Prediction of Retention Time and Collision Cross Section (CCS_{H+}, CCS_{H-} and CCS_{Na+}) of emerging contaminants using Multiple Adaptive Regression Splines

Alberto Celma
University Jaume I https://orcid.org/0000-0001-9763-8737
Richard Bade
Queensland Alliance for Environmental Health Sciences (QAEHS), The University of Queensland
Juan V. Sancho
Research Institute for Pesticides and Water, University Jaume I
Félix Hernández
Research Institute for Pesticides and Water, University Jaume I
Melissa Humpries
School of Mathematical Sciences, University of Adelaide
Lubertus Bijslma (bijlsma@uji.es)
Research Institute for Pesticides and Water, University Jaume I

Article

Keywords: Multivariate Adaptive Regression Splines (MARS), collision cross section prediction, retention time prediction, liquid chromatography, ion mobility separation, high resolution mass spectrometry

Posted Date: January 13th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1249834/v1

Version of Record: A version of this preprint was published at Journal of Chemical Information and Modeling on October 24th, 2022. See the published version at https://doi.org/10.1021/acs.jcim.2c00847.

Abstract

Ultra-high performance liquid chromatography coupled to ion mobility separation and high-resolution mass spectrometry instruments have proven very valuable for screening of emerging contaminants in the aquatic environment. However, when applying suspect or non-target approaches (*i.e.* when no reference standards are available) there is no information on retention time (RT) and collision cross section (CCS) values to facilitate identification. *In-silico* prediction tools of RT and CCS can therefore be of great utility to decrease the number of candidates to investigate. In this work, Multiple Adaptive Regression Splines (MARS) was evaluated for the prediction of both RT and CCS. MARS prediction models were developed and validated using a database of 477 protonated molecules, 169 deprotonated molecules and 249 sodium adducts. Multivariate and univariate models were evaluated showing a better fit for univariate models to the empirical data. The RT model (R^2=0.855) showed a deviation between predicted and empirical data of ± 2.32 min (95% confidence intervals). The deviation observed for CCS data of protonated molecules using CCS_H model (R^2=0.966) was ± 4.05% with 95% confidence intervals. The CCS_H model was also tested for the prediction of deprotonated molecules resulting in deviations below ± 5.86% for the 95% of the cases. Finally, a third model was developed for sodium adducts (CCS_{Na}, R^2=0.954) with deviation below ± 5.25% for the 95% of the cases. The developed models have been incorporated in an open access and user-friendly online platform which represents a great advantage for third-party research laboratories for predicting both RT and CCS data.

1. Introduction

In the last decade, considerable effort has been devoted to enhance the performance of high resolution mass spectrometry (HRMS) suspect screening (SS) and non-target screening (NTS) strategies [1–3]. The instrumental improvements of HRMS instruments has required the development of more sophisticated algorithms to be able to handle the large amount of data generated [3, 4]. Therefore, the development of open-access scripts for data processing and *in-silico* prediction tools represents a step-forward into the applicability of SS and NTS in wide-scope campaigns by facilitating the identification process [5–7]. Furthermore, the establishment of community-adopted levels of confidence for the identification of compounds using chromatography coupled to HRMS has been of paramount importance for the comparison of data across studies [8].

Recently, ion mobility separation (IMS) coupled to HRMS instruments (IMS-HRMS) has proven promising for SS and NTS strategies [9]. It permits, in theory, to resolve co-eluting compounds with same nominal or exact mass that could not be previously separated with solely the chromatographic method, such as isobaric or isomeric compounds [9–11]. Moreover, it allows the removal of mass spectrometric peaks that do not correspond to the feature of interest, which is particularly beneficial in data independent acquisition (DIA) experiments [9, 10, 12]. As a consequence, there is a reduction in the necessity of data-dependent analysis since the full-spectrum HRMS acquisition can be filtered on both RT and ion mobility data [12, 13].

Collision Cross Section (CCS) values, derived from drift time (DT) measured by IMS, are known to be system and matrix-independent and, therefore, empirical CCS data can be included in home-made or online databases with an expected deviation below 2% for most cases [9, 14, 15]. However, this is not the case for chromatographic retention time (RT) which cannot easily be compared between instrumental configurations. Thus, reference standards are practically required for building a home-made database. However, SS and NTS strategies for the identification of emerging contaminants are commonly applied prior to the acquisition of the corresponding reference standards [1, 3] and, therefore, lacking any information on empirical RT and CCS. In this sense, *in-silico* prediction tools of either chromatographic retention data or ion mobility data are of great utility to decrease the number of candidates to investigate and, therefore, increase the chance of correct identification of features [6].

Several studies have predicted RT [6, 16–21], CCS values [7, 22–28] or both [13]. Predictors of RT have been developed mainly to model RT data in reverse-phase liquid chromatography (RPLC) and hydrophilic interaction liquid chromatography (HILIC) with prediction accuracy between approximately ± 1 to ± 3 min. However, there is no clear agreement in the literature on how to express the prediction accuracy of the models or which should be the most appropriate statistical descriptor representing the prediction power of the system developed [6]. Although CCS could be theoretically modelled from the three-dimensional and chemical structure using super-computing systems [27, 29–31], data-driven predictive models have also been developed showing predictive accuracies in the range of 3 –6 % for Travelling Wave Ion Mobility instruments (TWIMS) [22, 24, 25, 28] and

Drift Tube Ion Mobility instruments (DTIMS) [23, 24, 26]. Similar prediction accuracy was obtained by Mollerup et al. in their study for the simultaneous prediction of RT and CCS [13]. However, these data-driven models were fed with data generated using different instruments depending on the output parameter. For the RT prediction they used data gathered from an ultra-high performance liquid chromatography (UHPLC)-HRMS instrument, while for CCS prediction they modelled CCS data generated with a UHPLC-IMS-HRMS instrument. Since RT variations could probably be observed across instruments, the utility of predicted RT in the identification of UHPLC-IMS-HRMS features is limited.

In general, the reported models were based on univariate or multivariate regressions [20, 28], artificial neural networks (ANNs) [13, 18, 21, 22, 24], quantitative structure-retention relationships (QSRR) [6, 17, 32], supported vector regression (SVR) [23, 26] or statistical analysis [25, 28]. However, there has been no prior exploration of Multivariate Adaptive Regression Splines (MARS) for the prediction of RT and CCS. MARS is a multivariate non-parametric regression procedure that was first proposed by Friedman [33]. One of the biggest advantages of MARS compared to the 'black box' methods of ANNs is that they are easy to interpret, with the interactions between variables clearly indicated [34]. MARS has previously been applied in the chemical sciences for quantitative structure-retention relationships [35]. However, the application of MARS for the prediction of chromatographic and ion mobility data of emerging contaminants has not previously been evaluated and reported in the literature.

In this work, a prediction model for both RT and CCS has been developed using MARS for the identification of candidates in SS and NTS strategies using UHPLC-IMS-HRMS. To facilitate other laboratories implementing this predictive tool in their workflows, a free online-available application has been released. This is, to best of the authors knowledge, the first application of MARS for the prediction of RT and CCS data. Additionally, it is the first parallel RT and CCS predictive model for the same instrument facilitating the identification process of emerging contaminants in SS and NTS strategies.

2. Materials And Methods

2.1 Chemicals and materials

A set of 556 reference standards encompassing illicit drugs, hormones, mycotoxins, new psychoactive substances, pesticides and pharmaceuticals were injected for the development of a CCS and RT library [9]. **Table S1** of the **Supporting Information** shows the complete set of compounds used in the study with their SMILES (simplified molecular-input line-entry system) representation, and measured RT and CCS data. This database is also available on the *Zenodo* online repository [36]. Within this dataset, 477 protonated adducts ($[M+H]^+$), 169 deprotonated adducts ($[M-H]^-$) and 249 sodium adducts ($[M+Na]^+$) were used for the development and validation of the CCS predictive models.

2.2 Instrumentation

Retention time and CCS data were obtained with a Waters Acquity I-Class UPLC system (Waters, Milford, MA, USA) coupled to a VION IMS-QTOF mass spectrometer (Waters, Milford, MA, USA), using an electrospray ionization (ESI) interface operating in positive and negative ionisation mode and following the method presented in Celma et al. (2020) [9].

The chromatographic column used was a CORTECS® C18 2.1 x 100 mm, 2.7 µm fused core column (Waters) at a flow rate of 300 µL min^{-1}. Gradient elution was performed using H_2O (A) and MeOH (B) as mobile phases, both with 0.01% formic acid. The percentage of B was initially set to 10%, and it was immediately linearly increased to 90% over 14 min, followed by a 2 min isocratic period, and then returned to initial conditions (at 16.1 min) with a 2 min equilibration of the column. The total run time was 18 min. The injection volume was 5 µL.

A capillary voltage of 0.8 kV and cone voltage of 40 V were used. The desolvation temperature was set to 550 °C, and the source temperature to 120 °C. Nitrogen was used as drying and nebulizing gas. The cone gas flow was 250 L h^{-1} and desolvation gas flow of 1000 L h^{-1}. The column temperature was set to 40 °C and the sample temperature to 10 °C. MS data

were acquired using the VION in HDMSe mode, over the range *m/z* 50-1000, with N_2 as the drift gas, an IMS wave velocity of 250 m s^{-1} and wave height ramp of 20-50 V. Leucine enkephalin (*m/z* 556.27658 and *m/ z* 554.26202) was used for mass correction in positive and negative ionization modes, respectively. Two independent scans with different collision energies were acquired during the run: a collision energy of 6 eV for low energy (LE) and a ramp of 28-56 eV for high energy (HE). A scan time of 0.3 s was set in both LE and HE functions. Nitrogen (≥ 99.999%) was used as collision-induced dissociation (CID) gas. All data were examined using an in-house built accurate mass screening workflow within the UNIFI software (version 1.9.4) from Waters Corporation.

2.3 Retention time and Collision Cross Section Modelling

2.3.1 Molecular descriptors

A total of 1666 molecular descriptors were downloaded from Dragon v5.4 integrated within OChem website (Online Chemical Database with modelling environment, www.ochem.eu) [37]. The complete set of descriptors for the molecules used in the study is available in **Table S1**.

2.3.2 Prediction model

Multivariate Adaptive Regression Splines (MARS) analysis was applied to predict both RT and CCS for protonated adducts ($[M+H]^+$) in a single multivariate model. Additionally, univariate models for individual RT and CCS for protonated adducts ($[M+H]^+$) (CCS_H) and sodium adducts ($[M+Na]^+$) (CCS_{Na}) were also performed. Due to the expected low correlation between RT and CCS (*r=0.354*), a multivariate model was not considered essential. As a further justification for this decision, the cross-validated R^2 values for the multivariate model were 0.798 for RT and 0.964 for CCS_H. This suggests instability on the data that is varying the accuracy of the model fits (particularly for RT). Therefore, the development of a multivariate MARS model able to predict simultaneously RT and CCS simultaneously was discarded.

MARS was able to select the most suitable molecular descriptors for each model (**Table 1**) and predictive interval bands were constructed for the univariate cases assuming a linear model variance structure. To meet this assumption, the square root of RT was modelled.

The CCS_H prediction model was also explored for the prediction of CCS for deprotonated adducts ($[M-H]^-$) and sodium adducts ($[M+Na]^+$). CCS_H accurately modelled $[M-H]^-$ data, but could not predict data at acceptable levels of accuracy for $[M+Na]^+$. Therefore, a exclusive univariate model was considered for the prediction of CCS data for sodium adducts (CCS_{Na}).

All analyses were complete using R [38] and MARS analysis was completed using the earth package with variance structure defined using the linear model (lm) option [39].

3. Results And Discussion

3.1. Development and validation of prediction models

3.1.1. Individual RT and CCS model development

There is no assumption of an underlying variance structure with the multivariate MARS analysis, and there was no facility to define one within the earth package at the time of implementation. However, for the univariate analyses, a linear model variance structure was defined. This meant the standard deviation was estimated as a function of the predicted response and, hence, allowed for the construction of prediction intervals.

It is essential to use prediction intervals, rather than confidence intervals, in cases where the goal is to predict future values. A prediction interval is wider than a confidence interval and, at the 95% level, will provide bounds within which 95% of predicted values should fall.

All analyses considered the whole set of 1666 molecular descriptors as possible inputs to be used in the models. The assumptions of normality, linearity and homoscedasticity were assessed for the univariate models which held those assumptions. The univariate MARS fit to RT violated the assumptions of linearity and homoscedasticity, so a square root transform was applied. This then reasonably met assumptions.

In summary, three different univariate models were developed for the prediction of RT (**Equation 1**), CCS data for (de)protonated molecules (CCS_H) (**Equation 2**) and CCS data for sodium adducts (CCS_{Na}) (**Equation 3).** As an example and to assist with interpretation, in **equation 1**, the term 0.099·max(0,(nDB-3)) is equal to 0 for nDB ≤ 3, and equal to 0.099·(nDB-3) for nDB > 3.

The univariate models obtained a cross validated R^2=0.855 for the RT model, R^2=0.966 for the CCS_H model and R^2=0.954 for the CCS_{Na} model. **Table 1** reveals that the univariate models (RT and CCS_H) do not share a single descriptor, lending weight toward the argument that univariate models provide better fits to the data than previously explored multivariate model.

RT model

$$\sqrt{RT} = 2.343 - 0.171 \cdot \max(0, (4.22 - ALOGPS\_logP)) + 0.099 \cdot \max(0, (nDB - 3)) - 0.086 \cdot \max(0, (3 - nDB)) - 0.451 \cdot \max(0, (N.068 - 1)) + 0.725 \cdot \max(0, (1 - N.068)) + 0.632 \cdot \max(0, (1 - nRNHR)) - 2.177 \cdot \max(0, (BEHm4 - 3.582)) - 0.533 \cdot \max(0, (3.582 - BEHm4)) - 1.565 \cdot \max(0, (Mor16m - 0.54)) + 0.111 \cdot \max(0, (ALOGP - 2.719)) - 0.234 \cdot \max(0, (2.719 - ALOGP)) + 0.114 \cdot \max(0, (O.059 - 1)) - 0.138 \cdot \max(0, (1 - O.059)) - 3.185 \cdot \max(0, (GATS1m - 1.422)) - 0.132 \cdot \max(0, (STN - 6.985)) \quad \textbf{(Eq. 1)}$$

CCS_H model

$$CCS_H = 203.344 + 0.482 \cdot \max(0, (AMR - 94.347)) - 0.524 \cdot \max(0, (94.347 - AMR)) - 0.002 \cdot \max(0, (Whetp - 1940.49)) - 0.836 \cdot \max(0, (9.95 - L1m)) - 14.618 \cdot \max(0, (PCR - 1.109)) + 36.31 \cdot \max(0, nRCHO) + 0.361 \cdot \max(0, (LPRS - 171.967)) - 0.157 \cdot \max(0, (171.967 - LPRS)) - 0.74 \cdot \max(0, (28.622 - MDDD)) \quad \textbf{(Eq. 2)}$$

CCS_{Na} model

$$CCS_{Na} = 197.356 - 0.252 \cdot \max(0, (102.616 - AMR)) + 0.575 \cdot \max(0, (\mathrm{Har2} - 117.656)) - 0.793 \cdot \max(0, (117.656 - Har2)) - 5.873 \cdot \max(0, (nR09 - 1)) - 5.475 \cdot \max(0, (1 - nR09)) + 0.046 \cdot \max(0, (158.403 - QXXv)) + 0.074 \cdot \max(0, (527.605 - \mathrm{ZM1V})) - 0.038 \cdot \max(0, (470.721 - QZZm)) + 8.192 \cdot \max(0, (\mathrm{Mor17m} - 0.302)) + 12.649 \cdot \max(0, (-0.302 - Mor17m)) + 0.116 \cdot \max(0, (171.057 - \mathrm{piID})) + 1.392 \cdot \max(0, (MAXDN - 2.491)) + 4.682 \cdot \max(0, (2.491 - MAXDN)) - 0.442 \cdot \max(0, (\mathrm{RDF065v} - 6.402)) \quad \textbf{(Eq. 3)}$$

3.1.2. RT, CCS_H and CCS_{Na} model validation

MARS models were fitted using a 3-fold cross validation with thirty iterations. This procedure splits the data into three sections, fits the model to two of those sections (*training data*) and then tests the accuracy of the resulting model on the final section (*test data*). This procedure is then repeated thirty times, each time randomly dividing the data into three sections. The measure of accuracy used to assess goodness of fit is the cross-validated R^2, which looks at the average R^2 value obtained across all thirty iterations when the model was fit to the test data. This value is usually lower than the R^2 for the best model fit but dramatic changes suggest volatility in the data or overfitting in the modelling procedure.

In order to perform an additional model validation and to obtain an overview of the model performance, RT and CCS data was predicted for the molecules used for model development. By comparing predicted and empirical RT data (**Figure 1A, top**), it was observed that the average deviation obtained using RT model (**eq. 1**) was ± 0.72 min as shown in **Table 2**. Yet, 95% of the predictions fell within ± 2.32 min. Additionally, it could also be observed that deviations in predicted data distributed normally around 0% deviation (marked as a red line in **Figure 1A, bottom**) The prediction accuracy obtained is an improvement for the 95% intervals in previously developed models (± 4.0 min using *logKow* predictor [20], ± 2.80 min using ANNs [21]) and in line with the model developed by means of ANN by Mollerup et al. (over ± 2 min deviation) [13]. The developed model herein presented also improves the prediction accuracy compared to Barron et al. where they obtained average deviation of ± 1.02 min [18]. As another way of presenting prediction accuracy, **Figure 2** plots the predicted vs. empirical data with the 95% prediction intervals (blue coloured area) for the univariate MARS analysis of the . Approximately, only 8% of predicted RT were more than 2 min away from empirical ones.

Prediction accuracy for CCS data was also studied. The deviation observed for CCS data of $[M+H]^+$ using CCS_H model averaged ± 1.23 %, being ± 4.05 % within 95% of the cases (**Table 2**). **Figure 1B,** bottom shows that deviations randomly distributed around 0% (marked as a red line) value without biasing predicted data. When compared with previous models, CCS data for protonated molecules could be predicted using ANNs with an accuracy of ± 5 – 6% for 95% of the cases [13,22] or slightly over ± 5 % deviation (95% confidence interval) using machine learning [25]. **Figure 3A** shows the 95% prediction intervals (blue coloured area) for the univariate MARS analysis on CCS_H model. The blue lines are placed at predicted values ± 2 $Å^2$ and the purple are ± 5 $Å^2$. It is clear that the model is still predicting well at higher values. However, since there is less data, the prediction intervals are much larger to accommodate the uncertainty. This vast improvement in the accuracy could be explained because of the larger database used for the model development as well as the better fitting of empirical data with MARS than ANNs.

Additionally, application of the CCS_H model for the prediction of CCS values for deprotonated molecules was tested, yielding highly accurate predictions (**Figure 1C, top**). By predicting mobility data for a set of 169 molecules ionized in negative mode, it was observed that the differences between the observed and predicted CCS for the $[M-H]^-$ fell, 95% of the time, within -13.4 and 9.3 $Å^2$, with a slight tendency to under-predict CCS values (**Figure 1C, bottom**). In relative terms, average deviation for $[M-H]^-$ data was ± 2.79 % (± 5.86% for the 95% of the cases, **Table 2**). Although these deviations seem larger than the ones observed for $[M+H]^+$ data, this increase in the deviations observed for $[M-H]^-$ was expected since the model was developed with $[M+H]^+$ data. However, it was assumed that the predictions of CCS_H model developed with $[M+H]^+$ data could also be extrapolated to the prediction of CCS data for $[M-H]^-$, as no remarkable improvement was expected if a model was exclusively developed for deprotonated molecules.

Ideally, a unique model for the prediction of CCS for (de)protonated molecules and sodium adducts was intended. Therefore, the CCS_H model was also tested against $[M+Na]^+$ data. However, high deviations were observed (± 4.77 % average, ± 10.86 % for the 95% of the cases, **Table 2**) which could be expected due to the likely higher impact of the volume of the sodium atom in the overall CCS of the molecule. In light of this data, $[M+Na]^+$ data required a separate model for CCS prediction that was different to the one initially developed. The procedure for CCS_{Na} model development was equivalent to the process described above (section 2.3) but using as input a dataset of 249 CCS values for $[M+Na]^+$ ions. The accuracy of the model was evaluated by also comparing predicted and empirical data (**Table 2**). Prediction deviations were ± 2.08 % on average (± 5.25 % for the 95% of the cases) showing a great improvement compared with predicted data using the CCS_H model. **Figure 3B** depicts the predicted vs. empirical CCS values comparing the 95% prediction intervals (blue coloured area) for the univariate MARS analysis on CCS_{Na} model. The fact that different predicted values can be obtained for both protonated molecules and sodium adducts is of great help for empirical observations of both species for a suspect substance. Hence, increased confidence on the tentative identification can be garnered by matching both of the CCS values observed with predicted data.

The CCS_{Na} model herein presented also improves the prediction accuracy of previously developed model by the authors [22]. In that work, we evaluated the performance of the ANN predictive model for sodium adducts finding that deviations between predicted and empirical data were below 8.7% for the 95% of the cases. However, the development of an exclusive model for the sodium adducts by MARS improves the prediction accuracy.

3.1.3. Blind testing of the models

Several reference standards were purchased from different research projects during the development of the predictors based on MARS. Hence, they were not included in the training and validation datasets used. These compounds were used to verify the utility of our prediction models for chemicals not previously considered in the training steps. Thus, model applicability can be extrapolated for upcoming RT and CCS predictions of real suspect compounds. Therefore, we calculated deviations between predicted and empirical data for this dataset, and compared the observed deviations with previously calculated accuracies at different percentiles (shown in **Table 2**). **Table 3** depicts the empirical and predicted values of RT and CCS for the different adducts observed for the additional set of 25 reference standards. Moreover, the deviation between empirical and predicted is shown and as it can be observed the RT predictions are generally in agreement with the empirical data with the 95th percentile of the observed deviations (± 4.15 min) being in the same range than that observed during validation. Furthermore, the vast majority of CCS values for $[M+H]^+$ are in agreement with the values calculated using the CCS_H model. For these compounds, 95% of the cases showed deviations below ± 3.71 %, yielding even better results than the initial database during model validation. Only *3,4-dichloroaniline* shows a deviation greater than 4%, which could be explained by the small CCS value calculated. When evaluating CCS_{Na}, higher deviations are observed concretely for the case of *di(2-ethylhexyl) terephthalate* and *vildagliptin* (-8.61% and 8.74%, respectively). These deviations could be explained because of particular chemical structures of the molecule such as the presence of an adamantyl group in *vildagliptin*, which has a large and rigid structure, or the high rotatability of alkyl chains in the *di(2-ethylhexyl) terephthalate*. However, if these adducts would be treated as outliers, 95% of

the CCS_{Na} values show deviations of ± 3.15 %, which is in great accordance with the data obtained during method validation. Finally, for $[M-H]^-$, a small set of molecules was gathered, and all of them fit well within the ± 5.8 % deviation.

3.2. Open access prediction platform

To aid future researchers working with UHPLC-IMS-HRMS, a free online webpage incorporating these models has been released. The models are available for the scientific community through https://datascience-adelaideuniversity.shinyapps.io/Predicting_RT_and_CCS/ . **Figure 4** illustrates the layout of the online platform for the prediction of RT and CCS for both (de)protonated molecules or sodium adducts.

The operational of the platform is user-friendly and easy-to-follow. As an example, the step-by-step method to obtain prediction for omeprazole is shown. First, selection of which parameter is going to be predicted need to be done (**Figure 4A**). In this case, CCS for protonated molecule is selected by indicating 'Select Response: Collision Cross Section' and 'Sodiated: No'. After downloading the appropriate descriptors for the molecule of interest using Dragon v5.4 integrated within OChem (www.ochem.eu) [37], those can be added in the corresponding editable fields (**Figure 4B**). The CCS value can, then, be predicted and the output is shown together with their corresponding prediction intervals (**Figure 4C**). In this case, the CCS predicted value for the protonated molecule of omeprazole is 181.51 $Å^2$ with a prediction interval of 171.93 – 190.08 $Å^2$. The empirical value for $[M+H]^+$ for omeprazole is 180.58 $Å^2$, denoting that the prediction only deviated 0.52% from the empirical value.

The ease of prediction as well as the open access for this online platform is of great help for those researchers working on UHPLC-IMS-HRMS instruments who do not have an in-house developed prediction model.

4. Conclusions

Three different prediction models using Multiple Adaptive Regression Splines have been developed for the prediction of RT, CCS for (de)protonated molecules and CCS for sodium adducts. This is the first application of MARS for the prediction of RT and CCS data. In addition, the reported models are the first parallel prediction of RT and CCS data for the same instrument, facilitating the identification process of chemicals of emerging concern in SS and NTS strategies. The developed predictive models make use of a set of 26 molecular descriptors to predict RT and/or CCS values. The prediction accuracy achieved with these models bettered previously reported models in the literature by reducing the deviation between predicted and empirical to ± 2.32 min for RT, ± 4.05% for CCS of protonated molecules, ± 5.86% for CCS of deprotonated molecules and ± 5.25% for CCS of sodium adducts (95% confidence intervals). Additionally, a free access online platform has been released to enable the application of these models to third-party laboratories interested in predicting RT and CCS data.

Declarations

Acknowledgements

A. Celma acknowledges the Spanish Ministry of Economy and Competiveness for his predoctoral grant (BES-2016-076914). L. Bijlsma acknowledges his fellowship funded by "la Caixa" Foundation. The project that gave rise to these results received the support of a fellowship from "la Caixa" Foundation (ID 10 0 010434). The fellowship code is LCF/BQ/PR21/11840012. Authors from University Jaume I acknowledge the financial support of Spanish Ministry of Science, Innovation and Universities (RTI2018-097417-B-100), of Generalitat Valenciana (Research Group of Excellence Prometeo 2019/040) and of University Jaume I of Castellón, Spain (project UJI-B2018-55 and UJI-B2020-19).

References

1. F. Hernández, J. Bakker, L. Bijlsma, J. de Boer, A.M. Botero-Coy, Y. Bruinen de Bruin, S. Fischer, J. Hollender, B. Kasprzyk-Hordern, M. Lamoree, F.J. López, T.L. te. Laak, J.A. van Leerdam, J. V. Sancho, E.L. Schymanski, P. de Voogt, E.A. Hogendoorn, The role of analytical chemistry in exposure science: Focus on the aquatic environment, Chemosphere. 222 (2019) 564–583. https://doi.org/10.1016/j.chemosphere.2019.01.118.
2. J. Hollender, B. van Bavel, V. Dulio, E. Farmen, K. Furtmann, J. Koschorreck, U. Kunkel, M. Krauss, J. Munthe, M. Schlabach, J. Slobodnik, G. Stroomberg, T. Ternes, N.S. Thomaidis, A. Togola, V. Tornero, High resolution mass spectrometry-based non-target screening can support regulatory environmental monitoring and chemicals management, Environ. Sci. Eur. 31 (2019). https://doi.org/10.1186/s12302-019-0225-x.
3. E.L. Schymanski, H.P. Singer, J. Slobodnik, I.M. Ipolyi, P. Oswald, M. Krauss, T. Schulze, P. Haglund, T. Letzel, S. Grosse, N.S. Thomaidis, A. Bletsou, C. Zwiener, M. Ibáñez, T. Portolés, R. De Boer, M.J. Reid, M. Onghena, U. Kunkel, W. Schulz, A. Guillon, N. Noyon, G. Leroy, P. Bados, S. Bogialli, D. Stipaničev, P. Rostkowski, J. Hollender, Non-target screening with high-resolution mass spectrometry: Critical review using a collaborative trial on water analysis, Anal. Bioanal. Chem. 407 (2015) 6237–6255. https://doi.org/10.1007/s00216-015-8681-7.
4. S. Samanipour, K. Langford, M.J. Reid, K. V. Thomas, A two stage algorithm for target and suspect analysis of produced water via gas chromatography coupled with high resolution time of flight mass spectrometry, J. Chromatogr. A. 1463 (2016) 153–161. https://doi.org/10.1016/j.chroma.2016.07.076.
5. N.A. Alygizakis, P. Oswald, N.S. Thomaidis, E.L. Schymanski, R. Aalizadeh, T. Schulze, M. Oswaldova, J. Slobodnik, NORMAN digital sample freezing platform: A European virtual platform to exchange liquid chromatography high resolution-mass spectrometry data and screen suspects in “digitally frozen” environmental samples, TrAC - Trends Anal. Chem. 115 (2019) 129–137. https://doi.org/10.1016/j.trac.2019.04.008.
6. R. Aalizadeh, M.-C. Nika, N.S. Thomaidis, Development and application of retention time prediction models in the suspect and non-target screening of emerging contaminants, J. Hazard. Mater. 363 (2019) 277–285. https://doi.org/10.1016/j.jhazmat.2018.09.047.
7. L. Bijlsma, M.H.G. Berntssen, S. Merel, A Refined Nontarget Workflow for the Investigation of Metabolites through the Prioritization by in Silico Prediction Tools, Anal. Chem. 91 (2019) 6321–6328. https://doi.org/10.1021/acs.analchem.9b01218.
8. E.L. Schymanski, J. Jeon, R. Gulde, K. Fenner, M. Ruff, H.P. Singer, J. Hollender, Identifying small molecules via high resolution mass spectrometry: Communicating confidence, Environ. Sci. Technol. 48 (2014) 2097–2098. https://doi.org/10.1021/es5002105.
9. A. Celma, J.V. Sancho, E.L. Schymanski, D. Fabregat-Safont, M. Ibáñez, J. Goshawk, G. Barknowitz, F. Hernández, L. Bijlsma, Improving Target and Suspect Screening High-Resolution Mass Spectrometry Workflows in Environmental Analysis by Ion Mobility Separation, Environ. Sci. Technol. 54 (2020) 15120–15131. https://doi.org/10.1021/acs.est.0c05713.
10. J. Regueiro, N. Negreira, M.H.G. Berntssen, Ion-mobility-derived collision cross section as an additional identification point for multiresidue screening of pesticides in fish feed, Anal. Chem. 88 (2016) 11169–11177. https://doi.org/10.1021/acs.analchem.6b03381.
11. M. McCullagh, K. Giles, K. Richardson, S. Stead, M. Palmer, Investigations into the performance of travelling wave enabled conventional and cyclic ion mobility systems to characterise protomers of fluoroquinolone antibiotic residues, Rapid Commun. Mass Spectrom. 33 (2019) 11–21. https://doi.org/10.1002/rcm.8371.
12. A. Celma, L. Ahrens, P. Gago-Ferrero, F. Hernández, F. López, J. Lundqvist, E. Pitarch, J.V. Sancho, K. Wiberg, L. Bijlsma, The relevant role of ion mobility separation in LC-HRMS based screening strategies for contaminants of emerging concern in the aquatic environment, Chemosphere. 280 (2021) 130799. https://doi.org/10.1016/j.chemosphere.2021.130799.
13. C.B. Mollerup, M. Mardal, P.W. Dalsgaard, K. Linnet, L.P. Barron, 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, J. Chromatogr. A. 1542 (2018) 82–88. https://doi.org/10.1016/j.chroma.2018.02.025.

14. V. Gabelica, E. Marklund, Fundamentals of ion mobility spectrometry, Curr. Opin. Chem. Biol. 42 (2018) 51–59. https://doi.org/10.1016/j.cbpa.2017.10.022.
15. J.W. Lee, Basics of ion mobility mass spectrometry, Mass Spectrom. Lett. 8 (2017) 79–89. https://doi.org/10.5478/MSL.2017.8.4.79.
16. D. Yeung, N. Klaassen, B. Mizero, V. Spicer, O. V. Krokhin, Peptide retention time prediction in hydrophilic interaction liquid chromatography: Zwitter-ionic sulfoalkylbetaine and phosphorylcholine stationary phases, J. Chromatogr. A. (2020) 1–8. https://doi.org/10.1016/j.chroma.2020.460909.
17. J.J. Yang, Y. Han, C.H. Mah, E. Wanjaya, B. Peng, T.F. Xu, M. Liu, T. Huan, M.L. Fang, Streamlined MRM method transfer between instruments assisted with HRMS matching and retention-time prediction, Anal. Chim. Acta. 1100 (2020) 88–96. https://doi.org/10.1016/j.aca.2019.12.002.
18. L.P. Barron, G.L. McEneff, Gradient liquid chromatographic retention time prediction for suspect screening applications: A critical assessment of a generalised artificial neural network-based approach across 10 multi-residue reversed-phase analytical methods, Talanta. 147 (2016) 261–270. https://doi.org/10.1016/j.talanta.2015.09.065.
19. J. Stanstrup, S. Neumann, U. Vrhovšek, PredRet: Prediction of Retention Time by Direct Mapping between Multiple Chromatographic Systems, Anal. Chem. 87 (2015) 9421–9428. https://doi.org/10.1021/acs.analchem.5b02287.
20. R. Bade, L. Bijlsma, J.V. Sancho, F. Hernández, Critical evaluation of a simple retention time predictor based on LogKow as a complementary tool in the identification of emerging contaminants in water, Talanta. 139 (2015) 143–149. https://doi.org/10.1016/j.talanta.2015.02.055.
21. R. Bade, L. Bijlsma, T.H. Miller, L.P. Barron, J.V. Sancho, F. Hernández, Suspect screening of large numbers of emerging contaminants in environmental waters using artificial neural networks for chromatographic retention time prediction and high resolution mass spectrometry data analysis, Sci. Total Environ. 538 (2015) 934–941. https://doi.org/10.1016/j.scitotenv.2015.08.078.
22. L. Bijlsma, R. Bade, A. Celma, L. Mullin, G. Cleland, S. Stead, F. Hernandez, J. V. Sancho, Prediction of Collision Cross-Section Values for Small Molecules: Application to Pesticide Residue Analysis, Anal. Chem. 89 (2017) 6583–6589. https://doi.org/10.1021/acs.analchem.7b00741.
23. Z. Zhou, J. Tu, X. Xiong, X. Shen, Z.-J. Zhu, LipidCCS: Prediction of Collision Cross-Section Values for Lipids with High Precision To Support Ion Mobility–Mass Spectrometry-Based Lipidomics, Anal. Chem. 89 (2017) 9559–9566. https://doi.org/10.1021/acs.analchem.7b02625.
24. P.-L. Plante, É. Francovic-Fontaine, J.C. May, J.A. McLean, E.S. Baker, F. Laviolette, M. Marchand, J. Corbeil, Predicting Ion Mobility Collision Cross-Sections Using a Deep Neural Network: DeepCCS, Anal. Chem. (2019) acs.analchem.8b05821. https://doi.org/10.1021/acs.analchem.8b05821.
25. D.H. Ross, J.H. Cho, L. Xu, Breaking Down Structural Diversity for Comprehensive Prediction of Ion-Neutral Collision Cross Sections., Anal. Chem. (2020). https://doi.org/10.1021/acs.analchem.9b05772.
26. Z. Zhou, X. Shen, J. Tu, Z.-J. Zhu, Large-Scale Prediction of Collision Cross-Section Values for Metabolites in Ion Mobility-Mass Spectrometry, Anal. Chem. 88 (2016) 11084–11091. https://doi.org/10.1021/acs.analchem.6b03091.
27. S.M. Colby, D.G. Thomas, J.R. Nunez, D.J. Baxter, K.R. Glaesemann, J.M. Brown, M.A. Pirrung, N. Govind, J.G. Teeguarden, T.O. Metz, R.S. Renslow, ISiCLE: A Quantum Chemistry Pipeline for Establishing in Silico Collision Cross Section Libraries, Anal. Chem. 91 (2019) 4346–4356. https://doi.org/10.1021/acs.analchem.8b04567.
28. G.B. Gonzales, G. Smagghe, S. Coelus, D. Adriaenssens, K. De Winter, T. Desmet, K. Raes, J. Van Camp, Collision cross section prediction of deprotonated phenolics in a travelling-wave ion mobility spectrometer using molecular descriptors and chemometrics, Anal. Chim. Acta. 924 (2016) 68–76. https://doi.org/10.1016/j.aca.2016.04.020.
29. S.A. Ewing, M.T. Donor, J.W. Wilson, J.S. Prell, Collidoscope: An Improved Tool for Computing Collisional Cross-Sections with the Trajectory Method, J. Am. Soc. Mass Spectrom. 28 (2017) 587–596. https://doi.org/10.1007/s13361-017-1594-2.
30. J.W. Lee, H.H.L. Lee, K.L. Davidson, M.F. Bush, H.I. Kim, Structural characterization of small molecular ions by ion mobility mass spectrometry in nitrogen drift gas: Improving the accuracy of trajectory method calculations, Analyst. 143 (2018) 1786–1796. https://doi.org/10.1039/c8an00270c.

31. L. Zanotto, G. Heerdt, P.C.T. Souza, G. Araujo, M.S. Skaf, High performance collision cross section calculation-HPCCS, J. Comput. Chem. 39 (2018) 1675–1681. https://doi.org/10.1002/jcc.25199.
32. F. Falchi, S.M. Bertozzi, G. Ottonello, G.F. Ruda, G. Colombano, C. Fiorelli, C. Martucci, R. Bertorelli, R. Scarpelli, A. Cavalli, T. Bandiera, A. Armirotti, Kernel-Based, Partial Least Squares Quantitative Structure-Retention Relationship Model for UPLC Retention Time Prediction: A Useful Tool for Metabolite Identification, Anal. Chem. 88 (2016) 9510–9517. https://doi.org/10.1021/acs.analchem.6b02075.
33. J.H. Friedman, Multivariate Adaptive Regression Splines, Ann. Stat. 19 (1991) 1–67.
34. R. Put, Q.S. Xu, D.L. Massart, Y. Vander Heyden, Multivariate adaptive regression splines (MARS) in chromatographic quantitative structure–retention relationship studies, J. Chromatogr. A. 1055 (2004) 11–19. https://doi.org/10.1016/j.chroma.2004.07.112.
35. R. Put, Y. Vander Heyden, Review on modelling aspects in reversed-phase liquid chromatographic quantitative structure–retention relationships, Anal. Chim. Acta. 602 (2007) 164–172. https://doi.org/10.1016/j.aca.2007.09.014.
36. A. Celma, D. Fabregat-Safont, M. Ibáñez, L. Bijlsma, F. Hernández, J. V. Sancho, S61 | UJICCSLIB | Collision Cross Section (CCS) Library from UJI (Version NORMAN-SLE S61.0.1.2) [Data set], Zenodo. (2019). https://doi.org/10.5281/zenodo.3966751.
37. I. Sushko, S. Novotarskyi, R. Körner, A.K. Pandey, M. Rupp, W. Teetz, S. Brandmaier, A. Abdelaziz, V. V. Prokopenko, V.Y. Tanchuk, R. Todeschini, A. Varnek, G. Marcou, P. Ertl, V. Potemkin, M. Grishina, J. Gasteiger, C. Schwab, I.I. Baskin, V.A. Palyulin, E. V. Radchenko, W.J. Welsh, V. Kholodovych, D. Chekmarev, A. Cherkasov, J. Aires-de-Sousa, Q.-Y. Zhang, A. Bender, F. Nigsch, L. Patiny, A. Williams, V. Tkachenko, I. V. Tetko, Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information, J. Comput. Aided. Mol. Des. 25 (2011) 533–554. https://doi.org/10.1007/s10822-011-9440-2.
38. R.C. Team, R: A language and environment for statistical computing., (2020). https://www.r-project.org/.
39. Stephen Milborrow. Derived from mda.mars by T. Hastie and R. Tibshirani, earth: Multivariate Adaptive Regression Splines (R package v5.3.0), (2011). http://cran.r-project.org/package=earth.
40. V.N. Viswanadhan, A.K. Ghose, G.R. Revankar, R.K. Robins, Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain, J. Chem. Inf. Model. 29 (1989) 163–172. https://doi.org/10.1021/ci00063a006.
41. I. V. Tetko, V.Y. Tanchuk, Application of Associative Neural Networks for Prediction of Lipophilicity in ALOGPS 2.1 Program, J. Chem. Inf. Comput. Sci. 42 (2002) 1136–1145. https://doi.org/10.1021/ci025515j.
42. I. V Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V.A. Palyulin, E. V Radchenko, N.S. Zefirov, A.S. Makarenko, V.Y. Tanchuk, V. V Prokopenko, Virtual computational chemistry laboratory–design and description., J. Comput. Aided. Mol. Des. 19 (2005) 453–63. https://doi.org/10.1007/s10822-005-8694-y.

Tables

Table 1: Descriptors needed for each of the univariate MARS models for retention time (RT) and collision cross section (CCS_H and CCS_{Na}). Note that there are no similarities between the three univariate models.

Molecular descriptors		
RT	**CCS_H**	**CCS_{Na}**
ALOGP	AMR	AMR
ALOGPS_LogP	L1m	Har2
BEHm4	LPRS	MAXDN
GATS1m	MDDD	Mor17m
Mor16m	nRCHO	nR09
N-068	PCR	piID
nDB	Whetp	QXXv
nRNHR		QZZm
O-059		RDF065v
STN		ZM1v

ALOGP: Ghose-Crippen octanol-water partition coefficient (logP) (calculation based on Viswanadhan et al. (1989) [40]; *ALOGPS_LogP*: Ghose-Crippen octanol-water partition coefficient (logP) (calculation based on Tetko and Tanchuk (2002) [41]; *AMR*: Ghose-Crippen molar refractivity; *BEHm4*:highest eigenvalue n. 4 of Burden matrix / weighted by atomic masses; *GATS1m*: Geary autocorrelation - lag 1 / weighted by atomic masses; *Har2*: square reciprocal distance sum index; *L1m*: 1st component size directional WHIM index / weighted by atomic masses; *LPRS*: log of product of row sums; *MAXDN*: maximal electrotopological negative variation; *MDDD*: mean distance degree deviation; *Mor16m*: 3D-MoRSE - signal 16 / weighted by atomic masses; *Mor17m*: 3D-MoRSE - signal 17 / weighted by atomic masses; *N-068*: Al3-N atom-centred fragment; *nDB*: number of double bonds; *nR09*: number of 9-membered rings; *nRCHO*: number of (aliphatic) aldehydes; *nRNHR*: number of secondary (aliphatic) amines; *O-059*: Al-O-Al atom-centered fragment; *PCR*: ratio of multiple path count over path count; *piID*: conventional bond-order ID number; *QXXv*: Qxx COMMA2 value / weighted by atomic van der Waals volumes; *QZZm*: Qzz COMMA2 value / weighted by atomic masses; *RDF065v*: radial distribution function - 6.5 / weighted by atomic van der Waals volumes; *STN*: spanning tree number (log); *Whetp*: Wiener-type index from polarizability weighted distance matrix; *ZM1v*: first Zagreb index by valence vertex degrees [42].

Table 2. Deviations at percentiles 50 (average), 95 and 99 for the predicted RT and CCS data during model validation.

Model		Average deviation	Deviation at 95%	Deviation at 99%
RT		± 0.72 min	± 2.32 min	± 3.82 min
CCS_H	$[M+H]^+$	± 1.23 %	± 4.05 %	± 6.33 %
	$[M-H]^-$	± 2.79 %	± 5.86 %	± 8.39 %
	$[M+Na]^+$	± 4.77 %	± 10.86 %	± 12.80 %
CCS_{Na}	$[M+Na]^+$	± 2.08 %	± 5.25 %	± 6.86 %

Table 3. Empirical and predicted values of RT and CCS for additional compounds not initially included in datasets. Investigation of the deviation of predicted values.

Compound	Retention Time (min)			CCS_H for $[M+H]^+$			CCS_{Na} for $[M+Na]^+$			CCS_H for $[M-H]^-$		
	Emp.	Pred.	Dev (min)	Emp.	Pred.	Dev (%)	Emp.	Pred.	Dev (%)	Emp.	Pred.	Dev (%)
(-)-Cotinine	0.87	3.05	2.18	141.48	136.12	-3.79%						
3,4-dichloroaniline	7.92	6.21	-1.71	137.10	125.87	-8.19%						
3-Hydroxyphenyl diphenyl phosphate	11.09	10.36	-0.73	174.31	178.94	2.66%	184.84	189.67	2.61%	180.89	178.94	-1.07%
5,6-Dimethylbenzotriazole	6.74	4.38	-2.36	129.73	127.21	-1.94%				129.38	127.21	-1.68%
8-hydroxyquinoline	1.51	4.57	3.06	125.01	123.54	-1.18%						
Amisulpride	2.46	1.99	-0.47	193.15	189.60	-1.84%						
Antiblaze V6	10.84	14.49	3.65	208.45	207.32	-0.54%	208.45	212.12	1.76%			
Benzotriazole	3.50	2.75	-0.75	121.49	117.94	-2.92%						
BClPHP phosphate [a]	8.07	7.69	-0.38	159.22	157.05	-1.36%	165.14	166.24	0.67%			
Caffeine	3.08	1.94	-1.14	136.62	136.37	-0.18%						
Chlorotoluron	2.54	6.79	4.25	146.29	146.00	-0.20%	155.35	157.96	1.68%			
Citalopram	6.49	5.21	-1.28	179.10	184.48	3.01%						
Di(2-ethylhexyl) terephthalate	16.86	15.02	-1.84				216.36	197.07.23	-8.61%			
Diphenyl hydrogen phosphate	12.46	5.06	-7.41	152.45	151.61	-0.55%	161.58	162.65	0.66%	152.18	151.61	-0.38%
Diphenylcresyl phosphate	7.36	11.07	3.72	175.28	178.08	1.60%						
Metolachlor ESA [b]	7.89	4.99	-2.90	168.38	171.29	1.73%	175.57	179.13	2.03%	174.30	171.29	-1.73%
Metoxuron	5.98	7.04	1.06	149.83	150.62	0.53%	158.51	161.17	1.68%			
Mono(2-ethylhexyl) phthalate	12.73	11.75	-0.98							170.91	167.767215	-1.84%
Monuron	6.68	5.67	-1.01	140.59	142.94	1.67%						
Nicotine	0.69	1.11	0.42	138.34	134.77	-2.58%						
Niflumic acid	11.51	10.86	-0.65	157.46	157.79	0.21%				156.92	157.79	0.55%
Pirbuterol	1.30	1.28	-0.02	153.78	156.91	2.04%	160.02	165.52	3.44%			
Prometon	6.74	7.50	0.76	156.67	155.56	-0.71%						
Trietazine	10.81	8.91	-1.90	150.63	151.12	0.33%						
Vildagliptin	1.38	1.67	0.29	176.98	174.62	-1.33%	172.29	187.35	8.74%			

[a] Bis(1-chloro-2-propyl) 1-hydroxy-2-propyl phosphate; [b] Metolachlor ethane sulfonic acid

Figures

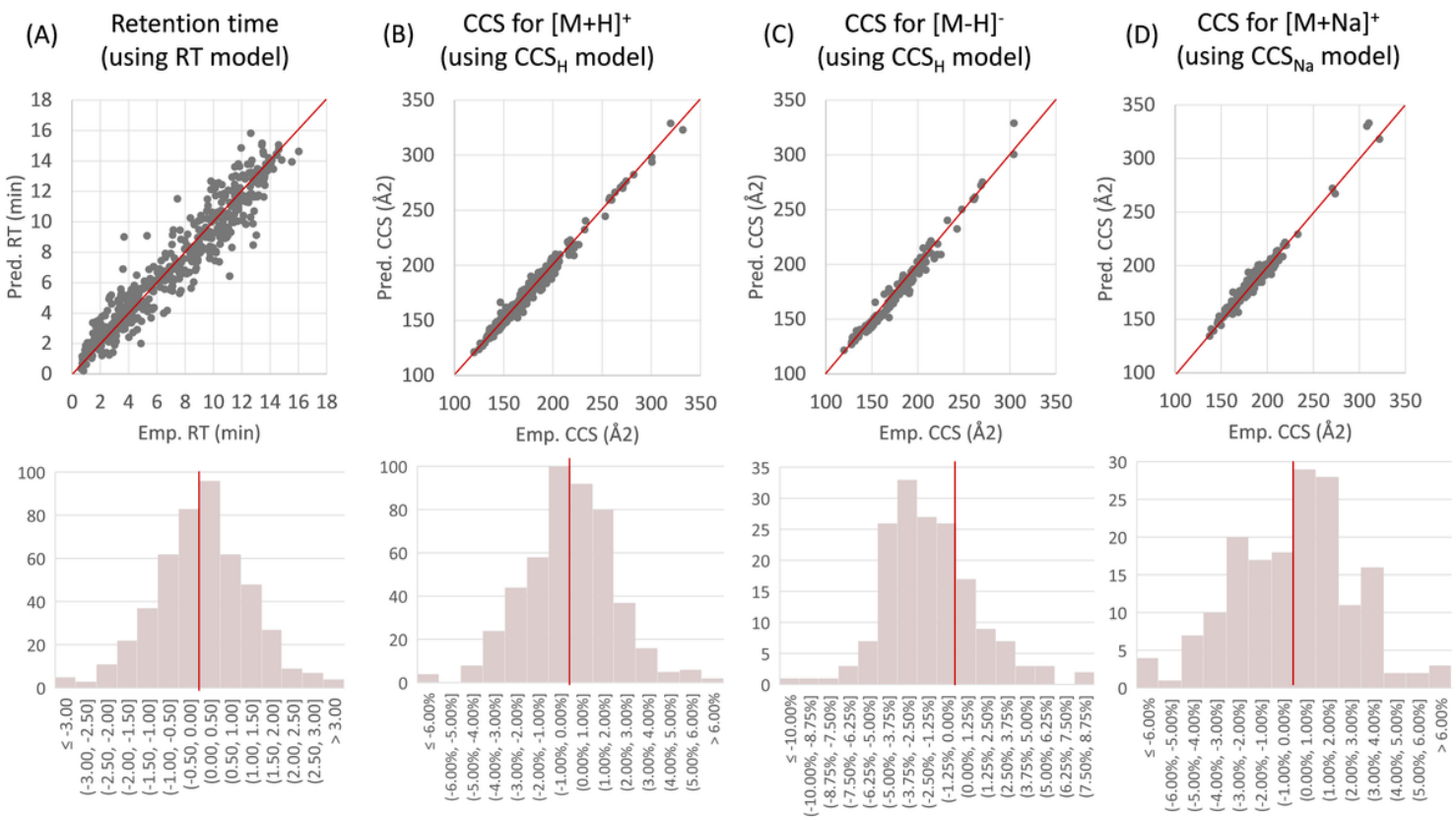

Figure 1

Top: Comparison of empirical and predicted RT data using RT model (A), CCS for protonated molecules using CCS_H model (B), CCS for deprotonated molecules using CCS_H model (C), and CCS for sodium adducts using CCS_{Na} model (D). (Red line indicates region where Empirical CCS = Predicted CCS) *Bottom:* Histogram distribution of deviations between empirical and predicted data for RT data using RT model (A), CCS for protonated molecules using CCS_H model (B), CCS for deprotonated molecules using CCS_H model (C), and CCS for sodium adducts using CCS_{Na} model (D). (Red vertical lines indicates 0% deviation)

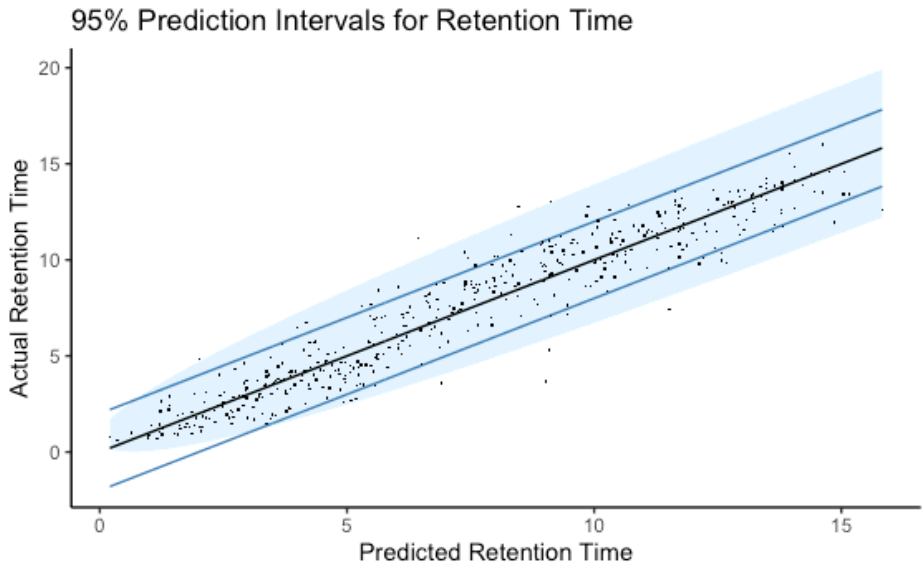

Figure 2

95% prediction intervals (blue coloured area) for the univariate MARS analysis on the square root of RT. The blue lines are placed at the predicted values ± 2 min. Approximately, only 8% of observed retention times were more than 2 minutes away from their predicted value.

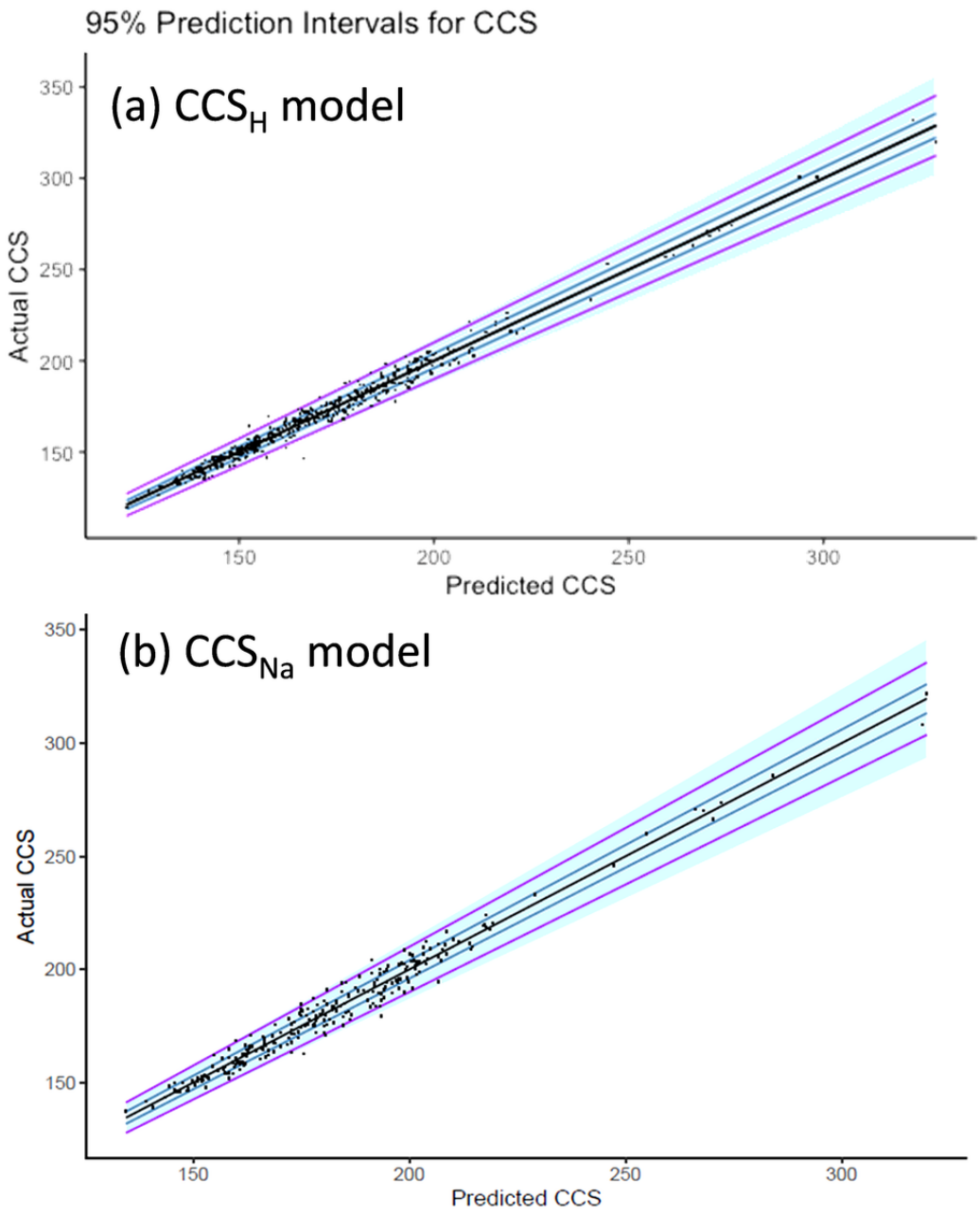

Figure 3

95% prediction intervals (blue coloured area) for the univariate MARS analysis on (a) CCS_H and (b) CCS_{Na} models. The blue lines are placed at 2% error bands and the purple at 5%. It is clear that the model is still predicting well at higher values where there is less data but the prediction intervals are much larger to accommodate the uncertainty due to lack of data.

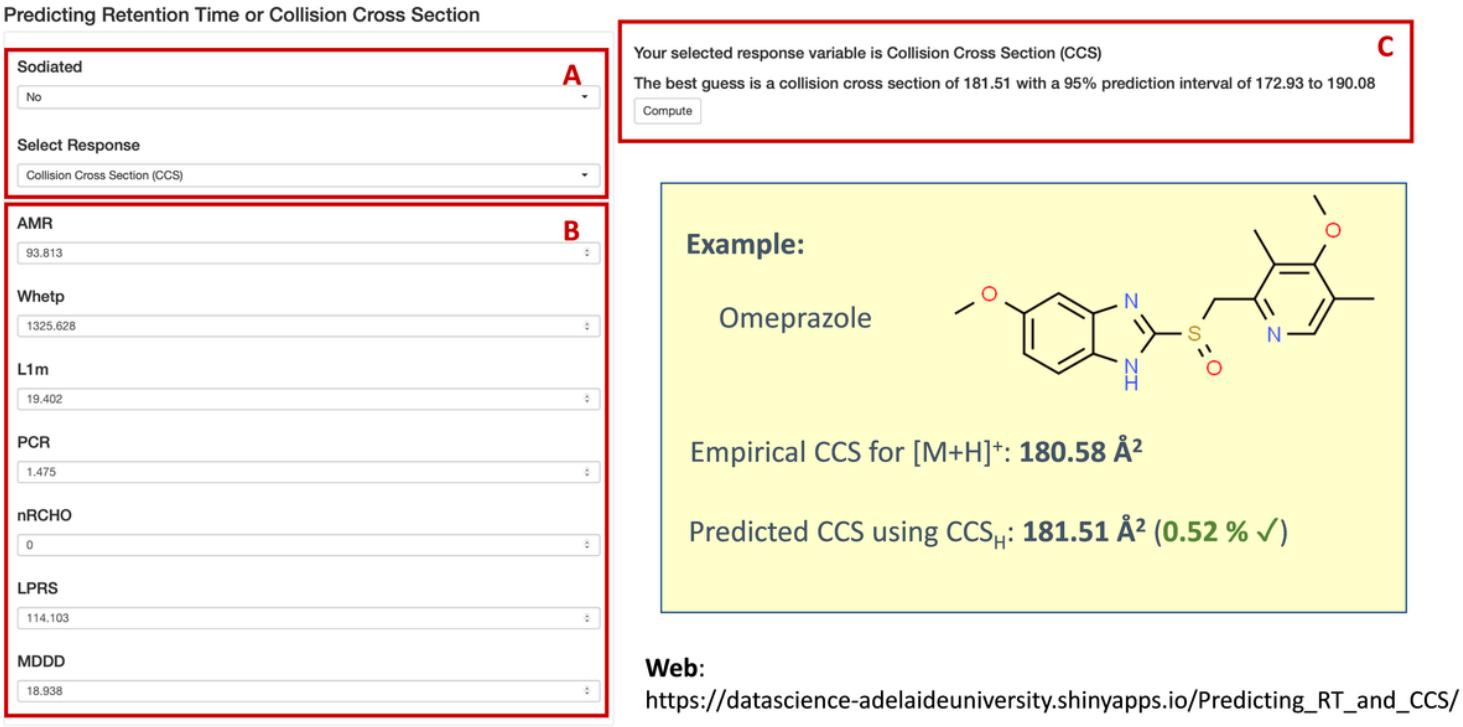

Figure 4

Online platform for the prediction of RT and CCS data using univariate models. (A) Selection of response to predict i.e. RT, CCS for (de)protonated molecules or CCS for sodium adducts; (B) Introduction molecular descriptors for the molecule of interest; (C) Output of the predictor model together with the prediction intervals. Example illustrated by omeprazole.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS1.xlsx