result of the cathode rays experiments[30] that proved particle theory with the discovery of the electron and its mass determination[31]. These works were the first charge-to-mass ratio (e/m) measurements ever recorded and led to the construction of the first proper mass spectrometer[32], which was used to quantify the masses of charged atoms. In the following years, the instrument was further developed and helped to prove the existence of elemental isotopes[33]. It was not until the late 1940's, when spectrometers became commercially available, that mass spectrometry started to be applied in other experimental sciences and chemists became aware of its potential for structural elucidation, as well as molecular characterization and identification[31].

The most used ionization method in LC-MS is **electrospray ionization** (ESI)[34] because it enables direct coupling with the LC's column effluent[35] either directly or through a flow divider. ESI is a soft desorption ion source that typically does not fragmentate molecular ions. It operates under atmospheric conditions of pressure and temperature. In a first step, the effluent goes through a nebulizing needle separated from a capillary electrode which generates a difference of potential between the two. This results in a high electric field charging the liquid's surface and forming a spray of droplets that are attracted to the capillary's entrance. A counterflow of gaseous nitrogen evaporates the droplets' liquid content, desolvating them to the point in which the repulsive electrostatic forces surpass the surface tension (Rayleigh's limit) and a coulombic explosion generates smaller droplets. This process takes place until gaseous molecular ions are formed and enter the capillary, where they are oriented, by application of electromagnetic fields, to reach the analyzer contained within a vacuum system[34,35].

$$E_{\mathrm{K}} = \frac{1}{2} \cdot m \cdot v^2 \quad \text{(Equation 1)}$$

E_{K}: kinetic energy, m: mass, v: velocity

One of the best and more used mass analyzers is the so-called **“time-of-flight”** (ToF) **analyzer**[36,37]. Its working principle is that all generated ions have the same kinetic energy (E_{K}, see Equation 1 above), so lighter ions travel faster than heavier ones and reach the detector earlier. The analyzer’s precision is increased in many instruments through reflexion of the generated ions inside the flight tube[37]. The ToF analyser is widely used, as it can be coupled to other MS analyzers for tandem MS[37]. In addition, it has no upper m/z limit, acquires spectra very rapidly and is extremely sensitive, showing a mass-resolving power that allows for the determination of exact mass-to-charge ratios[37] up to four decimals.

3.3. Quantitative Comparison of Chromatographic Methods

In HILIC, many experimental variables affect the analytes’ retention time. The most important ones are the mobile phase’s composition[20,38] and pH[20], the column temperature[17] and the stationary phase[20,38] used. In LC-MS, the analytes’ peaks are obtained from the total ions chromatogram (TIC) through accurate m/z searches and shown as extracted ion chromatograms (EICs), as in Figure 2, below.

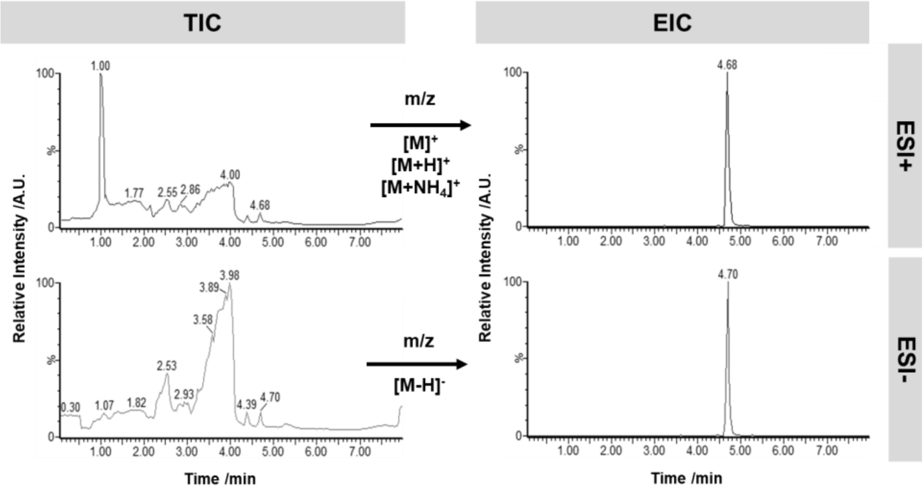


Figure 2: Chromatographic analysis for L-anserine (A8) in Method 3b.

When comparing different experimental conditions and determining which ones yield the best separation and detection, it is common to visually inspect the chromatograms and to calculate each pair of peaks' **chromatographic resolution,** R_i[16], as in Equation 2.

$$R_i = \frac{2 \cdot \left(t_{R,i+1} - t_{R,i}\right)}{w_i + w_{i+1}} \quad \text{(Equation 2)}$$

R_i: resolution; $t_{R,i}$: retention time; w_i: peak width of peak i

Assuming Gaussian peak shapes, R_i reflects the quality of the separation of two contiguous analytes; for instance, values of 0.5 and 1.0 translate to peak overlaps of around 16 and 2 %, and purities of 82 and 98 %, respectively[39]. Thereby, chromatographic peak resolution reflects the quality of the separation of two contiguous analytes but it is not an indicator of the whole chromatogram's quality[40]. Moreover, other factors, such as the number of detected molecules or the analysis time ought to be considered when assessing and quantitatively comparing the chromatographic quality from a set of experimental data[40]. As a result, many **chromatographic response functions**, CRFs, that is, mathematical equations which convert some of the previously mentioned criteria into a single measurable value, have been developed[40].

$$\mathrm{CRS} = \left[\sum_{i=1}^{N-1} \frac{\left(R_i - R_{opt}\right)^2}{R_i \cdot (R_i - R_{min})^2} + \sum_{i=1}^{N-1} \frac{(R_i)^2}{(N-1) \cdot \bar{R}^2}\right] \cdot \frac{t_\mathrm{L}}{N} \quad \text{(Equation 3)}$$

N: number of peaks; R_i: resolution between two adjacent peaks;

R_{opt}, R_{min}: optimal and minimum acceptable resolution; $\bar{R}$: mean resolution of the chromatogram

The **chromatography resolution statistic**, CRS, shown in Equation 3, was proposed by Schlabach and Excoffier in 1988[41] and may be used when comparing chromatograms with varying number of peaks in both untargeted and targeted analysis[40]. In a CRS comparison, the best experimental conditions are those yielding the smallest values, as can be deduced by analyzing its mathematical expression. The first summation in the equation tends to zero when the experimental resolutions (R_i) are similar to the analyst-defined optimal resolution (R_{opt}), while the second summation tries to ensure that peaks are equally distributed throughout the chromatogram. As long methods may result in better separations, both summations are corrected by the application of a factor which attempts to minimize the CRS value by favoring methods with many peaks and short analysis times (found in the denominator and numerator of the correction factor, respectively).

$$\mathrm{CRF(B)} = \sum_{i=1}^{N-1} R_i + N^{\alpha} - \beta \cdot |t_{\mathrm{M}} - t_{\mathrm{L}}| - \gamma \cdot |t_0 - t_1| \quad \text{(Equation 4)}$$

N: number of peaks; α,β,γ: chromatographic weighting factors; R_i: resolution between two adjacent peaks; t_0, t_{M}: minimum and maximum acceptable analysis time; t_1, t_{L}: retention time of the first and last peaks

Another very popular CRF was proposed by **Berridge** in 1982[42] and may be applied for the optimization of experimental conditions that result in chromatograms with a varying number of detected analytes[40]. As shown in Equation 4, it uses three different weighting factors (α, β, γ) which enable the analyst to modulate the relative importance of the number of peaks and the total analysis time against the summation of chromatographic resolutions. On one hand, the two first terms of the equation increase the CRF(B) value with growing resolutions and number of detected peaks. On the other hand, the two last terms decrease the CRF(B) value if the maximum acceptable analysis time (t_{M}) and the retention time of the last peak (t_{L}) are very different, and/or if the minimum acceptable analysis time (t_0) and the retention time of the first peak (t_1) differ significatively. As a result, the best experimental conditions from a set of chromatographic data yield higher CRF(B) values than the rest.

3.4. Retention Time Predictions

The annotation and identification of metabolites in untargeted LC-MS studies is currently based on **accurate mass searches** and, when possible, **structural elucidations** enabled by the analysis of tandem MS data[43]. Ideally, the experimental spectra of a given analyte is compared to reference spectra contained in libraries such as the HMDB[5], METLIN[6] or the MassBank[7]. Accurate mass searches yield long lists of candidate metabolites, even when accounting for their fragmentation patterns, which can be almost identical for similar molecules[43]. In addition, there is an important lack of available information because only a fraction of the metabolome is well documented, the experimental conditions of the reported spectra are not standardized and some metabolic areas are significantly poorly covered[44]. Consequently, many detected peaks may remain unidentified. In an attempt to improve the assignation ratio, **chromatographic retention time predictions** using quantitative structure-retention relationship *in silico* modellings have been proposed[43]. These calculations relate metabolites' physicochemical properties to their retention times under specific chromatographic conditions. To date, very few examples of such modelling approaches have been applied to HILIC metabolomics studies[43,45–47].

Retip is a recently developed and freely available R software-based package for retention time prediction aimed at facilitating peak annotations in RP-LC and HILIC MS metabolomics analyses[47]. Its predictions are based on physicochemical and structural molecular descriptors. The most relevant for HILIC RT predictions[47] have been found to be the octanol/water partition coefficient (XLogP[48]), the number of bonds of the largest π chain, the number of non-rotable bonds (nRotB), molecular shape indices (Kier[49,50]) and pK_a values. These descriptors are calculated by the R-based Chemistry Development Kit platform (rCDK)[51], a tool that computes them from chemical structure identifiers, e.g. the simplified molecular input line entry system (SMILES)[52] and the dashed international chemical identifier (InChIKey)[53]. The Retip app uses many molecular descriptors to generate models with five different non-linear machine learning algorithms to adequately fit the complex datasets derived from chromatographic separations[54,55]. All Retip models have been shown to predict most metabolites' RTs within a range of ±1 minutes from their experimental RTs[47], which in current HPLC and UPLC-MS methods is enough to significatively discard potential candidates and improve peak annotation.

Three of the algorithms included in the Retip app (which can be thought of as "black boxes", the grey-shadowed areas in Figure 3, next page) are the **Bayesian-regularized neural network** (BRNN), the **random decision forest** (RF) and the **extreme gradient boost** (XGBoost).

Artificial neural networks are algorithms formed by one or more layers of connected neurons, the circles in Figure 3A, which are processing units that convert various inputs into a single output through a non-linear parametrized function[54]. The system learns by backpropagation[54], that is, by calculating the errors of each neuron starting from those in the last layer and then using this information to modify their behavior and reduce the global error. The Bayesian approach ensures that the final result is the most probable, given the data used[56].

RF and **XGBoost** are decision tree-based machine learning algorithms applicable to the prediction of continuous values when used in regression mode[55]. Both RF and XGBoost assume that many decision trees combined make up a forest that learns more strongly than the separate trees and cancels out the overfitting individual trees may present[55]. On one hand, RF grows large trees, which are completely independent form each other, in parallel, and gives the same weight to their individual predictions when reaching a final result. On the other hand, XGBoost generates shorter trees sequentially, one after the other, so that each of them accounts for and tries to correct the errors of the previous tree. Contrary to RF, XGBoost gives each tree a different weight according to the quality of its predictions (varying sizes in Figure 3C). Trees yielding values closer to the those considered real have bigger weights, thus improving the model's final results.

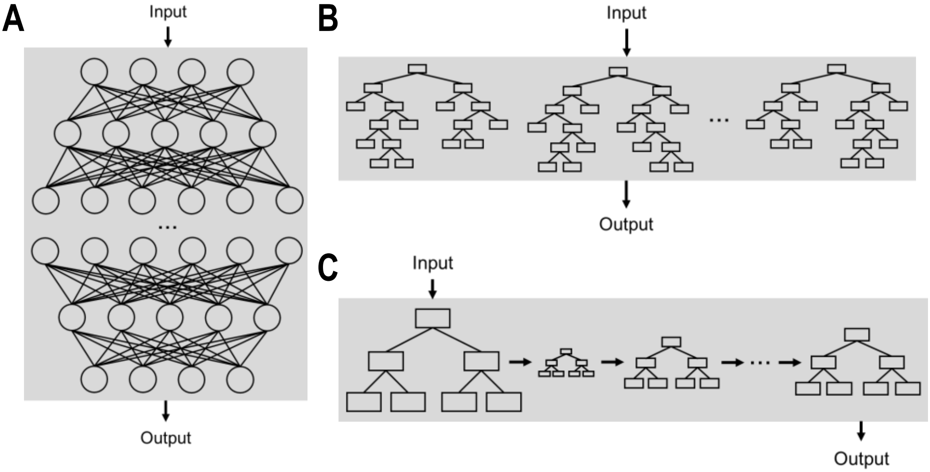


Figure 3: Schemes representing the structure and functioninf of the neural network (A), random forest (B) and gradient boost (C) algorithms.

4. OBJECTIVES

The initial goal of the present project was to **develop and optimize a HILIC-MS method** which could resolve, detect and unequivocally identify as many metabolites as possible so that it could later be applied in untargeted metabolomics analyses. However, the lockdown derived from the SARS-CoV-19 pandemic prevented the realization of much of the planned experimental work.

Consequently, aiming to make the most out of the already acquired HILIC-ESI/ToF experimental data, the objectives were reevaluated and redefined towards:

1. The **analysis and comparison of six different chromatographic methods**, both qualitatively and quantitatively,
2. The **selection of the best elution conditions** and
3. The **retention time prediction** of unidentified metabolites in the best chromatographic conditions so as **to improve peak detection and annotation**.

5. MATERIALS AND METHODS

5.1. SOLVENTS AND REAGENTS

The chemicals used for the preparation of mobile phases were of **analytical grade**: acetonitrile (≥99.9 %, Fischer Scientific), water (ToF quality, Fischer Scientific), hydrochloric acid (37 %, Acros Organics), ammonia and ammonium acetate (28 % and ≥98 %, respectively, Sigma Aldrich), acetic acid (glacial, Panreac) and formic acid (98 %, Merck).

5.2. METABOLITES

The **three different metabolite solutions**, containing amino acids, nucleosides and a mix of various types of metabolites (e.g. saccharides and small organic acids) used are described below. The molecules' structure and identifiers used all over this work can be found in Appendix 1.

Amino acids standard (Sigma Aldrich): β-alanine, L-alanine, L-α-aminoadipic acid, L-α-amino-n-butyric acid, γ-amino-n-butyric acid, D,L-β-aminoisobutyric acid, NH_4Cl, L-anserine, L-arginine, L-aspartic acid, L-carnosine, L-citrulline, creatinine, L-cystathionine, L-cystine, ethanolamine, L-glutamic acid, glycine, L-histidine, L-homocystine, δ-hydroxylysine, hydroxy-L-proline, L-isoleucine, L-leucine, L-lysine, L-methionine, 1-methyl-L-histidine, 3-methyl-L-histidine, L-ornithine, L-phenylalanine, L-proline, L-sarcosine, L-serine, taurine, L-threonine, L-tryptophan, L-tyrosine, urea and L-valine. The analytes were present in a concentration equal to 0.5 $\mu mol \cdot mL^{-1}$ ± 4 % in 0.2 N lithium citrate buffer, pH 2.2, 2 % w/v thiodiglycol and 0.1 % w/v phenol.

Nucleosides test mix (Sigma Aldrich): 50 ppm cytidine, 25 ppm guanosine, 25 ppm inosine, 25 ppm 1-methyladenosine, 100 ppm 5-methylcytidine, 20 ppm 2'-O-methylcytidine, 100 ppm 3-methylcytidine methosulphate, 25 ppm 7-methylguanosine, 50 ppm 5-methyluridine, 25 ppm β-pseudouridine, 10 ppm 2-thiocytidine dihydrate and 25 ppm uridine in 1 % NaCOOH.

Other metabolites (Sigma Aldrich)*:* Individual solutions from solid 3,4-dihydroxy-D,L-phenylalanine, D-gluconic acid sodium salt, L-glutamine, L-ornithine monohydrochloride, sucrose, D-(+)-galactose, citric acid, succinic acid, malic acid, itaconic acid, fumaric acid, pimelic acid, D-maltose, tryptamine hydrochloride, oxidized glutathione, nicotinamide adenine dinucleotide hydrate, D-glucose 6-phosphate sodium salt, D-(-)-ribose, L-norleucine, cytidine, uridine, inosine, dithiothreitol, N-acetyl-cysteine, L-pyroglutamic acid, taurine, 2-ketobutyric acid, α-ketoglutaric acid, uridine 5'-monophosphate disodium salt hydrate, cytidine 5'-monophosphate disodium salt, guanosine 5'-monophosphate disodium salt hydrate and adenosine monophosphate disodium salt were prepared dissolving the chemicals in water to final concentrations of 1,000 ppm.

5.3. Chromatographic Methods

Independent separations of the three different metabolite solutions were carried out in a Waters Acquity UPLC system where autosampler temperature was set at 10 ºC. The analyses were performed using either a **TSKgel Amide-80 HPLC column** (length: 250 mm, inner diameter: 2.1 mm, particle size: 5 µm) from Tosoh Bioscience or a **BEH HILIC Acquity UPLC column** (length: 100 mm, inner diameter: 2.1 mm, particle size: 1.7 µm) from Waters (Figure 1 in page 9). The chromatographic system was connected to a Waters LCT Premier orthogonal accelerated **ToF mass spectrometer** equipped with an **ESI ionization source** operated both in positive (ESI+) and negative (ESI-) modes and acquiring full scan spectra from 80 to 1,800 m/z. The spectrometer working parameters were an electrospray voltage of 3.0 kV or 2.5 kV for ESI+ and ESI- respectively, a sheath gas flow rate of 600 A.U., an auxiliary gas flow rate of 10 A.U. and a heated capillary temperature of 350 ºC.

Six different chromatographic conditions (Methods 1, 2, 3a-c and 4) were assayed. Methods 1[22] and 2[25] had previously been developed and published by IDAEA's Chemometrics group, where the present work was developed. Methods 3a-c are adaptations of Method 2 to UPLC conditions with the objective to **maintain a gradient elution profile starting with a small water percentage while shortening the total run time**. Method 4 is recommended by the UPLC column's manufacturer for the separation of amino acids. All these conditions were assayed and compared aiming to **further optimize** those that yielded better results. A thorough comparison of the methods described below can be found in Appendix 2.

In **Methods 1**[22] and **2**[25], the TSKgel Amide-80 column at room temperature with a flow rate of 0.15 $mL \cdot min^{-1}$ was used with an injection volume of 5 µL (the three metabolites solutions were three-fold diluted with water to final concentrations of 5-40 ppm). The organic component of the mobile phase (solvent A) was acetonitrile, with the aqueous component (solvent B) being 5 mM ammonium acetate pH 5.5. The gradient elution of **Method 1** was 0-8 min 25-30 % B, 8-10 min 30-60 % B, 10-12 min 60 % B, 12-14 min 60-25 %, 14-20 min 25 % B. The gradient elution of **Method 2** was 0-3 min 5 % B, 3-27 min 5-70 % B, 27-30 min 70-5 % B, 32-40 min 5 % B.

Methods 3a-c (adapted from Navarro-Reig *et al.*[25]) and **4** (Gradient Separation of Amino Acids on AQCITY UPLC BEH HILIC, Waters) used the BEH HILIC Acquity UPLC column at 30 ºC. The amino acid standards and nucleosides test mix were injected as is, and the "other metabolites" stock was diluted to 20 ppm (ACN/water 3:2). In **Methods 3a-c**, solvents A and B were the same as for Methods 1 and 2, and a flow rate of 0.3 mL·min^{-1} was used. **Method 3a** had an injection volume of 10 μL and an elution gradient of 0-0.6 min 5 % B, 0.6-5.4 min 5-70 % B, 5.4-6 min 70 % B, 6-7 min 70-5 % B, 7-10 min 5 % B. **Method 3b** had an injection volume of 5 μL and an elution gradient of 0-5 min 5-60 % B, 5-6 min 60 % B, 6-6.5 min 60-5 % B, 6.5-8 min 5 % B. **Method 3c** had an injection volume of 5 μL and an elution gradient of 0-6 min 5-60 % B, 6-7 min 60 % B, 7-7.5 min 60-5 % B, 7.5-9 min 5 % B.

In **Method 4**, a flow rate of 0.4 mL·min^{-1} and an injection volume of 5 μL were used. The organic component of the mobile phase (solvent A) was 90:10 ACN/H_2O 0.2 % HCOOH 10 mM NH_4COOH, while the aqueous component (solvent B) was 50:50 ACN/H_2O 0.2 % HCOOH 10 mM NH_4COOH. The gradient elution of **Method 4** was 0-4.36 min 0.1 % B, 4.36-11.88 min 0.1-99.9 % B, 11.88-13 min 99.9 % B, 13-13.2 min 99.9-0.1 % B, 13.2-15 min 0.1 % B.

The resulting chromatograms were analyzed with MassLynx™ Software (Version 4.1, Waters). **EICs** were obtained from **TICs** by searching for the m/z with a value equal to the exact monoisotopic molecular mass of the loss-of-a-proton adduct ($[M-H]^-$) in ESI-, and the loss-of-an-electron ($[M]^+$), gain-of-a-proton ($[M+H]^+$) or gain-of-an-ammonium group ($[M+NH_4]^+$) adducts in ESI+. The experimental data obtained from the six assayed methods were compared visually and qualitatively through calculation of their **chromatographic resolution statistics**[41] and **Berridge's chromatographic response function**[42].

5.4. Retention Time Predictions

Retention time predictions were calculated using the Retip app[47] as shown in Figure 4, in the next page. The **inputs** used were the InChIKeys, SMILES and experimental retention times of the identified metabolites of **the two best experimental methods**. From the 291 molecular descriptors (MD) the Retip package calculates, 56 were filtered out to eliminate information that was either redundant or constant for the considered metabolites. Three algorithms (BRNN[54,56], RF[55] and XGBoost[55], represented by circles, squares and triangles in Figure 4) were used to **train** models from a random selection of 80 % of the assigned metabolites (to simplify nomenclature, the models will be referred to with the name of the algorithm used to create them). 10 independent permutations of each model were calculated and **validated** with the information (experimental RTs and Retip-calculated MDs) of the 20 % remaining analytes. After model training and validation, each permutations' **relative mean standard error** (RMSE), coefficient of determination (R^2), mean absolute error (MAE) and the 95 % confidence of the predicted RTs (95%±min), in minutes, were calculated. For the three models, the two permutations whose validations yielded **better results, that is, lower RMSEs**, were selected to carry out the predictions. To perform them, the molecular descriptors of the unassigned metabolites (the "test" data in Figure 4) of the two best chromatographic methods, obtained as described above, were used as inputs to **predict the RTs of the unassigned metabolites**, that is, the final **outputs** of the calculations.

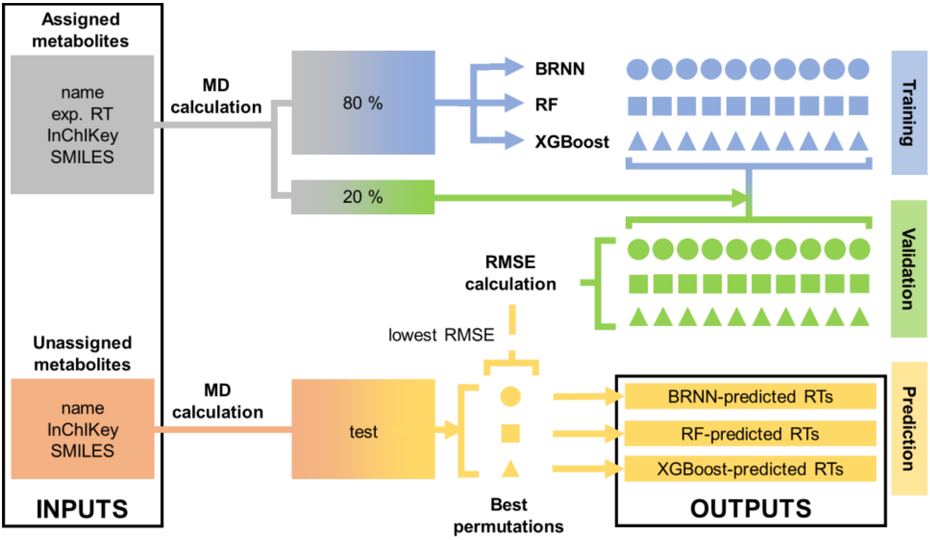


Figure 4: Retention time prediction using the Retip app[47].

The intervals defined by the predicted RTs were reexamined in the experimental chromatograms using the MassLynx™ Software (Version 4.1, Waters) to **unequivocally annotate peaks** generated by analytes with equal monoisotopic molecular masses, as well as to try and **detect previously unfound metabolites**.

6. Results and Discussion

6.1. Peak Detection and Assignment

The data obtained from independent injections of the three different metabolite solutions were examined as described in the experimental section. For all elution methods, the analyzed chromatograms were recorded both in positive and negative ionization modes. In most separations, **ESI- chromatograms showed equal or improved peak detection**, typically by a 10 % increase, when compared to ESI+. This could be due to the fact that metabolites are polar molecules which usually present acid/base equilibria and, thereby, can easily lose a proton and generate the [M-H]$^-$ adduct. All work described from here on in was performed using the datasets from ESI- chromatograms.

So as to compare the **peak detection** of the six assayed elution conditions, the information from the chromatograms corresponding to the amino acids, nucleosides and other metabolites solutions was combined, and is summarized in Table 1. For **HPLC separations**, Method 1 enables the detection of more peaks (N) than Method 2 (for detailed peak detection and annotation, see Tables A1-A3 in Appendix 1). However, Method 1's shorter run time results in peak overlaps and poorer separations when compared to Method 2, as shown by its lower mean chromatographic resolution (R_i) and larger mean peak widths (w_i). **UPLC separations** are more alike among each other than HPLC's, as is to be expected from similar run times and gradient elutions. Methods 3c and 4 were able to detect less peaks than Methods 3a and 3b, which consequently yield worse analyte separations. When comparing the latter, although Method 3a detects more peaks, Method 3b shows higher resolutions and leaner peaks.

	HPLC		UPLC			
Variables	**Method 1**	**Method 2**	**Method 3a**	**Method 3b**	**Method 3c**	**Method 4**
N	52	34	56	52	36	44
% Ass. metabolites	58	37	68	57	40	48
t_{L}	15.90	32.68	5.38	5.50	5.97	8.73
t_1	5.21	6.63	1.01	1.04	1.32	0.76
$\sum R_i$	16.51	59.25	8.50	21.97	26.18	25.27
Mean R_i (min)	0.3 ± 0.4	2 ± 3	0.2 ± 0.2	0.4 ± 0.5	0.7 ± 0.7	0.5 ± 0.7
Mean w_i (min)	0.8 ± 1.1	0.4 ± 0.2	0.6 ± 0.4	0.2 ± 0.2	0.2 ± 0.1	0.3 ± 0.3

Table 1: Relevant data of all assayed chromatographic methods.

Metabolite annotation was performed based on the monoisotopic mass of the loss-of-a-proton adduct. Even though the two columns used presented different stationary phases, the metabolites' distribution along the total run times was comparable in all chromatograms (Figure 5A). Moreover, the elution order and patterns shown by both columns under different elution conditions are practically identical. This is exemplified in Figure 5B, where the peaks of pimelic acid (O15), 1-methyladenosine (N4) and L-phenylalanine (A30) in a HPLC method using an amide-functionalized silica column (Method 1) and an UPLC method with a BEH column (Method 3b) are superimposable. These observations suggest that the assayed HILIC separations are based on **similar retention processes**.

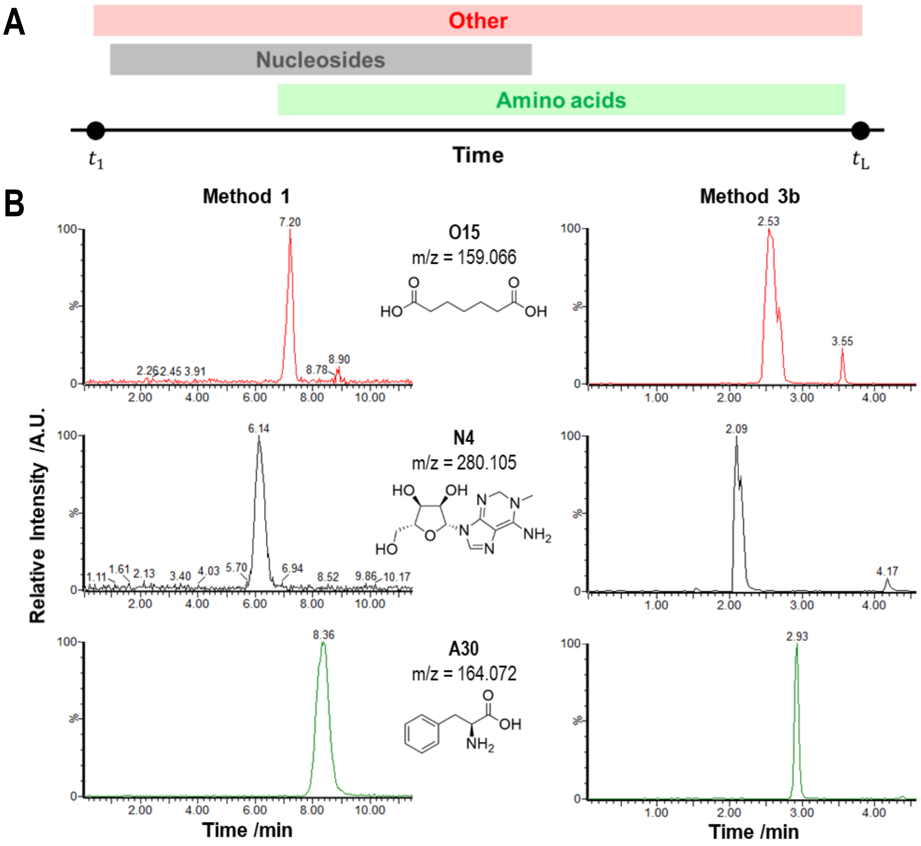


Figure 5: (A) Typical retention time distributions of the injected metabolites.
(B) Selected EICs of pimelic acid (O15), 1-methyladenosine (N4) and L-phenylalanine (A30), extracted from the TICs of Method 1 (left) and Method 3b (right).
The small peaks in O15 and N4 do not contain the m/z of $[M-H]^-$.

Most metabolites were found in at least one of the chromatograms obtained with different elution methods. However, molecules such as β-alanine (A1), L-alanine (A2), ethanolamine (A16), glycine (A18), L-sarcosine (A32) and nicotine adenine dinucleotide (O21) **were not detected in any chromatographic analysis**. But for NADH, which may be poorly ionized under the experimental conditions, their detection was expected to be difficult because they present molecular weights well below 100 g·mol^{-1}.

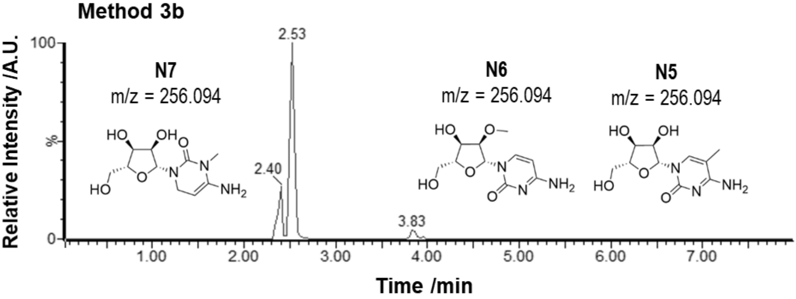


Figure 6: Method 3b's EIC showing the peaks of three isobaric nucleosides whose [M-H]$^-$ mass-to-charge ratio equals 256.094.

The peak assignation of metabolites whose MWs differ from the rest is very high throughput, but whenever two or more **isobars** (molecules with the exact same molecular mass) are present, their annotation based solely on retention times, without additional information such as the fragmentation patterns derived from tandem MS, is a challenge. Figure 6 illustrates this issue with 5-methylcytidine (N5), 2'-O-methylcytidine (N6) and 3-methylcytidine (N7), three isobaric nucleosides whose [M-H]$^-$ monoisotopic mass is 256.0939 g·mol^{-1}. Although three peaks containing ions with this m/z are visible in Method 3b's EIC, it is impossible to assign them without any complementary information.

Metabolite	Method 1	Method 2	Method 3a	Method 3b	Method 3c	Method 4
N3	7.90	-	2.12	2.16	2.09	1.39
O26	7.97	-	2.06	2.13	2.16	1.42

Table 2: Inosine RTs in the nucleoside (N3) and other metabolites (O26) chromatograms.

Retention times were found to be reproducible, even if only one chromatogram was recorded for each metabolite solution and elution method. The "other metabolites" mixture contained two amino acids present in the amino acids commercial solution (L-ornithine, A29/O4, and taurine, A34/O34) and three nucleosides (cytidine, N1/O24, inosine, N3/O26, and uridine, N12/O25) also found in the nucleosides mix. As an example, inosine's RTs, given in Table 2, differ only in 0.05 ± 0.02 min (Experimental RTs can be found in Tables A1-A3 of Appendix 1, where rows corresponding to the repeated metabolites have been shaded with the same color).

6.2. Quantitative Comparison of Chromatographic Methods

In an attempt to establish the methods providing the best separations, the **chromatography resolution statistic**, CRS, and **Berridge's chromatographic response function**, CRF(B), of all assayed separation conditions were calculated.

The **CRS** values plotted in Figure 7 consider a minimal acceptable resolution (R_{min}) of 0.5 and an optimal resolution (R_{opt}) of 1.5, as suggested by its developers[41]. The results showed that, for HPLC conditions, Method 1 yielded better separations than Method 2, with a difference of one order of magnitude between the two. For UPLC, Methods 3a and 4 performed similarly, in spite of the difference in the number of peaks they detected (56 and 44, respectively). In addition, CRS labeled Method 3b as the worse UPLC set of conditions because 20 analytes eluted in the 2.8-3.6 min region, and Method 3c as the best among UPLC conditions, even though it detected more than 15 metabolites less than Methods 3a and 3b. All of these suggests that CRS gave chromatographic resolution more importance than desired.

$$\mathrm{CRS} = \left[\sum_{i=1}^{N-1} \frac{\left(R_i - R_{opt}\right)^2}{R_i \cdot (R_i - R_{min})^2} + \sum_{i=1}^{N-1} \frac{(R_i)^2}{a \cdot \bar{R}^2}\right] \cdot \frac{t_L}{N}$$

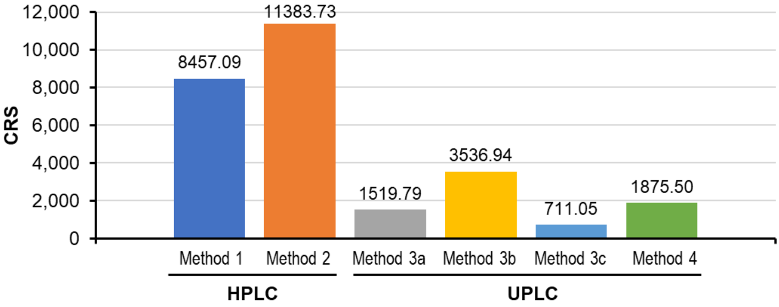
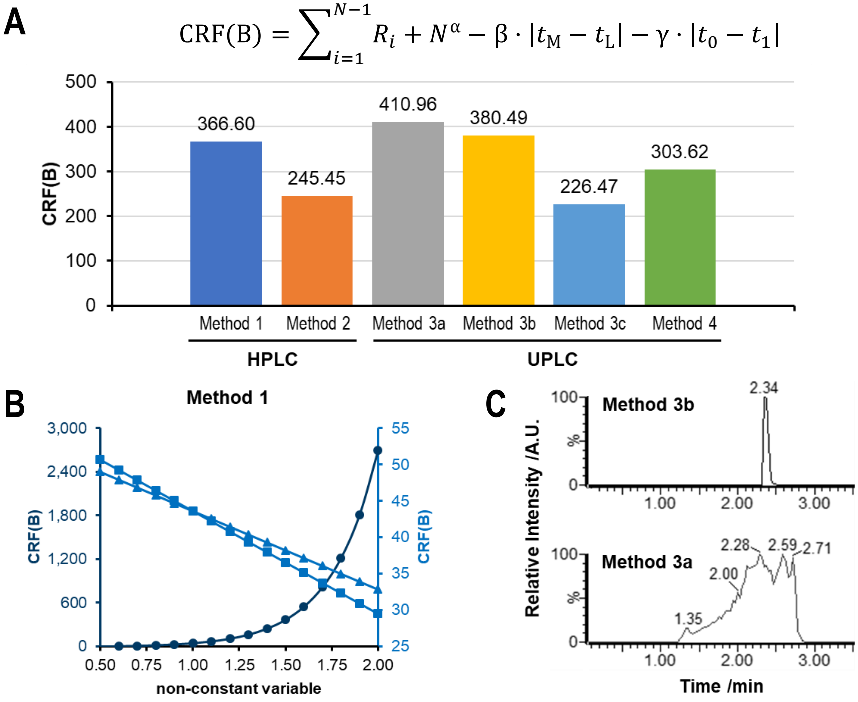


Figure 7: Chromatographic resolution statistic values of all assayed methods.

The **CRF(B)** calculations were carried out keeping either two or one of the three chromatographic weighing factors (α, β and γ) constant and equal to 1.0, while sweeping the other(s) from 0.5 to 2.0. As expected from the function's expression, α was the variable playing the most important role, so much so that when either β or γ were swept, the CRF(B) values were practically only affected by α. Figure 8B exemplifies the general trends with the CRF(B) values of Method 1. In order to account for the different running times of HPLC and UPLC, for Methods 1 and 2, the minimum (t_{M}) and maximum (t_0) acceptable retention times were set at 16 and 30 min, respectively, whereas for Methods 3a-c and 4, the times used were 8 and 15 min. Figure 8A shows the CRF(B) values calculated with $\alpha = 1.5$, to increase the relative importance of the number of detected peaks, and $\beta = \gamma = 1.0$. **Method 1 was confirmed as the best of the two HPLC methods**, as it presented a higher CRF(B) value, detected more analytes and had a shorter analysis time than Method 2. According to CRF(B), Methods 3a and 3b yielded comparable separations. Although Method 3a detected more peaks than Method 3b, they were wider (Figure 8C) and resulted in poorer separations. Taking all these into account, **Method 3b was chosen as the best UPLC method**.



$$\mathrm{CRF(B)} = \sum_{i=1}^{N-1} R_i + N^{\alpha} - \beta \cdot |t_{\mathrm{M}} - t_{\mathrm{L}}| - \gamma \cdot |t_0 - t_1|$$

Figure 8: (A) CRF(B) values of all assayed methods. (B) CRF(B) value variation for Method 1 when sweeping a chromatographic weighing factor (α: ●, β: ■, γ: ▲) while keeping the other two constant and equal to one. (C) Guanosine (N2) peak comparison in Methods 3a and 3b.

6.3. Retention Time Predictions

In either of the two best chromatographic methods, **less than 60 % of the 78 injected metabolites were unequivocally assigned**. Aiming to annotate isobaric analytes whose peaks were visible (like those in Figure 6, page 22) and to increase peak detection, the retention times of unidentified metabolites in Methods 1 and 3b were predicted using the Retip app[47] as described in the experimental section.

The RMSE, R^2, MAE and 95%±min values of the ten generated permutations of each of the three models (BRNN, RF and XGBoost) were calculated. When comparing their RMSEs (Figure 9), all models showed similar prediction properties after validation. The predictions based on training data (calc) always performed better than those used to validate them (val), as expected, and RF and XGBoost behaved very similarly from one another, differing a bit from BRNN's performance. The **permutations whose validations yielded lower RMSEs** and, consequently, better R^2, were used to carry out the RT predictions; permutations 5 and 3 for Methods 1 and 3b, respectively (The calculated statistics of the generated models are collected in Appendix 3.1, where those used for RT predictions are shadowed in light grey).

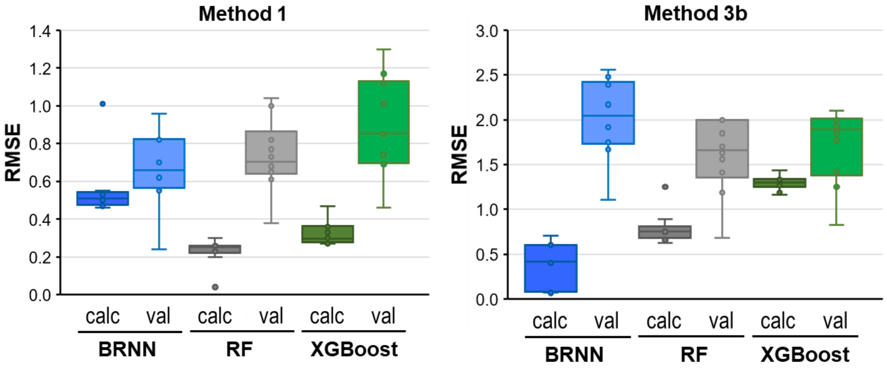


Figure 9: RMSE values of the models' calculation and validation.

Due to the way these three algorithms are built, it is possible to detect the variables (in this work, the molecular descriptors) which are important for the models' generation. The most relevant molecular descriptor in the best permutations was the **octanol/water partition coefficient**, XLogP[48], with a relative importance typically above 90 %. The **number of non-rotable bonds**, nRotB, was also significant, to a lesser extent, in most models. These observations are in agreement with the results of the Retip app developers[47]. However, molecular shape indices (Kier2[49] and Kier3[50]), the number of atoms in the largest π chain (nAtomP) and pK_a values were not so decisive as expected. Moreover, the tspaEfficiency (the molecule's polar surface area divided by its MW) played a key role in Method 1's models, as did the number of basic groups (nBase) in those of Method 3b. These differences may be due to the fact **that a very little number of molecules**, which covered **a small fraction of the total metabolome** and, for the most part, did not present many conjugated or aromatic systems, **were used to calculate and validate the models**. The 20 most important molecular descriptors for each models' best permutation and their relative importance are shown in Appendix 3.2.

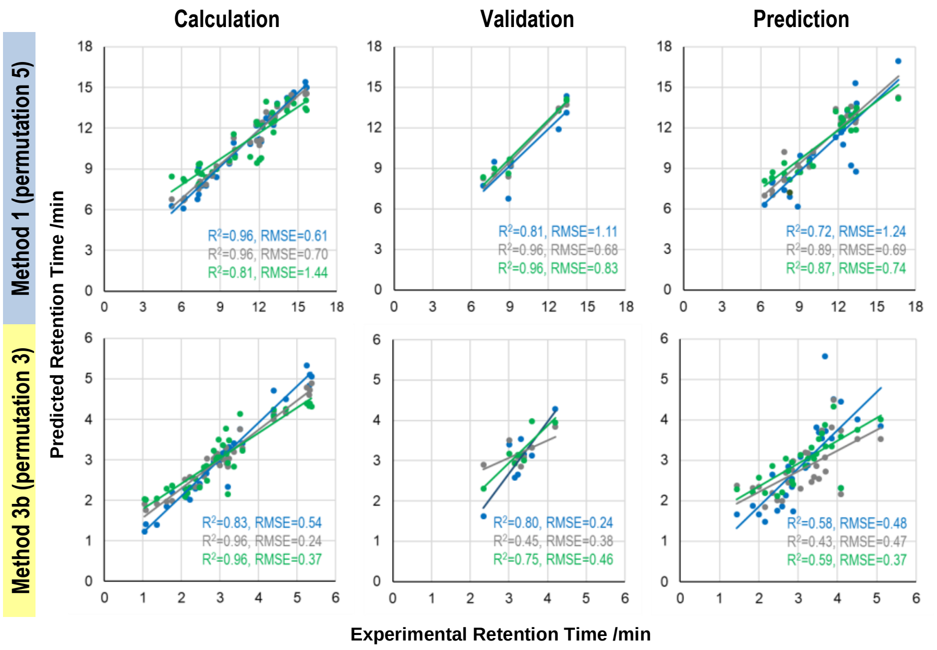


Figure 10: Model calculation and validation, and RT prediction with BRNN, RF and XGBoost.

Figure 10 shows the generation of the models (calculation and validation), as well as their application to the prediction of the retention time of unassigned metabolites. The validation data yielded better results in Method 1 than in Method 3b, which presented lower R^{2}s, specially for the RF model. Still, **the three generated models were able to predict the retention time of most metabolites with an error of around ±1 min**. Once the BRNN, RF and XGBoost models were validated, the retention times of the metabolites which had not been found or unequivocally identified were predicted as described in the experimental section.

The experimental chromatograms were carefully inspected again in search of the unannotated metabolites, this time taking into consideration the intervals defined by the predicted RTs (Appendix 3.3). This allowed for the **finding of the [M-H]$^-$ m/z of some previously undetected metabolites** in the predicted intervals and for the **discernment of some of the isobaric compounds** (Figure 11, shaded with the same colors). For instance, the assignment of the isomers 5-methylcytidine (N5), 2'-O-methylcytidine (N6) and 3-methylcytidine (N7) (EIC shown in Figure 6, page 22) was performed based on the predictions' intervals, which established that the elution order was N7<N5<N6.

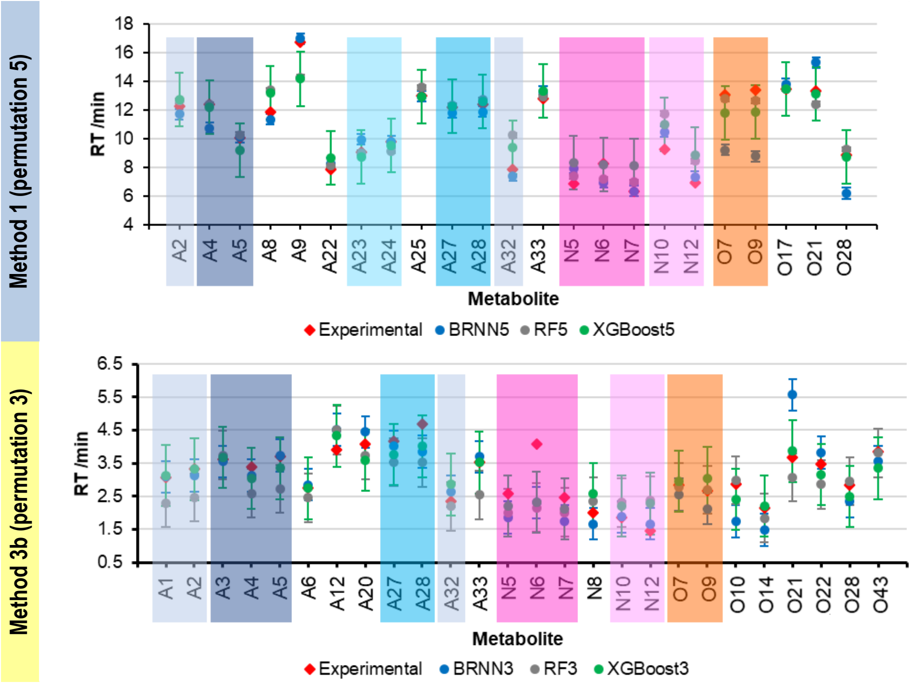


Figure 11: Predicted vs experimental RTs of previously unassigned metabolites.
Error bars plot the 95%±min of the models' validation.

The quality of the predicted RTs for Methods 1 and 3b could also be assessed when the two ESI- peaks with m/z=243.0623, which belong to either β-pseudouridine or uridine, were analyzed. Both the nucleoside and other metabolites solutions contained uridine, labeled as N12 and O25, respectively. From the "other metabolites" chromatogram, it was evident that uridine was the molecule presenting a shorter RT. The predictions not only showed this tendency, but also included the experimental RTs within the models' 95%±min validation.

	BRNN	RF	XGBoost
Method 1	0.5 ± 1.2	0.3 ± 0.7	0.4 ± 0.7
Method 3b	0.5 ± 0.5	0.5 ± 0.5	0.4 ± 0.4

Table 3: Absolute errors of the predicted retention times.

When comparing the performance of the three models (Table 3 and Figure 11, in the previous page), **BRNN yielded the less accurate predictions**. **RF and XGBoost results were very similar**, **but XGBoost presented more trueness**, that is, lower absolute errors, in both Methods 1 and 3b. Even though XGBoost showed lower precision in the validation step, as represented by the error bars of Figure 11, its predictions included the experimental retention times of almost all detected m/z values of the loss-of-a-proton adduct of previously unfound or unassigned metabolites.

The metabolite annotation after the chromatogram's reinspection improved for both methods. In Method 1, it increased from 58 to 69 % due to the assignment of the peaks corresponding to three amino acids, four nucleosides, two carbohydrates (sucrose and D-maltose, O7 and O9, respectively) and a small peptide (oxidized glutathione, O17). In Method 3b, the increase was more pronounced, from 57 to 81 %, because 19 additional peaks were unequivocally annotated to their corresponding analytes. These included eight amino acids, six nucleosides, one nucleotide (adenosine 5'-monophosphate, O22), the two carbohydrates also identified in Method 1 and two small organic acids (citric acid, O10, and fumaric acid, O14). Thus, the RT predictions helped improve the prediction of all injected metabolic families alike.

In summary, **the additional information provided by the retention time prediction allowed for an improvement greater than 10 % in the number of assigned metabolites**.

7. CONCLUSIONS

In this work, hydrophilic interaction liquid chromatography coupled to electrospray ionization and mass spectroscopy detection after a time-of-flight analyzer has successfully been applied to the separation of solutions containing metabolites with varying characteristics. An amide-functionalized silica column and an ethylene bridged hybrid stationary phase have shown superimposable chromatograms, suggesting very similar retention mechanisms. Electrospray ionization in the negative mode has yielded better results for the studied analytes, which were 78 small organic molecules, typically presenting various acid/base equilibria.

Six different gradient elution conditions, all of which used mobile phases whose organic component was acetonitrile, were assayed and compared aiming to find those that resulted in better separations. This was performed through visual inspection and the calculation of indicators of chromatographic quality, such as peak widths, chromatographic resolutions, the chromatography resolution statistic and Berridge's chromatography response function. Careful examination of all this information determined that Method 1, a 20 minute-long HPLC method which uses an amide-functionalized silica stationary phase, and Method 3b, an 8 minute-long UPLC method with an ethylene bridged hybrid column, were the best and yielded similar chromatographic separations. Method 3b starts with a lower water percentage in the mobile phase than Method 1, thereby ensuring that not very polar analytes can be retained and do not elute with the eluent in the dead time. In addition, considering that shorter run times require less reagents and allow for the analysis of more samples in less time, thus reducing the cost of the analysis, Method 3b shows promising applications in future metabolomics studies.

Retention time predictions carried out with different algorithms (Bayesian-regularized neural network, random decision forest and extreme gradient boost) enabled an improvement greater than 10 % in peak detection and annotation. Even though all models showed useful predictive power, and RF and XGBoost performed similarly, XGBoost presented a lower absolute error, making it the best of the three used algorithms. Given the fact that the retention time prediction of unidentified analytes in Method 3b allowed for an improvement above 20 % in the number of unequivocally assigned metabolites, future work could include the application of Method 3b to the analysis of real samples in either targeted or untargeted approaches. It would be interesting to complement those analyses with information provided by XGBoost models, whose predictive power could be further improved by adding experimental data from other metabolic families.

8. References

1. Mishra, N. Science of omics: Perspectives and Prospects for human health care. *Integr. Mol. Med.* **3**, 1–8 (2016).
2. Manzoni, C. *et al.* Genome, transcriptome and proteome: The rise of omics data and their integration in biomedical sciences. *Brief. Bioinform.* **19**, 286–302 (2018).
3. Putri, S. P., Yamamoto, S., Tsugawa, H. & Fukusaki, E. Current metabolomics: Technological advances. *J. Biosci. Bioeng.* **116**, 9–16 (2013).
4. Patti, G. J., Yanes, O. & Siuzdak, G. Innovation: Metabolomics: the apogee of the omics trilogy. *Nat. Rev. Mol. Cell Biol.* **13**, 263–269 (2012).
5. Wishart, D. S. *et al.* HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Res.* **46**, D608–D617 (2018).
6. Guijas, C. *et al.* METLIN: A Technology Platform for Identifying Knowns and Unknowns. *Anal. Chem.* **90**, 3156–3164 (2018).
7. Horai, H. *et al.* MassBank: A public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **45**, 703–714 (2010).
8. Dudley, E., Yousef, M., Wang, Y. & Griffiths, W. J. Targeted metabolomics and mass spectrometry. *Adv. Protein Chem. Struct. Biol.* **80**, 45–83 (2010).
9. Vinayavekhin, N. & Saghatelian, A. Untargeted metabolomics. *Curr. Protoc. Mol. Biol.* **90**, 1–24 (2010).
10. Büscher, J. M., Czernik, D., Ewald, J. C., Sauer, U. & Zamboni, N. Cross-platform comparison of methods for quantitative metabolomics of primary metabolism. *Anal. Chem.* **81**, 2135–2143 (2009).
11. Touchstone, J. C. History of Chromatography. *J. Liq. Chromatogr.* **16**, 1647–1665 (1993).
12. Martin, A. J. P. & Synge, R. L. M. A new form of chromatogram employing two liquid phases. 1. A theory of chromatography 2. Application to the micro-determination of the higher monoamino-acids in proteins. *Trends Biochem. Sci.* **2**, (1977).
13. Huber, J. F. K. & Hulsman, J. A. R. J. A study of liquid chromatography in columns, the time of separation. *Anal. Chim. Acta* **38**, 305–313 (1967).
14. Kirkland, J. J. A High-Performance Ultraviolet Photometric Detector for Use with Efficient Liquid Chromatographic Columns. *Anal. Chem.* **40**, 391–396 (1968).
15. Mazzeo, J. R., Neue, U. D., Kele, M. & Plumb, R. S. Advancing LC performance with smaller particles and higher pressure. *Anal. Chem.* **77**, 460–467 (2005).
16. Ettre, L. S. Nomenclature for chromatography (IUPAC recommendations, 1993). *Pure Appl. Chem.* **65**, 819–872 (1993).
17. Alpert, A. J. Hydrophilic-interaction chromatography for the separation of peptides, nucleic acids and other polar compounds. *J. Chromatogr. A* **499**, 177–196 (1990).
18. Buszewski, B. & Noga, S. Hydrophilic interaction liquid chromatography (HILIC)-a powerful separation technique. *Anal. Bioanal. Chem.* **402**, 231–247 (2012).
19. Jandera, P. Stationary and mobile phases in hydrophilic interaction chromatography: A review. *Anal. Chim. Acta* **692**, 1–25 (2011).
20. Guo, Y. & Gaiki, S. Retention and selectivity of stationary phases for hydrophilic interaction chromatography. *J. Chromatogr. A* **1218**, 5920–5938 (2011).

21. Weiss, J. *Hydrophilic interaction liquid chromatography*. *Handbook of Ion Chromatography, Fourth Edition* (2016).
22. Ortiz-Villanueva, E. *et al.* Assessment of endocrine disruptors effects on zebrafish (Danio rerio) embryos by untargeted LC-HRMS metabolomic analysis. *Sci. Total Environ.* **635**, 156–166 (2018).
23. Ortiz-Villanueva, E. *et al.* Metabolic disruption of zebrafish (Danio rerio) embryos by bisphenol A. An integrated metabolomic and transcriptomic approach. *Environ. Pollut.* **231**, 22–36 (2017).
24. Navarro-Reig, M., Ortiz-Villanueva, E., Tauler, R. & Jaumot, J. Modelling of hydrophilic interaction liquid chromatography stationary phases using chemometric approaches. *Metabolites* **7**, 6–9 (2017).
25. Navarro-Reig, M. *et al.* Metabolomic analysis of the effects of cadmium and copper treatment in: Oryza sativa L. using untargeted liquid chromatography coupled to high resolution mass spectrometry and all-ion fragmentation. *Metallomics* **9**, 660–675 (2017).
26. Grumbach, E. S., Diehl, D. M. & Neue, U. D. The application of novel 1.7 µm ethylene bridged hybrid particles for hydrophilic interaction chromatography. *J. Sep. Sci.* **31**, 1511–1518 (2008).
27. Wyndham, K. D. *et al.* Characterization and Evaluation of C18 HPLC Stationary Phases Based on Ethyl-Bridged Hybrid Organic/Inorganic Particles. *Anal. Chem.* **75**, 6781–6788 (2003).
28. Hsieh, Y., Galviz, G. & Hwa, J. J. Ultra-performance hydrophilic interaction LC - MS/MS for the determination of metformin in mouse plasma. *Bioanalysis* **1**, 1073–1079 (2009).
29. Nguyen, H. P. & Schug, K. A. The advantages of ESI-MS detection in conjunction with HILIC mode separations: Fundamentals and applications. *J. Sep. Sci.* **31**, 1465–1480 (2008).
30. Thomson, J. J. XL. Cathode Rays. *London, Edinburgh, Dublin Philos. Mag. J. Sci.* **44**, 293–316 (1897).
31. Griffiths, J. A brief history of mass spectrometry. *Anal. Chem.* **80**, 5678–5683 (2008).
32. Aston, F. W. LXXIV. A positive ray spectrograph. *London, Edinburgh, Dublin Philos. Mag. J. Sci.* **38**, 707–714 (1919).
33. Aston, F. W. Neon. *Nature* **104**, 334 (1919).
34. Yamashita, M. & Fenn, J. B. Electrospray ion source. Another variation on the free-jet theme. *J. Phys. Chem.* **88**, 4451–4459 (1984).
35. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246**, 64–71 (1989).
36. Wolff, M. M. & Stephens, W. E. A pulsed mass spectrometer with time dispersion. *Rev. Sci. Instrum.* **24**, 616–617 (1953).
37. Marshall, A. G. & Hendrickson, C. L. High-Resolution Mass Spectrometers. *Annu. Rev. Anal. Chem.* **1**, 579–599 (2008).
38. Hao, Z., Xiao, B. & Weng, N. Impact of column temperature and mobile phase components on selectivity of hydrophilic interaction chromatography (HILIC). *J. Sep. Sci.* **31**, 1449–1464 (2008).
39. Bidleman, T. F. The relationship between resolution and percent band overlap. *J. Chem. Educ.* **56**, 293 (1979).

40. Matos, J. T. V., Duarte, R. M. B. O. & Duarte, A. C. Chromatographic response functions in 1D and 2D chromatography as tools for assessing chemical complexity. *TrAC - Trends Anal. Chem.* **45**, 14–23 (2013).
41. Schlabach, T. D. & Excoffier, J. L. Multi-variate ranking function for optimizing separations. *J. Chromatogr. A* **439**, 173–184 (1988).
42. Berridge, J. C. Unattended optimisation of normal phase high-performance liquid chromatography separations with a microcomputer controlled chromatograph. *Chromatographia* **16**, 172–174 (1982).
43. Witting, M. & Böcker, S. Current status of retention time prediction in metabolite identification. *J. Sep. Sci.* **43**, (2020).
44. Frainay, C. *et al.* Mind the gap: Mapping mass spectral databases in genome-scale metabolic networks reveals poorly covered areas. *Metabolites* **8**, (2018).
45. Cao, M. *et al.* Predicting retention time in hydrophilic interaction liquid chromatography mass spectrometry and its use for peak annotation in metabolomics. *Metabolomics* **11**, 696–706 (2015).
46. Creek, D. J. *et al.* Toward global metabolomics analysis with hydrophilic interaction liquid chromatography-mass spectrometry: Improved metabolite identification by retention time prediction. *Anal. Chem.* **83**, 8703–8710 (2011).
47. Bonini, P., Kind, T., Tsugawa, H., Barupal, D. K. & Fiehn, O. Retip: retention time prediction for compound annotation in untargeted metabolomics. *Anal. Chem.* **92**, 7515–7522 (2020).
48. Wang, R., Fu, Y. & Lai, L. A new atom-additive method for calculating partition coefficients. *J. Chem. Inf. Comput. Sci.* **37**, 615–621 (1997).
49. Kier, L. B. A Shape Index from Molecular Graphs. *Quant. Struct. Relationships* **4**, 109–116 (1985).
50. Kier, L. B. Shape Indexes of Orders One and Three from Molecular Graphs. *Quant. Struct. Relationships* **5**, 1–7 (1986).
51. Steinbeck, C. *et al.* The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics. *J. Chem. Inf. Comput. Sci.* **43**, 493–500 (2003).
52. Weininger, D. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
53. Heller, S. R., McNaught, A., Pletnev, I., Stein, S. & Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. Cheminform.* **7**, 1–34 (2015).
54. Marini, F. Neural Networks. in *Comprehensive Chemometrics,* vol. 3, 477–505 (Elsevier Inc., 2009).
55. Stavropoulos, G., van Voorstenbosch, R., van Schooten, F.-J. & Smolinska, A. Random Forest and Ensemble Methods. in *Comprehensive Chemometrics,* 661–672 (Elsevier Inc., 2020).
56. Kubat, M. Probabilities: Bayesian Classifiers. in *An Introduction to Machine Learning,* 19–42 (Springer International Publishing, 2017).

9. ACRONYMS

ACN: acetonitrile

A.U.: arbitrary units

BEH: ethylene bridged hybrid

BRNN: Bayesian-regularized neural network

CRF: chromatographic response function

CRF(B): Berridge's chromatographic response function

CRS: chromatographic resolution statistics

EIC: extracted ion chromatogram

ESI: electrospray ionization

HILIC: hydrophobic interaction liquid chromatography

HMDB: human metabolome database

HPLC: high-performance liquid chromatography

LC: liquid chromatography

LC-MS: liquid chromatography coupled to mass spectroscopy

MAE: mean absolute error

METLIN: metabolite and chemical entity database

$[M\text{-}H]^-$: loss-of-a-proton adduct

MW: molecular weight

MS: mass spectroscopy

m/z: mass to charge ratio

NP-LC: normal phase liquid chromatography

ppm: parts per million ($mg{\cdot}L^{-1}$)

R^2: coefficient of determination

rCDK: R chemistry development kit

RF: random forest

RT: retention time

RMSE: root mean square error

RP-LC: reverse phase liquid chromatography

SMILES: simplified molecular input line entry system

TIC: total ion chromatogram

ToF: time-of-flight

UPLC: ultra-high-performance liquid chromatography

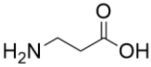
XGBoost: extreme gradient boost

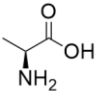
95%±min: 95 % confidence in the predicted retention times

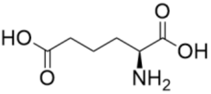
APPENDICES

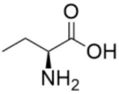
APPENDIX 1: STRUCTURES AND RETENTION TIMES

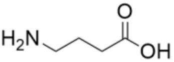
A1.1. AMINO ACIDS

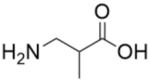

A1 β-alanine

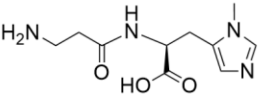

A2 L-alanine

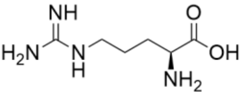

A3 L-α-aminoadipic acid

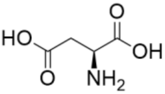

A4 L-α-amino-n-butyric acid

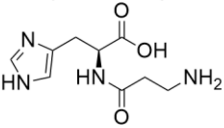

A5 γ-amino-n-butyric acid

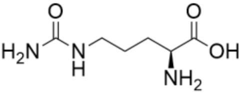

A6 D,L-β-aminoisobutyric acid

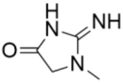

A8 L-anserine

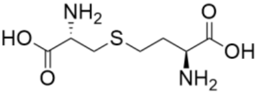

A9 L-arginine

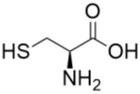

A10 L-aspartic acid

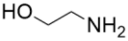

A11 L-carnosine

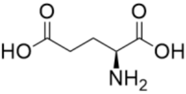

A12 L-citrulline

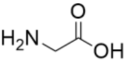

A13 Creatinine

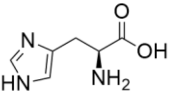

A14 L-cystathionine

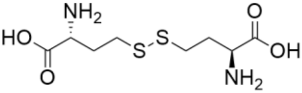

A15 L-cysteine

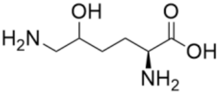

A16 Ethanolamine

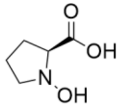

A17 L-glutamic acid

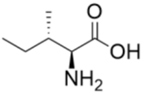

A18 Glycine

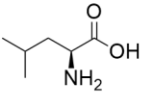

A19 L-histidine

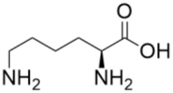

A20 L-homocystine

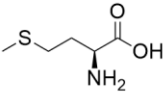

A21 δ-hydroxylysine

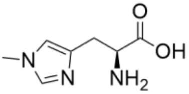

A22 Hydroxy-L-proline

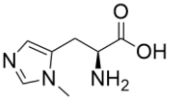

A23 L-isoleucine

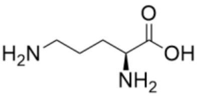

A24 L-leucine

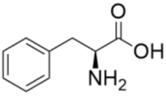

A25 L-lysine

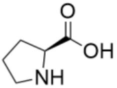

A26 L-methionine

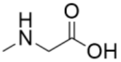

A27 1-methyl-L-histidine

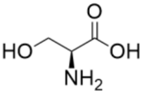

A28 3-methyl-L-histidine

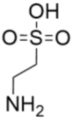

A29 L-ornithine

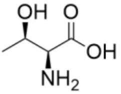

A30 L-phenylalanine

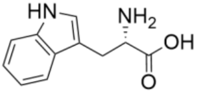

A31 L-proline

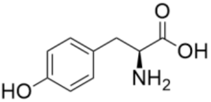

A32 L-sarcosine

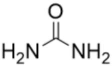

A33 L-serine

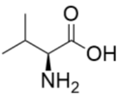

A34 Taurine


A35 L-threonine


A36 L-tryptophan


A37 L-tyrosine

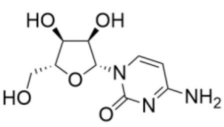

A38 Urea


A39 L-valine

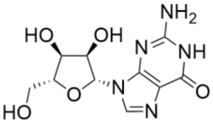
ID	Monoisotopic mass /g·mol^{-1}	Experimental retention times and peak widths /min					
		Method 1	Method 2	Method 3a	Method 3b	Method 3c	Method 4
A1	89.0477	-	-	-	-	-	-
A2	89.0477	-	-	-	-	-	-
A3	161.0688	12.77, 0.23	19.50, 0.32	3.40, 0.26	-	-	6.76, 0.37
A4	103.0633	12.56[a], 0.12	-	-	-	3.46[j], 0.16	-
A5	103.0633	12.56[a], 0.12	-	-	-	3.46[j], 0.16	-
A6	103.0633	12.56[a], 0.12	-	-	-	3.46[j], 0.16	-
A8	240.1222	-	26.20, 0.67	4.77, 0.40	4.71, 0.15	5.15, 0.15	-
A9	174.1117	-	32.68, 0.38	5.26, 1.05	5.25, 0.36	5.78, 0.34	-
A10	133.0375	14.20, 0.50	-	3.32, 0.34	3.08, 0.14	3.31, 0.10	7.40, 0.25
A11	226.1066	15.66, 0.65	24.81, 0.52	4.51, 0.47	4.40, 0.15	4.82, 0.23	8.73, 0.16
A12	175.0957	13.39, 0.32	20.99, 0.38	3.80, 0.34	-	-	7.43, 0.15
A13	113.0589	7.53, 0.73	15.38, 0.22	-	2.65, 0.15	2.75, 0.18	-
A14	222.0674	14.51, 0.50	23.94, 0.67	4.11, 0.60	4.20, 0.24	-	-
A15	121.0197	14.70, 0.75	-	-	-	-	-
A16	61.0528	-	-	-	-	-	-
A17	147.0532	12.55, 0.30	19.19, 0.22	3.35, 0.28	3.15, 0.13	3.40, 0.11	7.07, 0.22
A18	75.0320	-	-	-	-	-	-
A19	155.0695	15.59, 0.45	24.77, 0.42	4.30, 0.90	4.39, 0.14	4.83, 0.16	8.05, 0.19
A20	268.0551	14.16, 0.50	23.01, 0.36	3.93, 0.38	-	4.08, 0.13	7.95, 0.15
A21	162.1004	15.63, 0.60	31.54, 0.68	5.28, 1.18	5.33, 0.44	5.82, 0.33	-
A22	131.0582	-	20.31, 0.34	3.62, 0.49	3.36, 0.10	3.58, 0.10	6.28, 0.32
A23	131.0946	9.05[b], 0.80	17.40, 0.46 or 17.76, 0.33	3.31, 0.40	3.02[h], 0.16	-	3.30[k], 1.10
A24	131.0946	9.05[b], 0.80	17.40, 0.46 or 17.76, 0.33	-	3.02[h], 0.16	-	3.30[k], 1.10
A25	146.1055	-	32.06, 0.94	5.38, 1.23	5.38, 0.45	5.97, 0.34	-
A26	149.0510	9.67, 0.70	18.26, 0.45	3.34, 0.31	3.01, 0.10	3.21, 0.10	3.60, 0.67
A27	169.0851	15.90[c], 1.60	-	4.33, 0.51 or 4.76, 0.17	4.17, 0.25 or 4.70, 0.21	4.51, 0.20 or 5.13, 0.14	-
A28	169.0851	15.90[c], 1.60	-	4.33, 0.51 or 4.76, 0.17	4.17, 0.25 or 4.70, 0.21	4.51, 0.20 or 5.13, 0.14	-
A29	132.0899	9.36, 1.40	-	5.27, 1.12	5.32, 0.37	5.84, 0.30	7.85, 1.00
A30	165.0790	8.36, 0.95	16.90, 0.54	3.24, 0.35	2.93, 0.09	-	3.06, 0.13
A31	115.0633	-	19.35, 0.25	3.59, 0.31	3.39, 0.11	-	-
A32	89.0477	-	-	-	-	-	-
A33	105.0426	-	-	-	-	3.49, 0.14	-
A34	125.0147	10.10, 1.08	18.20, 0.25	2.93, 1.10	2.59, 0.11	2.72, 0.11	-
A35	119.0582	13.02, 0.26	-	-	3.24, 0.17	-	5.70, 1.09
A36	204.0899	7.81, 0.95	-	3.06, 0.45	2.84, 0.14	3.03, 0.13	2.87, 0.30
A37	181.0739	10.01, 0.95	18.17, 0.50	3.28, 0.39	2.96, 0.11	3.15, 0.11	3.21, 0.26
A38	60.0324	-	-	-	-	-	-
A39	117.0790	11.30, 0.80	-	3.46, 0.44	3.16, 0.12	-	-

Table A1: Amino acids' experimental retention times and peak widths in all assayed chromatographic conditions. Shaded analytes are also present in the "other metabolites" solution. Isobaric molecules are indicated with the same superscript letter.

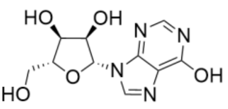
A1.2. NUCLEOSIDES



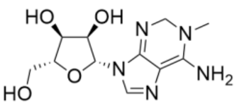
N1 Cytidine



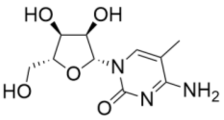
N2 Guanosine



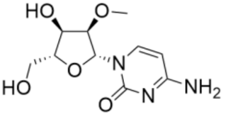
N3 Inosine



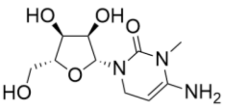
N4 1-methyladenosine



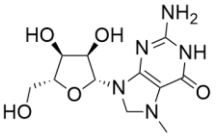
N5 5-methylcytidine



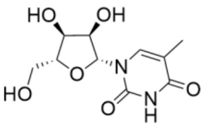
N6 2'-O-methylcytidine



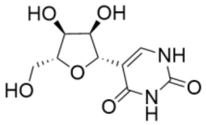
N7 3-methylcytidine



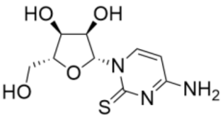
N8 7-methylguanosine



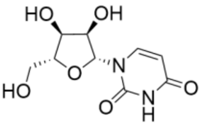
N9 5-methyluridine



N10 β-pseudouridine



N11 2-thiocytidine

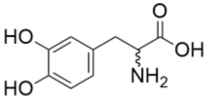


N12 Uridine

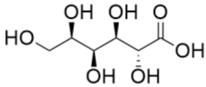
ID	Monoisotopic mass /g·mol^{-1}	Experimental retention times and peak widths /min					
		Method 1	Method 2	Method 3a	Method 3b	Method 3c	Method 4
N1	243.0855	8.70, 0.77	16.71, 0.58	2.65, 0.70	2.47, 0.13	2.53, 0.16	1.82, 0.28
N2	283.0917	8.99, 1.20	-	2.59, 1.60	2.37, 0.13	2.42, 0.22	1.64, 0.30
N3	268.0808	7.90, 0.95	-	2.12, 0.19	2.16, 0.19	2.09, 0.06	1.39, 0.25
N4	281.1124	6.14, 0.80	-	2.22, 1.35	2.12, 0.19	2.06, 0.15	1.26, 0.25
N5	257.1012	6.88[d], 0.70	25.52[e], 0.44	2.70[f], 0.54	2.39, 0.13, 2.53, 0.10 or 3.85, 0.13	2.47, 0.21, 2.59, 0.16 or 4.08, 0.16	1.61, 0.26, 1.95, 0.30 or 3.62, 0.51
N6	257.1012	6.88[d], 0.70	25.52[e], 0.44	2.70[f], 0.54	2.39, 0.13, 2.53, 0.10 or 3.85, 0.13	2.47, 0.21, 2.59, 0.16 or 4.08, 0.16	1.61, 0.26, 1.95, 0.30 or 3.62, 0.51
N7	257.1012	6.88[d], 0.70	25.52[e], 0.44	2.70[f], 0.54	2.39, 0.13, 2.53, 0.10 or 3.85, 0.13	2.47, 0.21, 2.59, 0.16 or 4.08, 0.16	1.61, 0.26, 1.95, 0.30 or 3.62, 0.51
N8	298.1151	-	-	2.23, 0.12	-	-	1.30, 0.33
N9	258.0852	6.20, 0.90	-	1.30, 0.60	1.36, 0.35	1.32, 0.44	0.99, 0.27
N10	244.0695	6.91, 0.82 or 9.24, 1.05	14.32, 0.73 or 16.89, 0.52	1.42, 0.80 or 2.00, 0.48	1.47, 0.30 or 1.87, 0.29	1.43, 0.38 or 1.85, 0.43	1.01, 0.23 or 1.25, 0.23
N11	259.0627	6.94, 0.75	19.53, 0.14	1.40, 0.71	1.61, 0.28	1.63, 0.17	1.12, 0.21
N12	244.0695	6.91, 0.82 or 9.24, 1.05	14.32, 0.73 or 16.89, 0.52	1.42, 0.80 or 2.00, 0.48	1.47, 0.30 or 1.87, 0.29	1.43, 0.38 or 1.85, 0.43	1.01, 0.23 or 1.25, 0.23

Table A2: Nucleosides' experimental retention times and peak widths in all assayed chromatographic conditions. Shaded analytes are also present in the "other metabolites" solution. Isobaric molecules are indicated with the same superscript letter.

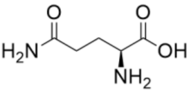
A1.3. Other Metabolites



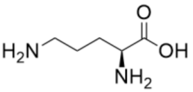
O1 3,4-dihydroxy-D,L-phenylalanine



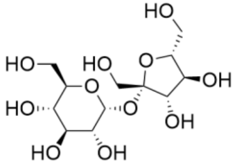
O2 D-gulonic acid



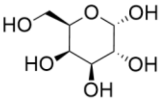
O3 L-glutamine



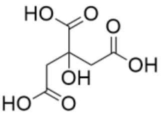
O4 L-ornithine



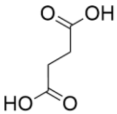
O7 Sucrose



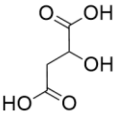
O8 D-galactose



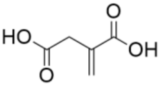
O10 Citric acid



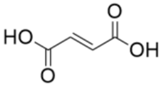
O11 Succinic acid



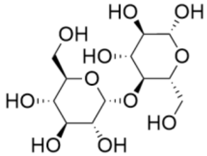
O12 Malic acid



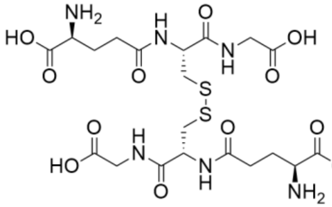
O13 Itaconic acid



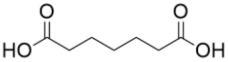
O14 Fumaric acid



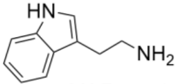
O9 D-maltose



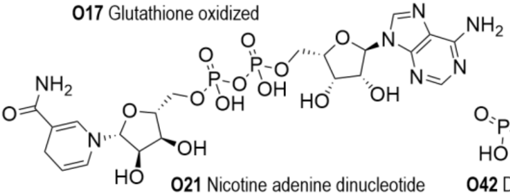
O17 Glutathione oxidized



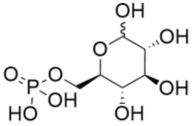
O15 Pimelic acid



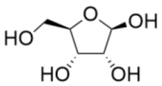
O20 Tryptamine



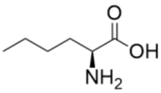
O21 Nicotine adenine dinucleotide



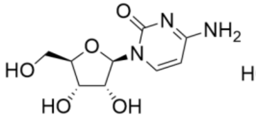
O42 D-glucose 6-phosphate



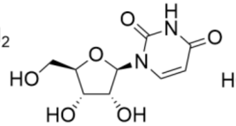
O23 D-ribose



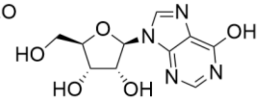
O43 L-norleucine



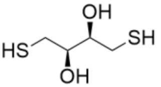
O24 Cytidine



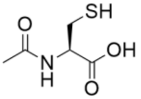
O25 Uridine



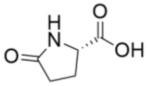
O26 Inosine



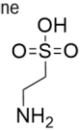
O28 Dithiothreitol



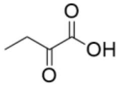
O29 N-acetyl-cysteine



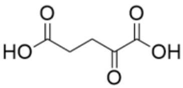
O31 L-pyroglutamic acid



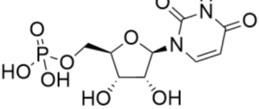
O34 Taurine



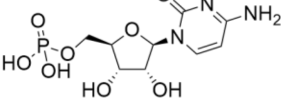
O38 2-ketobutyric acid



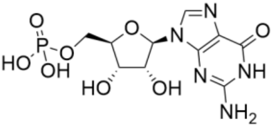
O40 α-ketoglutaric acid



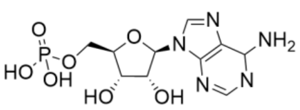
O45 Uridine 5'-monophosphate



O36 Cytidine 5'-monophosphate



O35 Guanosine 5'-monophosphate



O22 Adenosine 5'-monophosphate

ID	Monoisotopic mass /g·mol^{-1}	Experimental retention times and peak widths /min					
		Method 1	Method 2	Method 3a	Method 3b	Method 3c	Method 4
O1	197.0688	12.87, 1.10	-	2.55, 0.31	3.24, 0.86	-	-
O2	196.0583	11.80, 1.51	-	256, 0.38	2.90, 0.62	-	3.16, 0.27
O3	146.0691	13.40, 0.40	-	3.71, 0.33	3.52, 0.26	3.82, 0.18	6.88, 0.14
O4	132.0899	8.99, 0.88	-	5.32, 0.80	5.50, 0.35	-	-
O7	342.1162	13.03, 0.30 or 13.04, 0.35	20.15, 0.39 or 20.84, 0.58	2.83^{g}, 0.40	2.65^{i}, 0.11	-	2.72, 0.17 or 3.09, 0.33
O8	180.0634	13.18, 0.75	-	3.02, 0.34	2.10, 0.18	-	-
O9	342.1162	13.03, 0.30 or 13.04, 0.35	20.15, 0.39 or 20.84, 0.58	2.83^{g}, 0.40	2.65^{i}, 0.11	-	2.72, 0.17 or 3.09, 0.33
O10	192.0270	-	-	3.50, 0.61	-	-	-
O11	118.0266	7.29, 0.60	-	1.51, 0.93	1.76, 0.26	2.24, 0.06	0.92, 0.16
O12	134.0215	12.16, 1.10	-	2.54, 0.42	3.20, 0.77	-	1.80, 0.46
O13	130.0266	11.85, 0.90	6.63, 0.50	1.07, 0.29	1.04, 0.22	-	0.89, 0.14
O14	116.0110	12.00, 1.30	-	2.53, 0.59	-	-	-
O15	160.0736	7.20, 0.53	-	2.90, 0.37	2.66, 0.12	3.91, 0.12	0.80, 0.06
O17	612.1520	-	-	3.64, 0.39	3.60, 0.40	-	-
O20	160.1000	8.90, 0.41	16.90, 0.32	3.80, 0.20	3.58, 0.10	3.88, 0.15	-
O21	745.0838	-	-	-	-	-	-
O22	347.0631	11.79, 1.40	-	-	-	-	7.30, 0.30
O23	150.0528	-	-	-	-	-	1.17, 0.18
O24	243.0855	8.91, 0.55	16.74, 0.30	2.78, 0.36	2.47, 0.13	-	1.87, 0.20
O25	244.0695	7.01, 0.73	14.35, 0.48	1.41, 0.55	1.48, 0.30	-	1.01, 0.15
O26	268.0808	7.97, 0.62	-	2.06, 1.11	2.13, 0.16	2.16, 0.19	1.42, 0.14
O28	154.0122	-	-	1.01, 0.06	-	-	0.76, 0.07
O29	163.0303	5.21, 0.40	-	1.82, 1.45	2.22, 0.19	-	1.54, 0.28
O31	129.0426	7.32, 0.55	16.45, 0.36	2.96, 0.36	-	-	-
O34	125.0147	10.79, 1.40	-	2.80, 0.44	2.59, 0.11	-	2.41, 0.20
O35	363.0580	12.9, 0.55	19.66, 0.29	3.38, 0.49	3.18, 0.23	2.58, 0.15	7.26, 0.27
O36	323.0519	12.93, 1.10	-	3.45, 0.41	3.31, 0.28	-	7.56, 0.30
O38	102.0317	-	-	1.80, 2.10	1.07, 0.11	-	-
O40	146.0215	7.41, 0.32	16.96, 0.25	1.20, 0.60	2.34, 0.16	-	1.70, 0.17
O42	260.0297	13.20, 0.27	20.22, 0.24	3.35, 0.35	3.21, 0.26	-	7.28, 0.24
O43	131.0946	9.06, 0.50	-	-	-	-	-
O45	324.0359	12.28, 0.90	18.98, 0.35	2.99, 1.25	3.00, 0.25	-	6.57, 0.62

Table A3: Other metabolites experimental retention times and peak widths in all assayed chromatographic conditions. Green- and purple-shaded analytes are also present in the “amino acids” and “nucleosides” solutions, respectively. Isobaric molecules are indicated with the same superscript letter.

APPENDIX 2: COMPARISON OF THE METHODS

		Method 1	Method 2	Method 3a	Method 3b	Method 3c	Method 4
LC type		HPLC		UPLC			
Stationary phase		Amide-functionalized silica		Ethylene bridged hybrid			
Column temperature /ºC		Room temperature		30			
Flow rate /mL·min^{-1}		0.15		0.30			0.40
Mobile phase	**Organic solvent (A)**	ACN					90:10 ACN/H_2O, 0.2 % HCOOH and 10 mM NH_4COOH
	Aqueous solvent (B)	H_2O, 5 mM NH_4Ac, pH 5.5 (HAc)					50:50 ACN/H_2O, 0.2 % HCOOH and 10 mM NH_4COOH
Total run time /min		20	40	10	8	9	15
Metabolites' concentration /ppm		4-50		20			
Injection volume /µL		5		10	5		

Table A4: Comparison of the assayed chromatographic methods.

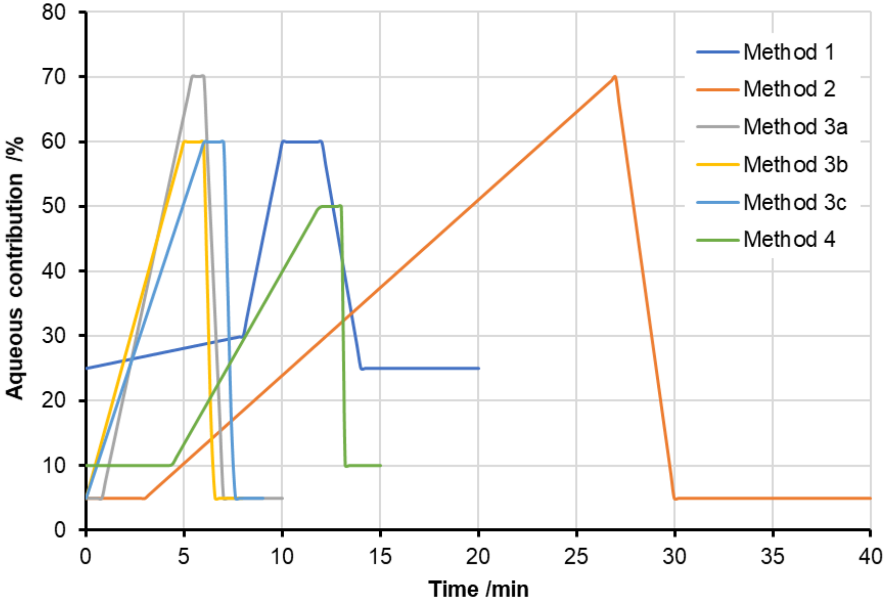


Figure A1: Percentage of water in the mobile phases of each method's gradient elution.

APPENDIX 3: RETENTION TIME PREDICTIONS

A3.1. MODEL CALCULATION AND VALIDATION

		Method 1						Method 3b					
		BRNN		RF		XGBoost		BRNN		RF		XGBoost	
Permutation	**Statistic**	**calc**	**val**	**calc**	**val**	**calc**	**val**	**calc**	**val**	**calc**	**val**	**calc**	**val**
1	**RMSE**	0.07	2.48	0.66	2.00	1.28	2.05	1.01	0.71	0.30	0.77	0.47	1.30
	R^2	1.00	0.21	0.97	0.65	0.84	0.67	0.33	0.80	0.96	0.81	0.85	0.27
	MAE	0.05	1.85	0.51	1.52	0.98	1.76	0.75	0.53	0.23	0.64	0.35	1.00
	95%±min	0.12	3.61	1.09	1.38	2.09	1.80	1.65	0.93	0.49	0.98	0.79	1.84
2	**RMSE**	0.43	1.93	0.76	1.85	1.33	1.93	0.53	0.96	0.20	1.00	0.33	1.17
	R^2	0.98	0.66	0.96	0.79	0.81	0.77	0.77	0.92	0.96	0.46	0.95	0.43
	MAE	0.32	1.73	0.63	1.23	1.11	1.20	0.40	0.73	0.15	0.91	0.23	0.97
	95%±min	0.71	3.79	1.28	2.30	2.21	2.18	0.87	1.94	0.33	1.89	0.55	2.35
3	**RMSE**	0.71	1.75	1.25	1.41	1.32	1.42	0.54	0.24	0.24	0.38	0.37	0.46
	R^2	0.96	0.54	0.92	0.67	0.84	0.67	0.83	0.80	0.96	0.45	0.96	0.75
	MAE	0.56	1.46	0.99	1.08	1.05	1.12	0.43	0.21	0.18	0.35	0.30	0.41
	95%±min	1.15	3.22	2.22	2.46	2.21	2.44	0.90	0.48	0.41	0.73	0.61	0.93
4	**RMSE**	0.40	2.39	0.69	2.00	1.19	2.00	0.48	0.84	0.26	1.04	0.30	1.12
	R^2	0.98	0.43	0.95	0.60	0.85	0.61	0.82	0.89	0.93	0.87	0.96	0.81
	MAE	0.26	2.07	0.51	1.77	0.90	1.70	0.37	0.75	0.17	0.88	0.23	0.96
	95%±min	0.68	3.91	1.19	3.49	1.96	3.64	0.79	1.37	0.45	1.51	0.50	1.73
5	**RMSE**	0.61	1.11	0.70	0.68	1.44	0.83	0.46	0.55	0.23	0.68	0.28	0.69
	R^2	0.96	0.81	0.96	0.96	0.81	0.96	0.87	0.68	0.96	0.65	0.96	0.56
	MAE	0.45	0.90	0.57	0.60	1.16	0.75	0.35	0.47	0.17	0.52	0.23	0.53
	95%±min	1.02	1.87	1.16	0.37	2.40	0.16	0.77	0.58	0.38	0.51	0.44	0.60
6	**RMSE**	0.60	1.92	0.79	1.56	1.28	1.85	0.47	0.57	0.04	0.65	0.27	0.85
	R^2	0.96	0.77	0.95	0.83	0.82	0.71	0.88	0.64	1.00	0.54	0.96	0.54
	MAE	0.43	1.76	0.61	1.28	0.95	1.66	0.34	0.51	0.03	0.54	0.21	0.67
	95%±min	1.01	3.73	1.33	3.13	2.12	3.48	0.78	1.02	0.06	1.22	0.44	0.86
7	**RMSE**	0.45	1.67	0.89	1.19	1.34	1.25	0.55	0.82	0.25	0.82	0.36	1.01
	R^2	0.98	0.57	0.96	0.73	0.83	0.70	0.76	0.92	0.94	0.67	0.95	0.63
	MAE	0.34	1.19	0.69	0.97	1.05	0.89	0.40	0.63	0.19	0.63	0.26	0.87
	95%±min	0.75	2.68	1.45	2.14	2.22	2.15	0.91	1.55	0.44	1.64	0.61	2.05
8	**RMSE**	0.07	2.17	0.75	1.70	1.34	1.97	0.52	0.70	0.26	0.73	0.27	0.74
	R^2	1.00	0.54	0.97	0.72	0.82	0.59	0.79	0.89	0.94	0.63	0.96	0.80
	MAE	0.05	1.84	0.59	1.40	1.01	1.67	0.39	0.56	0.19	0.55	0.21	0.56
	95%±min	0.12	4.27	1.24	2.96	2.19	3.39	0.87	1.38	0.44	1.46	0.43	1.48
9	**RMSE**	0.11	2.40	0.79	1.63	1.27	1.77	0.49	0.57	0.25	0.61	0.29	0.86
	R^2	1.00	0.82	0.96	0.75	0.85	0.63	0.87	0.43	0.95	0.32	0.97	0.09
	MAE	0.07	1.99	0.62	1.24	1.02	1.38	0.39	0.36	0.19	0.38	0.23	0.65
	95%±min	0.17	4.48	1.36	3.22	2.08	3.58	0.80	0.81	0.43	0.85	0.47	1.37
10	**RMSE**	0.09	2.56	0.63	2.00	1.16	2.10	0.50	0.62	0.26	0.66	0.28	0.70
	R^2	1.00	0.44	0.97	0.42	0.88	0.34	0.83	0.76	0.94	0.76	0.96	0.76
	MAE	0.06	1.97	0.46	1.55	0.91	1.57	0.38	0.42	0.19	0.49	0.21	0.50
	95%±min	0.15	5.16	1.05	3.86	1.94	3.88	0.83	1.05	0.44	1.13	0.47	1.22

Table A5: Calculated statistics for the generation of each model's ten permutations. Shaded values correspond to the best permutations, those selected to perform RT predictions.

A3.2. Most Relevant Molecular Descriptors

	BRNN		RF		XGBoost	
	Molecular descriptor	Relative importance /%	Molecular descriptor	Relative importance /%	Molecular descriptor	Relative importance /%
Method 1 (permutation 5)	XLogP	100.00	XLogP	100.00	XLogP	100.00
	nHBAcc	59.53	VP.5	33.02	nBase	10.39
	tpsaEfficiency	58.39	SP.5	25.73	tpsaEfficiency	8.31
	khs.sNH2	57.27	nHBDon	21.93	SPC.6	6.51
	SC.5	50.49	tpsaEfficiency	21.28	Kier3	5.57
	MDEC.22	47.08	ALogp2	20.41	ALogP	4.42
	nRotB	45.36	VC.3	20.08	nAcid	3.82
	BCUTc.1h	45.05	Kier3	16.64	khs.sNH2	2.75
	SC.3	42.50	MDEC.13	16.59	BCUTp.1l	2.58
	Kier2	41.17	ALogP	14.67	BCUTc.1h	1.83
	VC.5	39.21	Zagreb	14.60	SPC.5	1.78
	SPC.4	37.94	khs.sNH2	13.89	VPC.5	1.73
	TopoPSA	37.82	nAcid	13.31	ATSc4	1.67
	ATSm3	37.82	VCH.6	13.23	Fsp3	1.53
	Kier3	36.77	VC.5	13.22	SCH.7	1.47
	VCH.6	36.66	MW	12.85	nHBAcc	1.40
	VC.3	34.79	Fsp3	12.77	BCUTw.1h	1.36
	MDEC.33	34.70	MDEN.11	12.33	nRotB	1.35
	ATSc1	33.75	TopoPSA	12.32	ALogp2	1.35
	ALogP	32.37	ATSm5	12.19	BCUTw.1l	1.15
Method 3b (permutation 3)	MDEC.22	100.00	nBase	100.00	XLogP	100.00
	khs.sNH2	90.63	XLogP	92.43	nBase	50.24
	XLogP	86.19	khs.ssCH2	85.02	MDEC.22	36.06
	nBase	81.66	khs.sNH2	83.36	BCUTw.1h	18.64
	BCUTw.1h	79.69	MDEN.11	78.22	WTPT.5	14.63
	nRotB	69.74	BCUTw.1h	77.56	MDEO.11	13.29
	SP.6	68.39	VPC.5	72.81	Kier3	13.13
	BCUTp.1l	67.59	nRotB	71.81	nRotB	8.04
	VP.6	67.14	WTPT.5	70.35	BCUTw.1l	7.55
	MDEO.11	66.42	fragC	68.33	khs.sNH2	6.97
	ATSp4	64.77	nHBDon	65.75	khs.ssCH2	5.41
	VP.2	64.38	khs.sssCH	65.20	WTPT.4	4.83
	VP.3	63.57	MDEC.22	65.05	khs.dO	4.65
	VP.5	59.40	Kier3	64.53	nAtomP	3.97
	WTPT.4	58.62	VPC.6	64.29	Kier2	3.72
	VP.1	58.25	VP.4	62.65	tpsaEfficiency	2.25
	VP.4	58.20	SP.7	62.24	ALogp2	1.99
	Fsp3	56.28	VP.6	61.60	ATSc5	1.96
	HybRatio	56.28	VCH.7	61.47	Fsp3	1.59
	khs.ssCH2	56.01	ATSm4	61.14	VP.1	1.55

Table A6: 20 most important molecular descriptors for the generation of the BRNN, RF and XGBoost models from Methods 1 and 3b chromatographic data. Common descriptors are indicated with the same colors.

A3.3. Improved Metabolite Detection and Annotation

ID	MM /g·mol^{-1}	Method 1					Method 3b				
		Predicted RT /min				Exp. RT /min	Predicted RT /min				Exp. RT /min
		BRNN5	RF5	XGBoost5	Interval		BRNN3	RF3	XGBoost3	Interval	
A1	89.0477	9.82	10.49	9.86	9.8-10.5	-	3.09	2.30	3.13	2.3-3.1	3.06^{m}
A2	89.0477	11.69	12.63	12.74	11.7-12.7	12.24^{m}	3.13	2.46	3.32	2.5-3.3	3.34^{m}
A32	89.0477	8.27	8.12	8.62	7.4-10.2	7.84^{m}	2.65	2.19	2.86	2.2-2.9	2.35^{m}
A4	103.0633	10.74	12.37	12.18	10.7-12.4	12.40^{m}	3.13	2.59	3.04	2.6-3.1	3.40^{m}
A5	103.0633	9.27	9.49	9.70	9.3-9.7	-	3.73	2.72	3.35	2.7-3.7	3.70^{m}
A6	103.0633	10.11	10.26	9.19	9.2-10.1	10.07^{m}	2.85	2.45	2.74	2.5-2.9	2.75^{m}
A3	161.0688	AA&UMG					3.55	3.74	3.67	3.5-3.7	3.61^{p}
A8	240.1222	11.32	13.38	13.20	11.3-13.4	11.84^{m}	AA&UMG				
A9	174.1117	16.96	14.27	14.15	14.2-17.0	16.70^{m}	AA&UMG				
A12	175.0957	AA&UMG					4.51	4.50	4.33	4.3-4.5	3.90^{p}
A15	121.0197	AA&UMG					3.83	3.63	3.68	3.6-3.8	-
A16	61.0528	8.55	10.49	9.84	8.6-10.5	-	3.94	2.43	3.13	2.4-3.9	-
A18	75.0320	11.58	13.05	13.46	11.6-13.5	-	3.18	2.47	3.61	2.5-3.6	-
A20	268.0551	AA&UMG					4.45	3.74	3.59	3.6-4.5	4.08^{p}
A22	131.0582	8.27	8.12	8.62	8.2-8.6	7.84^{p}	AA&UMG				
A23	131.0946	9.94	9.01	8.73	8.7-9.9	9.05^{p}	2.81	3.20	2.86	2.8-3.2	2.99^{p}
A24	131.0946	9.78	9.13	9.49	9.1-9.8	9.76^{p}	2.88	3.18	3.04	2.9-3.2	
A25	146.1055	12.94	13.56	12.91	12.9-13.6	13.01^{m}	AA&UMG				
A27	169.0851	11.80	12.31	12.23	11.8-12.3	12.20^{m}	4.01	3.53	3.76	3.7-4.0	4.17^{p}
A28	169.0851	11.88	12.72	12.60	11.9-12.7	12.40^{m}	3.85	3.52	4.01	3.5-4.0	4.70^{p}
A31	115.0633	6.81	8.56	8.87	6.8-8.9	-	AA&UMG				
A33	105.0426	13.25	12.91	13.33	12.9-13.3	12.75^{m}	3.70	2.54	3.52	2.5-3.7	3.52^{p}
N5	257.1012	7.94	7.36	8.33	7.4-8.3	6.88^{p}	1.86	2.00	2.21	1.8-2.2	2.59^{p}
N6	257.1012	6.92	7.18	8.19	6.9-8.2	8.27^{p}	2.31	2.16	2.32	2.2-2.3	4.08^{p}
N7	257.1012	6.32	6.98	8.10	6.3-8.1	6.29^{m}	1.75	2.01	2.12	1.7-2.1	2.47^{p}
N8	298.1151	8.02	7.02	8.70	7.0-7.8	-	1.66	2.34	2.58	1.7-2.6	2.00^{p}
N10	244.0565	10.46	11.74	10.98	10.5-11.7	9.24^{p}	1.88	2.31	2.21	1.9-2.3	1.85^{p}
N12/O25	244.0695	7.32	8.43	8.88	7.3-8.9	6.91^{p}	1.66	2.37	2.28	1.7-2.4	1.44^{p}
O7	342.1162	9.21	12.76	11.80	9.2-12.8	13.03^{p}	2.55	2.77	2.95	2.6-3.0	2.84^{p}
O8	180.0633	7.42	8.2	8.38	7.4-8.4	-	AA&UMG				
O9	342.1162	8.75	12.64	11.86	8.8-12.6	13.40^{p}	2.13	2.7	3.05	2.1-3.1	2.68^{p}
O10	192.0270	14.02	9.23	10.1	9.2-14.0	-	1.74	2.99	2.41	1.7-3.0	2.87^{p}
O14	116.0110	AA&UMG					1.48	1.84	2.20	1.5-2.2	2.16^{p}
O17	612.1520	13.81	13.46	13.47	13.5-15.8	13.43^{p}	AA&UMG				
O21	745.0911	15.31	12.39	13.12	12.4-15.3	13.34^{m}	AA&UMG				
O22	347.0631	AA&UMG					3.82	2.86	3.16	2.9-3.8	3.46^{p}
O23	150.0528	6.72	8.27	8.29	6.7-8.3	-	2.27	2.93	2.05	2.1-2.9	-
O28	154.0122	6.18	9.25	8.74	6.2-9.3	8.87^{m}	AA&UMG				
O31	129.0426	AA&UMG					2.35	2.95	2.49	2.5-3.0	2.84^{p}
O38	102.0317	6.8	10.53	8.69	6.8-10.5	-	AA&UMG				
O43	131.0946	AA&UMG					3.55	3.81	3.35	3.4-3.8	3.86^{p}

Table A7: Predicted retention times, examined intervals and newly found m/z and annotated peaks (MM: monoisotopic mass, Exp. RT: experimental retention time, AA&UMG: metabolites already assigned and used for the models' generation, m: detected mass, p: visible peak).