Journal of Chromatography A, 1542 (2018) 82–88

Contents lists available at ScienceDirect

Journal of Chromatography A

journal homepage: www.elsevier.com/locate/chroma

Prediction of collision cross section and retention time for broad scope screening in gradient reversed-phase liquid chromatography-ion mobility-high resolution accurate mass spectrometry

Christian Brinch Mollerup[a,*], Marie Mardal[a], Petur Weihe Dalsgaard[a], Kristian Linnet[a], Leon Patrick Barron[b]

[a] Section of Forensic Chemistry, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Frederik V's vej 11, 3rd floor, Copenhagen DK-2100, Denmark
[b] Analytical & Environmental Sciences Division, Faculty of Life Sciences & Medicine, King's College London, Franklin-Wilkins Building, 150 Stamford Street, London SE1 9NH, United Kingdom

ARTICLE INFO

Article history:
Received 15 December 2017
Received in revised form 6 February 2018
Accepted 14 February 2018
Available online 15 February 2018

Keywords:
Collision cross section prediction
Retention time prediction
Artificial neural networks

ABSTRACT

Exact mass, retention time (RT), and collision cross section (CCS) are used as identification parameters in liquid chromatography coupled to ion mobility high resolution accurate mass spectrometry (LC-IM-HRMS). Targeted screening analyses are now more flexible and can be expanded for suspect and non-targeted screening. These allow for tentative identification of new compounds, and *in-silico* predicted reference values are used for improving confidence and filtering false-positive identifications. In this work, predictions of both RT and CCS values are performed with machine learning using artificial neural networks (ANNs). Prediction was based on molecular descriptors, 827 RTs, and 357 CCS values from pharmaceuticals, drugs of abuse, and their metabolites. ANN models for the prediction of RT or CCS separately were examined, and the potential to predict both from a single model was investigated for the first time. The optimized combined RT-CCS model was a four-layered multi-layer perceptron ANN, and the 95th prediction error percentiles were within 2 min RT error and 5% relative CCS error for the external validation set (n = 36) and the full RT-CCS dataset (n = 357). 88.6% (n = 733) of predicted RTs were within 2 min error for the full dataset. Overall, when using 2 min RT error and 5% relative CCS error, 91.9% (n = 328) of compounds were retained, while 99.4% (n = 355) were retained when using at least one of these thresholds. This combined prediction approach can therefore be useful for rapid suspect/non-targeted screening involving HRMS, and will support current workflows.

1. Introduction

Liquid chromatography coupled high resolution accurate mass spectrometry (LC-HRMS) has enabled comprehensive toxicological screening of large numbers of trace contaminants in complex matrices such as biological samples and environmental matrices [1–6]. The addition of ion mobility spectrometry (IMS) has recently represented a significant increase in capability and allows for separation of ions in the gas-phase based on their mobility differences in an applied electric field [7,8]. Ions are then measured by their drift times through a tube containing a buffer gas. While drift times are system dependent, the average collision cross sections (CCS) between the ion and buffer gas can be derived when using constant operating procedures. The CCS of an ion is correlated to its size, shape, and charge. After calibration, the drift times observed from a travelling-wave IMS (TW-IMS) system can be used to determine CCS values [9]. CCS from TW-IMS have been shown to be matrix and system independent [10,11], and the use of LC-IMS-HRMS have been used to reduce the number of false-positive identifications and can replace other screening metrics for confirmatory analysis as a result (*e.g.* isotopic pattern match and fragment ions) [12]. The use of RT and CCS for confirmatory analyses means there also exists a lesser need for data-dependent fragmentation as the full-scan HRMS fragmentation can be filtered both on RT and drift time alignment [7]. This can then be applied to targeted, suspect, and non-targeted screening as required using the same dataset.

A common challenge, particularly in forensic screening, is keeping methods updated with relevant compounds. More than two

* Corresponding author.
E-mail address: christian.brinch.mollerup@sund.ku.dk (C.B. Mollerup).

https://doi.org/10.1016/j.chroma.2018.02.025
0021-9673/

new psychoactive substances enter the American and European drug market every week, on average [13,14]. Also, with the increase of long-distance travel for vacations and work, local populations can be exposed to pollutants and drugs not prescribed in their home countries. Suspect and non-targeted screening approaches have been utilized for identification of compounds before acquisition of reference standards [4,15–17]. For this purpose, *in-silico* fragmentation matching [18,19] and prediction of retention time (RT) have been shown to reduce the list of potential compounds [1,20]. *In-silico* prediction of CCS and IMS drift times have utilized molecular modelling techniques [21–23]; however, models based on molecular descriptors have shown similar results while drastically reducing computing time [24–27], which corresponds to findings for prediction of the reduced ion mobility constants [28,29].

The aim of this work was to predict both RT and CCS with the use of artificial neural networks (ANNs), a machine learning technique that has been demonstrated for predicting analytical reference values, and has only very recently been utilized for prediction of either RT [1,30] or CCS [24] for use in screening. However, combination of these tools to understand their added value for preliminary suspect identifications has not yet been performed. A previously developed ANN model for RT prediction was trained and validated herein on a new, significantly larger dataset gathered under different LC conditions and in a different laboratory; ANN and linear regression models for prediction of CCS were trained and validated, and finally a combined model for prediction of both RT and CCS simultaneously was critically evaluated. This novel approach to *in silico* prediction of both RT and CCS alongside the use of HRMS data will markedly increase the speed and confidence in tentative identifications of potentially large numbers of new compounds.

2. Materials and methods

2.1. Chemicals

Reference standards of pharmaceuticals, drugs of abuse, and their metabolites were purchased from Lipomed GmbH (Bad Säckingen, Germany), Cerilliant (Round Rock, TX, USA), Toronto Research Chemicals (Toronto, Canada), and SelleckChem (Houston, TX, USA). All reference standards were of ≥98% purity. Methanol, water, acetonitrile, propanol, and formic acid (LC–MS grade) were obtained from Fisher Scientific (Loughborough, UK). Leucine enkephalin was purchased from Sigma-Aldrich (Copenhagen, Denmark).

2.2. Instrumentation

Analyses were performed on two separate systems, an ultra-high performance liquid chromatography-time-of-flight mass spectrometer (UHPLC-TOF; System 1) and a UHPLC-TW-IMS-TOF (System 2). RT were obtained on System 1 with an ACQUITY UPLC I-Class coupled with a Xevo G2-S QTOF (Waters MS Technologies, Manchester, United Kingdom), and CCS values were obtained on System 2: an ACQUITY UPLC H-Class coupled with a VION IMS QTOF (Waters MS Technologies, Manchester, United Kingdom). LC separations on both systems were achieved using an Acquity UPLC HSS C_{18} column (150 mm × 2.1 mm, 1.8 μm), which was maintained at a constant temperature of 50 °C and a flow rate of 0.4 mL/min. Mobile Phase A consisted of 5 mM aqueous ammonium formate buffer adjusted to pH 3 with formic acid, and Mobile Phase B consisted of acetonitrile with 0.1% v/v formic acid. The gradient was 0 min to 0.5 min: 13% (B); from 0.5 min to 10 min: 13% to 50% (B); from 10 min to 10.75 min: 50% to 95% (B); from 10.75 min to 12.25 min: 95% (B); and from 12.25 min to 12.5 min: 95% to 13% (B); from 12.5 min to 15 min: 13% (B). The total run time was 15 min, and the injection volume was 3 μL. Ion mobility (System 2) was calibrated with a Major Mix IMS/Tof Calibration Kit from Waters, drift times were measured, and CCS values were calculated by the UNIFI software (Waters MS Technologies, Manchester, United Kingdom). Nitrogen (N_2) was used as drift gas in the TW-IMS of System 2.

With respect to mass spectrometry, both systems were used in positive electrospray ionization (Z-spray) mode with the following settings: nebulization gas 1000 L/h (System 1) and 800 L/h (System 2), with a desolvation temperature of 400 °C; cone gas flow 10 L/h (System 1) and 20 L/h (System 2); source temperature 150 °C; capillary voltage 800 V; cone voltage 25 V; and argon as the collision gas. The low collision energy was set at 4 eV, and the high collision energy was ramped from 10 to 40 eV. The acquisition time was the entire run, with a scan time of 0.200 s. The minimum mass-to-charge (m/z) was 50 and the maximum was 950 (System 1) or 1000 (System 2). Mass calibration of System 1 was performed with 5 mM sodium formate solution in propanol: water (90:10, v/v), while System 2 was mass calibrated with the Major Mix IMS/Tof Calibration Kit from Waters. Lock mass was used with leucine enkephalin as a reference mass at m/z 556.2766 on both systems.

2.3. Reference values

In total, RTs for 869 compounds were determined from reference standards (Dataset I). Of these, the CCS of the proton adduct was determined for 364 compounds (Dataset II). For both datasets, compounds identified as multiple LC peaks were excluded. RTs were recorded on both systems, however, only RTs from System 1 was used for prediction. The differences in dataset sizes were primarily due to reference standards only being analyzed on System 1. Other factors were no observed protonation adducts, either due to high affinity for metal adducts or heavy in-source fragmentation.

2.4. Molecular descriptor generation

A total of 869 unique simplified molecular-input line-entry system (SMILES) strings were generated with ChemScript v16.0 from PerkinElmer (Waltham, MA, USA) from an in-house database of mol-files. Each SMILES string corresponded to a single compound and was used to generate a total of 105 molecular descriptors with Parameter Client freeware [31,32]. The selected descriptors were constitutional descriptors, functional group counts, and molecular properties. Additional descriptors were generated for each compound: thirteen descriptors from ACD/Percepta (ACD/Labs, Toronto, Canada) and six descriptors from ChemScript v16.0 from PerkinElmer (Waltham, MA, USA). The full list of SMILES and corresponding descriptor values are available in Table A.1. Compounds for which the descriptor generation failed were excluded from the ANN modelling.

2.5. Descriptor selection and ANN optimization

ANN modelling was performed using Trajan Neural Networks v6.0. Prior to any evaluation, Dataset I & II were split into optimization and external validation sets with compounds chosen at random, in proportions 80:20 and 90:10, respectively. RT values for compounds exclusive to Dataset I were added as an external validation set in Models RT2 & RT-CCS (only regarding the RT prediction). The external validation set were used to reduce the risk of overfitting to the optimization set.

In total, four ANNs were trained and optimized. Single-output models for RT or CCS included Models RT1 & RT2, which were used with Dataset I & II, respectively, to predict RT, and Model CCS, which used Dataset II for predicting CCS. Model RT-CCS was a two-output model for predicting both CCS and RT simultaneously