Analytica Chimica Acta 1184 (2021) 339035

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Development of a single retention time prediction model integrating multiple liquid chromatography systems: Application to new psychoactive substances

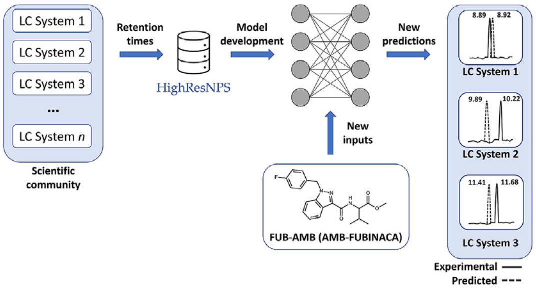
Daniel Pasin *, Christian Brinch Mollerup , Brian Schou Rasmussen , Kristian Linnet , Petur Weihe Dalsgaard

Section of Forensic Chemistry, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

HIGHLIGHTS

- A retention time prediction model was developed incorporating multiple LC systems.
- The model was trained using the retention times of new psychoactive substances.
- Retention times were obtained from the crowd-sourced database, HighResNPS.
- A singular model demonstrated improved performance over individual models.
- The model can be used to predict retention times for unique entries on HighResNPS.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 15 April 2021
Received in revised form
1 September 2021
Accepted 2 September 2021
Available online 6 September 2021

Keywords:
New psychoactive substances
Retention time prediction
High-resolution mass spectrometry
Suspect screening

ABSTRACT

Database-driven suspect screening has proven to be a useful tool to detect new psychoactive substances (NPS) outside the scope of targeted screening; however, the lack of retention times specific to a liquid chromatography (LC) system can result in a large number of false positives. A singular stream-lined, quantitative structure-retention relationship (QSRR)-based retention time prediction model integrating multiple LC systems with different elution conditions is presented using retention time data ($n = 1281$) from the online crowd-sourced database, HighResNPS. Modelling was performed using an artificial neural network (ANN), specifically a multi-layer perceptron (MLP), using four molecular descriptors and one-hot encoding of categorical labels. Evaluation of test set predictions ($n = 193$) yielded coefficient of determination (R^2) and mean absolute error (MAE) values of 0.942 and 0.583 min, respectively. The model successfully differentiated between LC systems, predicting 54%, 81% and 97% of the test set within ±0.5, ±1 and ±2 min, respectively. Additionally, retention times for an analyte not previously observed by the model were predicted within ±1 min for each LC system. The developed model can be used to predict retention times for all analytes on HighResNPS for each participating laboratory's LC system to further support suspect screening.

* Corresponding author.
E-mail address: daniel.pasin@sund.ku.dk (D. Pasin).

https://doi.org/10.1016/j.aca.2021.339035

1. Introduction

The last decade has seen the proliferation of new psychoactive substances (NPS) with 730 substances monitored by the European Monitoring Centre for Drugs and Drug Addiction (EMCDDA) as of 2018 [1] and 950 being identified by the United Nations Office of Drugs and Crime (UNODC) Early Warning Advisory (EWA) as of 2020 [2]. Interestingly, the latest report from the UNODC EWA highlights those new identifications, which have typically been synthetic cannabinoid-dominated, have now been superseded by synthetic opioids since 2018. This is a concerning shift since synthetic opioids have demonstrated the greatest potential to cause harm in users due to their potency. The combination of a dynamic NPS market and the lack of certified reference materials (CRMs) has made it challenging, if not virtually impossible, for laboratories to develop and validate up-to-date targeted screening methods for these analytes [3–5].

This has resulted in the development of alternative approaches such as suspect and non-targeted screening to attempt to detect analytes outside the scope of routine targeted methods, without the need for CRMs [6–10]. The use of high-resolution mass spectrometry (HRMS) has further enabled these approaches since it has the capacity to measure accurate masses and collect full scan mass spectrometry (MS) and MS/MS which can be retrospectively interrogated [11]. Suspect screening generally involves interrogation of HRMS data using databases containing masses of precursor and product ions from known or theoretical substances and can be collected from online repositories such as mzCloud (https://www.mzcloud.org/), Human Metabolome Database (HMDB) [12], MassBank [13] and MassBank of North America (MoNA, http://mona.fiehnlab.ucdavis.edu/), or from scientific literature. While the scope of these databases does not exclude NPS, other databases that focus on NPS and NPS related compounds also exist such as the RESPONSE project (https://www.policija.si/apps/nfl_response_web/seznam.php) and NPS Datahub [14]. Unfortunately, these databases are not available in the required formats that can be imported into HRMS vendor data analysis software. With these pitfalls in mind, the online crowd-sourced database, HighResNPS (https://highresnps.forensic.ku.dk/), was developed in 2016 by the current authors which offers the convenience of having the database converted into several major HRMS vendor formats allowing implementation in routine suspect screening [15]. Furthermore, the database is kept up-to-date with the latest NPS using information from a range of different sources, but most notably, those newly reported compounds by the EMCDDA. The use of this database has been reported in literature either as a reference for MS/MS data [8] or as a suspect screening library [16–18]. A disadvantage of using suspect screening databases is that they generally lack retention times specific to the user's liquid chromatography (LC) system and, therefore, have the potential to generate large numbers of false positives using MS data alone, particularly in complex biological matrices.

In order to circumvent this issue, *in silico* retention time prediction using quantitative structure-retention relationship (QSRR) models in combination with molecular descriptors have been developed with some success using artificial neural networks (ANN) [19–23]. However, these models are only able to provide predicted retention times for the LC system included in the model development. In addition, the development of these retention time models generally requires specialist software, which may involve expensive license agreements. Alternatively, models can be developed at no cost using programming languages such as R (https://www.r-project.org/) or Python (https://python.org/) with various machine learning packages; however, this approach requires personnel with experience in machine learning and computer programming which are skills often limited in routine analytical chemistry laboratories.

To overcome this, several tools have been made freely available to the scientific community including QSRR Automator by Naylor et al. [24] which is designed to be a user-friendly program that can create models for inexperienced users. For model transferability, PredRet was developed to “project” or “map” retention times from one LC system to another [25]. More recently, Bouwmeester et al. [26] reported a generalized calibration approach, calibrate all LC (CALLC), an extension of the approach that underpins PredRet, to predict retention times for different LC systems. While these approaches offer some solution to developing models capable of integrating multiple LC systems, they are however, limited in certain aspects. For example, PredRet can only map retention times of analytes to an LC system if they have been measured by at least one LC system in the database, therefore, this approach cannot be used to extrapolate retention times for unobserved analytes. On the other hand, CALLC extends on this approach by using a layered workflow which firstly develops a QSRR-based retention time prediction model for each LC system (layer 1) then performs the calibration using predicted retention times from each system rather than experimental retention times (layer 2). While CALLC is a useful approach for a single user, a retention time prediction model for each LC setup needs to be developed, which is a time-consuming process from a database perspective when considering a large number of LC systems. A review published in 2020 by Witting and Böcker [27], indicated that retention time prediction approaches which integrate multiple LC systems are still lacking but are of huge importance for untargeted metabolomics.

From a machine learning theory perspective, it is possible to simultaneously model retention times from multiple LC systems using a singular model rather than individual models for each LC system. This can be achieved using a fundamental machine learning technique called “one-hot encoding” which encodes categorical labels present in a dataset into numerical or indicator variables, this is especially important for models that can only interpret numerical input values, such as ANNs [28]. In this context, the categorical labels would indicate the origin of the retention time such as a laboratory or dataset name. Therefore, the present study aims to provide a stream-lined approach for retention time model development that can integrate multiple LC systems with different elution conditions through the use of one-hot encoding. The model was developed using entries from the HighResNPS database and aims to encourage the scientific community to contribute and further improve the model and in turn, receive predicted retention times specific to their LC system for all analytes on HighResNPS.

2. Experimental section

2.1. Retention time dataset

A total of 1281 retention times corresponding to 685 analytes were obtained from HighResNPS (Table S1 in the Supplementary Information). These retention times were selected from the nine main contributors, including the author's laboratory (CPH) and laboratories A to H, where the number of entries were greater than or equal to 50 (as of 12th June 2020). Details of LC conditions for each laboratory are provided in Table S2. The International Union of Pure and Applied Chemistry (IUPAC) name for each analyte was also obtained from HighResNPS and parsed through the Open Parser for Systematic IUPAC Nomenclature (OPSIN, https://opsin.ch.cam.ac.uk/) [29] to generate simplified molecular-input line-entry system (SMILES) strings. HighResNPS is currently a closed user group, however, access to the database can be granted by contacting the corresponding author.

2.2. Molecular descriptors

JChem for Excel was used to generate molecular descriptors for each SMILES string (JChem for Excel 20.11.0.644, ChemAxon, https://www.chemaxon.com). The descriptors used in this study were the logarithm of the distribution between octanol and water (logD), the logarithm of the partition coefficient between octanol and water (logP) and the number of carbon (nC) and oxygen atoms (nO). In this study, the logP values are calculated using a Consensus model which is developed using the ChemAxon and Klopman et al. [30] methods in combination with the PhysProp database. The ChemAxon method is derived from the Viswanadhan et al. [31] method with slight modifications. The Consensus and ChemAxon logP methods are similar to ClogP and AlogP methods, respectively [32]. These descriptors were selected based on their relative importance for retention time modelling outlined by Barron and McEneff [23].

2.3. Data analysis and model development

All data analysis and machine learning were performed using the Python programming language (Python 3.7.6) in the Scientific Python Development Environment (Spyder 4.0.1, https://spyder-ide.org/). Retention time modelling was performed using the TensorFlow (2.1.0, https://tensorflow.org/) [33] and Keras (1.0.8, https://keras.io/) [34] machine learning packages. Data analysis was performed using the scikit-learn (0.22.1, https://scikit-learn.org/) [34], pandas (1.0.1, https://pandas.pydata.org/) [35] and NumPy (1.18.1, https://numpy.org/) [36] packages with data visualization performed using the Matplotlib (3.1.3, https://matplotlib.org/) [37] and seaborn (0.10.0, https://seaborn.pydata.org/) [38] packages.

2.4. Data preprocessing

To account for the different LC conditions used by each laboratory and, therefore, different retention times observed for the same analyte across the dataset, the laboratory names ($n = 9$) were one-hot encoded as descriptors (variables) using the pandas.get_dummies function [39]. This takes n categories in a single variable and converts them to n-1 indicator variables which have a binary value (0 or 1). The same process was also applied for each drug class ($n = 13$) present in the dataset. The final dataset after preprocessing contained 24 input variables (four molecular descriptors, eight laboratory and 12 drug class variables). The dataset was then randomly split into training, optimization, validation and test sets at an approximate ratio of 55:15:15:15 (707:190:191:193). Finally, the variable values were centered and scaled using scikit-learn's StandardScaler [40] which subtracts the mean of the values in a variable in the training set and then divides it by the standard deviation of those values. The mean and standard deviation from the training set is then used to center and scale the optimization, validation and test sets. The final input data for the train and test sets are provided in Table S3 and Table S4 for pre- and post-standard scaling, respectively.

2.5. Architecture selection and evaluation of the singular model

Modelling of retention times was performed using a feedforward multilayer perceptron (MLP) with a gradient-based Adam optimizer and rectified linear unit (ReLU) activation function. During the model development using the training set, the mean absolute error (MAE) for the optimization set predictions for each training cycle (epoch) was evaluated. Training was limited to 500 epochs with an early-stopping mechanism implemented to prevent overfitting by ceasing training 50 epochs after the minimum MAE for the optimization set was recorded.

The optimum MLP architecture, i.e. the number of hidden layers and the number of neurons in each hidden layer was determined using an in-house application. This application created a set of architectures based on user-defined limits for the number of layers and neurons in each layer. For each architecture, the model was trained (as outlined above) and predictions were made using the validation set. This process was performed in replicate ($n = 5$) and each architecture was evaluated using the averaged MAE and coefficient of determination (R^2) values of the validation set predicted retention times (t_R^P). The optimum architecture was defined as the one that gave the lowest averaged MAE of the validation set. For this study, a total of 110 architectures were evaluated including both one- and two-layer MLPs and with up to 100 neurons in each hidden layer. The external held-out test set was then used to evaluate the performance of the model.

2.6. CALLC

The developed model was benchmarked against the freely available and user-friendly CALLC graphical user interface (https://github.com/RobbinBouwmeester/CALLC). CALLC is a layered approach which firstly trains five different QSRR-based models: Extreme Gradient Boost (XGBoost), Support Vector Regressor (SVR), Least Absolute Shrinkage and Selection Operator (LASSO), Adaptive Boosting (AdaBoost) and Bayesian Ridge Regression (BRR). Predictions are then made for the test set using each of the models and the predictions are then calibrated in the second layer using a generalized additive model (GAM) and finally linearly combined into a single retention time per analyte. CALLC was applied to each laboratory using the same train and test set analytes as used in the singular model.

3. Results and discussion

The dataset used for this study contained 1281 retention times that corresponded to 685 analytes comprising mostly NPS and their metabolites with a small number of naturally derived substances and common pharmaceuticals. The most represented classes in the dataset were cannabinoids ($n = 411$), cathinones ($n = 191$), phenethylamines ($n = 177$) and opioids ($n = 175$). It should be noted that the class for each entry is selected by the contributor, using a pre-defined list of classes provided on HighResNPS. Therefore, the class designation for each entry is the one that the contributor determines to best describe the analyte and may not be the true class designation according to NPS classification guidelines. Entries designated as "unknown" do not fall under any of the other classes. Fig. S1(a) shows the distribution of drug classes for each laboratory while Fig. S1(b) shows the distribution of drug classes for all and unique entries.

The number of entries provided by each laboratory ranged from 51 (G) to 282 (H) with retention time windows (i.e. the difference between the first and last eluting analyte) ranging from 2.78 (D) to 12.48 min (H). The total retention time window of the dataset was 13.27 min with the minimum and maximum retention times of 0.95 and 14.22 min corresponding to benzylpiperazine (Laboratory B) and EG-018 (Laboratory H), respectively. Retention times were acquired using seven different LC systems all utilizing reverse-phase C_{18} columns with a variety of specifications (i.e. length, internal diameter and particle size) and maintained at different temperatures (30–50 °C). Flow rates varied from 0.2 to 0.9 mL/min, with laboratory D having a flow rate increase from 0.7 to 0.9 mL/

min during the LC run. Mobile phases were generally comprised of aqueous ammonium formate buffers or 0.1% formic acid with acetonitrile or methanol organic phases containing formic acid (0.01% or 0.1%). Initial mobile phase composition ranged from 0 to 13% organic phase with various gradients employed between the seven different methods.

3.1. Architecture selection and test set evaluation of the singular model

The evaluation of the 110 different MLP architectures indicated that the dataset showed robust correlations between the selected descriptors and retention times. This was supported by Pearson correlation coefficients (i.e. Pearson's R) of 0.902 0.785, 0.779 and 0.125 for the correlations of retention time with logD, logP, nC and nO, respectively. For all architectures, the MAE values for the averaged t_R^P of the validation set were between 0.605 and 0.699 min with R^2 values between 0.900 and 0.924 (Table S5). From these results, the architecture that yielded the lowest MAE value for the averaged predictions (0.605 min, $R^2 = 0.924$) was a two-layer MLP with 40 and 50 neurons in the first and second hidden layer, respectively. Therefore, the results presented herein are based on the averaged t_R^P of the test set from the 24:40:50:1 MLP architecture. Fig. 1 illustrates the correlations of the experimental retention time (t_R^E) versus t_R^P for all predictions and each laboratory. The MAE and the R^2 values for the averaged test set t_R^P were 0.583 min and 0.942, respectively, with seven of the nine laboratories having R^2 values > 0.850 while Laboratory C and D had R^2 values of 0.696 and 0.127, respectively. It is important to note here that while these laboratories had lower R^2 values, the MAE values (0.348 and 0.375 min, respectively) were lower than the MAE values of the other laboratories. The lower R^2 values can be due to an uneven distribution of the test set analytes across the retention time range for the laboratory (Laboratory C). Additionally, outlier predictions can lower the R^2 values considerably when the test set contains a small number of analytes (both Laboratory C and D have seven analytes). This indicates that R^2 alone should not be used as a means to evaluate the predictive ability for different laboratories.

For the averaged t_R^P of the test set predictions for all laboratories, 54% of analytes ($n = 105$) were predicted within ±0.5 min of the t_R^E with 82% ($n = 158$) and 97% ($n = 188$) predicted within ±1 and ±2 min, respectively (Table S6). Only five analytes had t_R^P outside ±2 min of the t_R^E, with maximum negative and positive errors of −2.06 and 2.85 min, respectively. The outlier analytes included two synthetic cannabinoids, two benzodiazepines and a piperazine derivate with the former two classes demonstrating the largest error distributions (Fig. S2). For the predictions of each laboratory, the median errors ranged from −0.13 to 0.46 min with most laboratories having >50% and >75% of the test set t_R^P within ±0.5 and ±1 min of the t_R^E, respectively.

3.2. Singular model performance on identical analytes from different LC systems

The ability for the singular model to accurately predict retention times for the same analyte in different LC systems was also investigated. The synthetic cannabinoid, FUB-AMB, had the most entries in the dataset with seven laboratories providing retention times with a range of 7.51–11.68 min. These entries were removed from the dataset ($n = 1274$) and the model was retrained in replicate ($n = 5$) and evaluated as detailed above (MAE = 0.620 min, $R^2 = 0.917$). For each replicate, the retention times of FUB-AMB for each laboratory were then predicted and are summarized in Table 2. The mean errors for the averaged t_R^P were within ±0.73 min of the t_R^E for all laboratories with five of the seven laboratories within ±0.25 min. The t_R^P across the replicates for each laboratory were generally consistent with the minimum and maximum t_R^P ranging no more than 0.70 min. This is a significant result as it demonstrates the power of one-hot encoding to effectively allow the model to differentiate between LC systems and accurately predict retention times for the same analyte. It is also important to highlight here that these predictions were made for analytes not previously observed by the model demonstrating the ability to make predictions from a completely new set of molecular descriptors which contrasts to the previously mentioned mapping techniques.

3.3. Singular (combined) vs. individual model performance

To ascertain whether using a singular model improved or worsened performance compared to a model for each laboratory, individual models for each laboratory were trained and evaluated as previously detailed. The training, optimization and test sets for each laboratory consisted of the same analytes that were used to develop and evaluate the singular model, however, in this case the models were trained using only 16 variables (four molecular descriptors and 12 drug class variables). The same architecture was used for each laboratory as the singular model, therefore, the validation set was not used in this case. To compare the overall performance of the individual models with the singular model, R^2 and MAE values were determined using the test set t_R^P from all individual models (Table S7). In general, the individual models showed an overall poorer performance when compared to the singular model with an R^2 value of 0.896 and MAE value of 0.664 (Table 1). The number of analytes predicted within ±0.5 min was slightly greater in the individual models, however, there were 13 analytes with errors greater than ±2 min with maximum negative and positive errors of −3.69 and 5.38 min, respectively. More noticeably, however, were the difference in the R^2 and MAE values between the singular and individual models for laboratories with lower number of entries. For example, laboratories D and F showed no correlation between t_R^E and t_R^P ($R^2 = 0$) while laboratories F and G showed a 2.7 and 4.4 times increase in MAE values, respectively. Laboratories with higher numbers of entries such as CPH, A and H saw improved model performance based on decreases in MAE values, likely due to the increased number of training observations, while laboratory E showed similar performance compared to the singular model with laboratory B having a slight improvement in performance with the singular model. The overall improved performance of the singular model is also reflected in the comparison of error distributions shown in Fig. 2. It is important to consider; however, that while the singular model generally improved overall performance, this comparison was performed using an MLP architecture that was not optimized for each laboratory. Therefore, it may be possible for an individual model to provide improved results for those cases where performance was worse than the singular model, if an optimized architecture is determined for each laboratory; however, this would require *n* optimized models for *n* laboratories. Even though only nine laboratories were included in this study, the number of NPS and the number laboratories contributing their data to HighResNPS is constantly increasing. Therefore, a stream-lined model which can be trained once for *n* laboratories is preferred requiring less maintenance and computational time.

3.4. The use of one-hot encoding and indicator variables

This present study exploited the categorical information present for each entry on HighResNPS by encoding it as binary indicator variables using one-hot encoding in addition to four calculated

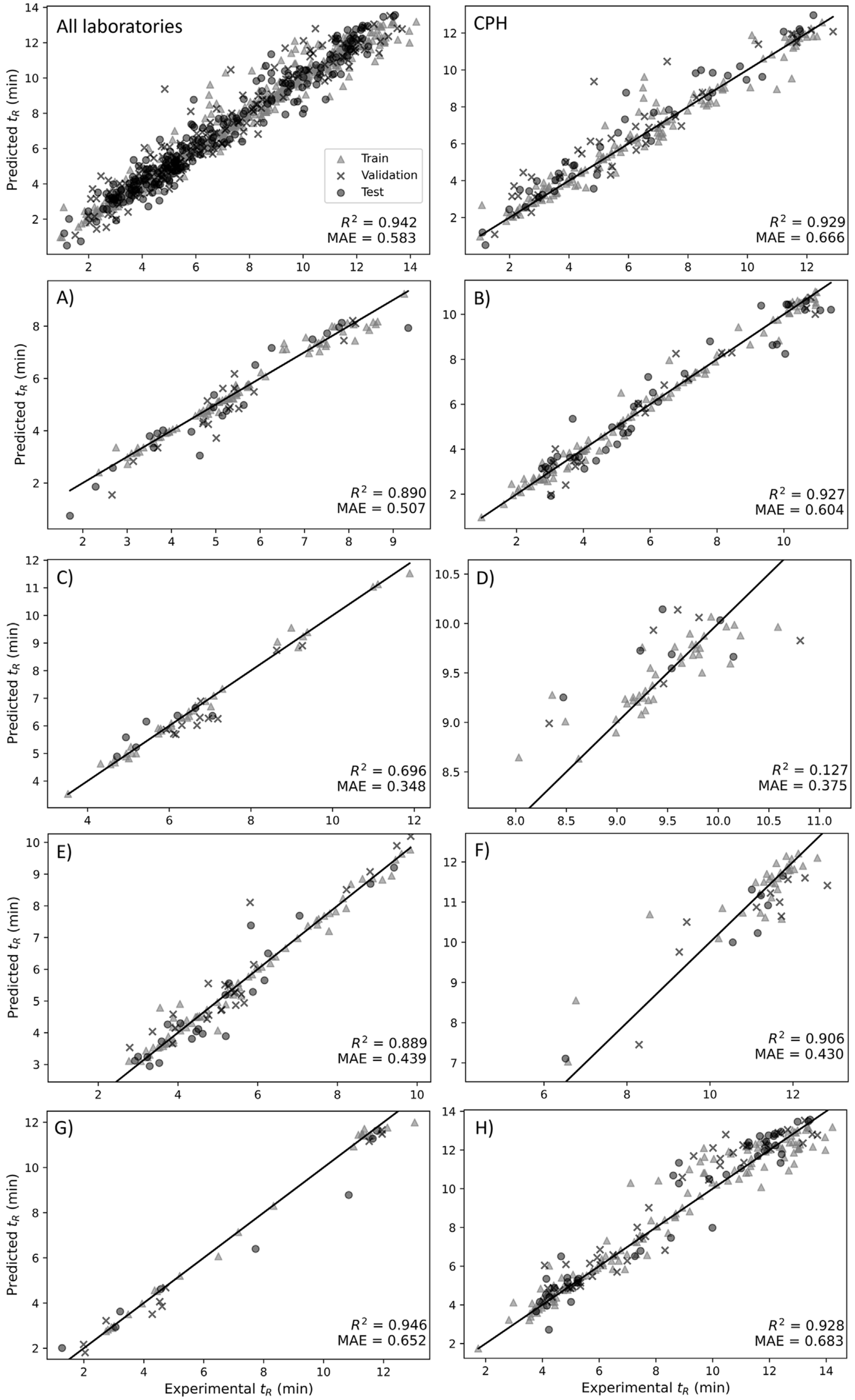


Fig. 1. t_r^E versus t_r^P plots for the train (grey circle), optimization (cross) and test (black circles) set averaged predictions ($n = 5$) for all predictions (top left) and individual laboratories with the respective R^2 and MAE values for the test set predictions.

Table 1
Summary of t_R^P performance for averaged predictions (n = 5) of the test set for singular and individual models.

Laboratory	Entries	t_R range (min)	Train	Optimization	Validation	Test	Model[a]	Model metrics, $n = 5$		Error (min), $n = 5$					t_R^P of test set analytes within t_R^E window (% of test set)			$t_R^P > 2$ min or $t_R^P < -2$ min from t_R^E (% of test set)
								R^2	MAE (min)	Mean	SD	Negative Max	Median	Positive Max	±0.5 min	±1 min	±2 min	N
ALL	1281	0.95–14.22	707	190	191	193	S	0.942	0.583	0.10	0.77	−2.06	0.13	2.85	105 (54)	158 (82)	188 (97)	5 (3)
							I[b]	0.896	0.664	−0.03	1.03	−3.69	0.00	5.38	115 (60)	155 (80)	180 (93)	13 (7)
CPH	263	1.01–12.88	147	34	40	42	S	0.929	0.666	0.43	0.75	−1.27	0.46	2.85	19 (45)	33 (79)	41 (98)	1 (2)
							I	0.929	0.583	0.18	0.84	−1.84	0.13	2.91	27 (64)	36 (86)	40 (95)	2 (5)
A	155	1.71–9.34	89	27	18	21	S	0.890	0.507	−0.16	0.64	−1.59	−0.05	0.90	14 (67)	19 (90)	21 (100)	
							I	0.840	0.478	−0.04	0.80	−1.41	−0.08	2.58	15 (71)	17 (81)	20 (95)	1 (5)
B	202	0.95–11.41	113	32	20	37	S	0.927	0.604	−0.10	0.76	−1.80	−0.08	1.67	21 (57)	28 (76)	37 (100)	
							I	0.892	0.673	−0.22	0.91	−2.39	−0.11	2.02	20 (54)	28 (76)	34 (92)	3 (8)
C	67	3.52–11.89	37	12	11	7	S	0.696	0.348	0.15	0.47	−0.70	0.17	0.71	4 (57)	7 (100)		
							I	0.725	0.384	−0.03	0.47	−0.70	0.00	0.53	5 (71)	7 (100)		
D	63	8.03. 10.81	40	10	6	7	S	0.127	0.375	0.23	0.45	−0.49	0.15	0.78	5 (71)	7 (100)		
							I	0.000	0.490	−0.21	0.57	−0.96	−0.45	0.71	4 (57)	7 (100)		
E	137	1.24–9.84	71	19	25	22	S	0.889	0.439	−0.07	0.58	−1.31	−0.07	1.55	14 (64)	20 (91)	22 (100)	
							I	0.882	0.453	−0.11	0.59	−1.37	−0.06	1.14	15 (68)	19 (86)	22 (100)	
F	61	4.61–12.83	32	10	12	7	S	0.906	0.430	−0.18	0.52	−0.92	−0.11	0.59	4 (57)	7 (100)		
							I	0.000	1.872	0.03	2.69	−2.44	0.28	5.38	2 (29)	3 (43)	4 (57)	3 (43)
G	51	1.27–13.04	25	8	10	8	S	0.946	0.652	−0.35	0.92	−2.06	−0.13	0.74	5 (63)	6 (75)	7 (87)	1 (13)
							I	0.666	1.784	0.43	2.43	−3.69	0.57	4.54	1 (13)	3 (38)	6 (75)	2 (25)
H	282	1.74–14.22	153	38	49	42	S	0.928	0.683	0.25	0.89	−2.00	0.24	2.53	19 (45)	31 (73)	39 (93)	3 (7)
							I	0.933	0.600	−0.10	0.89	−2.88	−0.05	2.85	26 (62)	35 (83)	40 (95)	2 (5)

[a] S = singular, I = individual.
[b] Combined predictions from each individual model.

Table 2
Prediction summary for FUB-AMB for each laboratory using a 24:40:50:1 MLP architecture.

Laboratory	t_R^E (min)	t_R^P per replicate (min)					Mean t_R^P (min)	SD (min)	Mean error (min)
		1	2	3	4	5			
CPH	11.68	10.85	11.21	10.89	10.84	11.08	10.95	0.17	−0.73
A	7.51	7.40	7.23	7.09	7.31	7.69	7.26	0.13	−0.25
B	10.22	9.96	10.33	10.02	10.17	10.17	10.12	0.17	−0.10
D	9.35	9.38	9.61	9.19	9.21	8.95	9.35	0.19	0.00
E	8.89	8.64	8.90	8.77	8.91	8.82	8.80	0.13	−0.09
F	11.66	12.19	12.43	12.73	12.08	12.51	12.36	0.29	0.70
H	10.77	10.81	10.59	10.69	10.79	10.82	10.72	0.10	−0.05

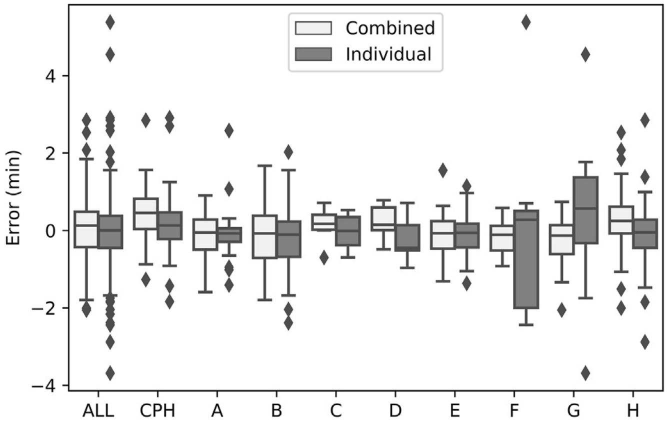


Fig. 2. Box and whisker plots for the errors of the averaged test set t_R^P (n = 5) for the combined (white) and individual (grey) models. Boxes represent the 25th and 75th percentile, whiskers represent 1.5 times the interquartile range (IQR), the solid lines within in the box and diamonds represent the median and outliers, respectively.

molecular descriptors. However, overall modelling performance was also evaluated using molecular descriptors only and the descriptors with laboratories encoded as binary variables. Unsurprisingly, the performance of the model using only the calculated descriptors was considerably poorer ($R^2 = 0.867$, MAE = 0.878 min, Table S8) compared to the performance of a model that used all categorical data ($R^2 = 0.942$, MAE = 0.583 min). This is due to different experimental retention times for identical analytes (i.e. identical set of descriptors), however, when the laboratories are encoded there is a marked improvement in performance ($R^2 = 0.919$, MAE = 0.683, Table S9) albeit poorer than when the drug classes were also encoded. The improved performance for the model that incorporated drug classes could be possibly explained by class-specific properties that influence retention time which are not explained by the selected molecular descriptors. This is evidenced by the imperfect correlation of logD with retention time (Pearson's R = 0.902). In addition, there are a broad set of chemistries represented in the dataset which can range from simple monocyclic phenethylamines to the more complex polycyclic synthetic cannabinoids and benzodiazepines. The authors do acknowledge, however, that due to the subjective selection of drug classes for each entry by contributors, there are likely to exist entries which are classified incorrectly and could affect their predictions based on the phenomena described above. It should also be noted that there were minimal interventions made on the dataset prior to modelling which only included ensuring that identical analytes had the same class designations. This was to attempt to simulate an automated process whereby modelling and predictions of the database are performed periodically with updated datasets with minimal human intervention. Alternative methods for objective and automated classification are currently being investigated as part of the future directions of HighResNPS. These include freely available tools such as ClassyFire [41] and the use of SMILES arbitrary target specification (SMARTS) substructure queries [42], therefore, removing the need for contributors to classify each entry.

The main advantages of the use of one-hot encoding over other reported techniques include the ability for a model to be developed using greater numbers of analytes since datasets from different LC systems can be combined. This combination of datasets also appears to improve model performance for datasets where there is a smaller number of analytes as previously observed. The major advantage, however, is the model's ability to make predictions simultaneously for each LC system included in the modelling process. The importance of this aspect cannot be understated from a database perspective, as it significantly reduces the time taken to make predictions for when there are many laboratories or LC systems. The major limitation of this approach, however, is that the model does not extract meaningful information about structural classes or LC system parameters and only uses the labels (names) to differentiate between the categories it has previously observed. Therefore, the current model is unable to generalize predictions to a new LC system or drug class. This would require the LC system parameters or structural information to be encoded into the model input. The LC system parameters may include those that generally influence retention time such as, but not limited to, column stationary phase, column specifications, flow rate, gradient and mobile phase composition. However, there still exists the issue that many of these parameters take in labels (as opposed to a continuous values). For example, a mobile phase composed of 10 mM ammonium acetate and acetonitrile would need to be encoded into a set of numerical descriptors that adequately represent this composition. With respect to structural information, the molecular structure could be directly encoded as fingerprint vector [43]. Additionally, the use of binary variables restricts the model from making predictions when new categories are present in the dataset. Consequently, there exists a trade-off between model performance and generalization capabilities when considering the use of indicator variables, however in this study, increased model performance was favored over the capacity to generalize since the NPS chemical space is considerably narrow with the majority of analogues falling under the classes that the model was trained on. Of course, the model can be adjusted by removing the indicator variables for the drug classes so that predictions can be made for analytes outside of the NPS chemical space based on descriptor values alone. In order to incorporate new laboratories or classes, the model would then need to be retrained with these new categories encoded as indicator variables, however, this issue is largely nontrivial since the modelling process can be performed in a matter of seconds. There is also the possibility that retraining a new model with these new categories, may improve the performance for existing categories.

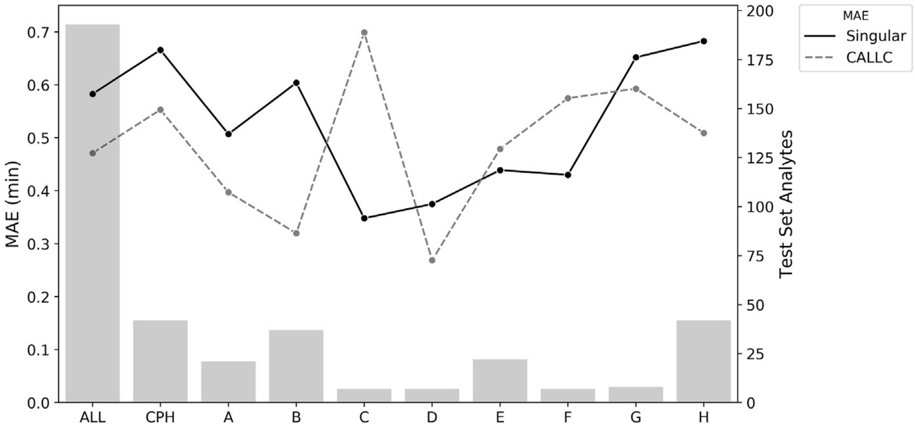


Fig. 3. The MAE value comparison between the singular (solid line) and CALLC (dashed) models for the overall results (ALL) and the individual laboratories. The number of analytes in the test set are denoted by the bar graph.

3.5. Comparison to existing methods

It is important to distinguish the work presented here from the existing techniques such as PredRet and CALLC. As previously mentioned, PredRet uses the t_R^E of several analytes from different LC systems to create projection models between the t_R^E for those LC systems. These projections can be used to predict retention times of other analytes if they have been measured in another system. Therefore, retention times cannot be predicted for analytes that do not have an t_R^E, this contrasts to the present model which can predict the retention time for any analyte (provided a logD and logP can be calculated) specific to the LC systems the model is trained on. As indicated previously, HighResNPS is kept up to date through the addition of novel analytes which have been reported by different sources. In some cases, these analytes may not have analytical data available and only a molecular formula and structure can be provided, therefore, there exists a number of analytes where no t_R^E observed on any LC system. Considering this, PredRet is insufficient to warrant further evaluation for the purposes of this study. On the other hand, CALLC aims to develop a retention time prediction model for a LC system of interest through calibration of t_R^P for a set of analytes using different LC systems. While similar to PredRet, it differs since it firstly trains a QSRR-based retention time model for each dataset and then performs the calibration using t_R^P rather than t_R^E. Therefore, this allows for the mapping of retention times where t_R^E may not be available; however, it is important to note that CALLC still trains individual models for each LC system and can only make predictions for one LC system at a time. The present model allows for a singular model to be trained on multiple LC systems which can simultaneously make predictions for new observations for the LC systems involved in the training process.

The performance of the singular model was benchmarked against the freely available CALLC (Tables S–10). Overall, the performance of CALLC was superior to that of the singular model with R^2 and MAE values of 0.958 and 0.471 min, respectively, corresponding to a decrease of 0.112 min (19.2%) in MAE. This increase in performance was also observed (Fig. 3) for six of the nine laboratories (A, B, CPH, D, G and H) while laboratories C, E and F demonstrated poorer performance with laboratory C having almost double the MAE when using CALLC (0.699 min). Whilst there was an overall increase in performance for CALLC, this increase in performance is outweighed by the requirement of having to repeat the CALLC process for each laboratory.

3.6. Implications for HighResNPS and retention time prediction

The development of a singular model and the ability to simultaneously predict retention times for any analyte specific to the LC systems the model is trained on has important implications for the future directions of HighResNPS. The model can be retrained on a periodical basis, therefore, allowing the inclusion of new LC systems and analytes. These contributors will then receive predicted retention times for the entire HighResNPS database ($n = 1803$ as of 12th June 2020) in the relevant LC-MS library format. The addition of this feature will hopefully enhance the suspect screening capabilities of the database and encourage more contributions from the scientific community. This in turn will improve the overall quality of the database, particularly through the addition of missing product ion data and newly reported analogues.

The authors do emphasize, however, that while satisfactory modelling can be achieved with as few as 50 analytes for a particular LC system, there is the possibility that the model may not generalize well for all unique HighResNPS analytes on that system. This is considered dependent on the number of entries provided and the diversity of drug classes to which they belong. The ideal application of the predicted retention times should be limited to ranking tentative identifications, with little to no emphasis put on small retention time differences, when candidates otherwise were equally likely. The use of predicted values in this manner, where all candidates would be inspected and evaluated by an expert is too time consuming for any daily application of suspect screening applied to complex matrices, such as human whole-blood in a forensic setting. In this setting the number of false-positive identifications will typically be overwhelming. The less ideal application of predicted retention times is the situation with an absolute error threshold, which will decrease the false-positive rate and increase the false-negative rate. The threshold therefore becomes a tool for balancing the false-positive and false-negative rates. Regardless, any application should allow for identifications which would have been judged likely without the predicted retention times, i.e. with several matched expected fragments, regardless of a predicted large retention time error, as there will be outliers, and these should not be overlooked simply because predicted retention times were employed.

This is the first report of a singular QSRR-based retention time model that integrates multiple LC systems and, from a broader perspective, has important implications for retention time prediction in suspect and non-targeted screening. Lastly, while the method presented here has a strong focus on NPS, this is not to say that the same modelling procedure can't be translated to environmental, metabolomic and lipidomic contexts where similar databases exist, facilitating multi-institutional collaborations between researchers with mutual analytical targets.

4. Conclusion

In the last decade there has been an increase in the use of *in silico* approaches to predict retention times for suspect and non-targeted screening; however, models integrating multiple LC systems that can predict retention times from new observations has been lacking. Here for the first time, we present a stream-lined approach for the development of a retention time prediction model capable of integrating multiple LC systems using one-hot encoding. This study emphasized the ability of the developed model to predict retention times for all unique analytes in the HighResNPS database specific to different LC systems with a focus on the translation of predicted retention times to suspect screening workflows. Finally, prediction of retention times for all unique analytes in the HighResNPS database specific to contributors LC systems will be made available.

CRediT authorship contribution statement

Daniel Pasin: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Christian Brinch Mollerup:** Formal analysis, Writing – original draft, Writing – review & editing, Visualization. **Brian Schou Rasmussen:** Formal analysis, Writing – original draft, Writing – review & editing, Visualization. **Kristian Linnet:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Petur Weihe Dalsgaard:** Conceptualization, Data curation, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to sincerely thank all the contributors of HighResNPS, without whom this study would not be possible.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.aca.2021.339035.

References

[1] European Monitoring Centre for Drugs and Drug Addiction, European Drug Report 2019: Trends and Developments, 2019. Accessed in 07-Jul-20, https://www.emcdda.europa.eu/publications/edr/trends-developments/2019.

[2] United Nations Office of Drugs and Crime, Current NPS threats: Volume II, Accessed in 07-Jul-20, https://www.unodc.org/documents/scientific/Current_NPS_Threats_Volume_II_Web.pdf, 2020.

[3] D. Pasin, A. Cawley, S. Bidny, S. Fu, Current applications of high-resolution mass spectrometry for the analysis of new psychoactive substances: a critical review, Anal. Bioanal. Chem. 409 (2017) 5821–5836.

[4] J.A. Baz-Lomba, M.J. Reid, K.V. Thomas, Target and suspect screening of psychoactive substances in sewage-based samples by UHPLC-QTOF, Anal. Chim. Acta 914 (2016) 81–90.

[5] R. Sutherland, S. Allsop, A. Peacock, New psychoactive substances in Australia: patterns and characteristics of use, adverse effects, and interventions to reduce harm, Curr. Opin. Psychiatr. 33 (2020) 343–351.

[6] M. Ibáñez, J.V. Sancho, L. Bijlsma, A.L.N. van Nuijs, A. Covaci, F. Hernández, Comprehensive analytical strategies based on high-resolution time-of-flight mass spectrometry to identify new psychoactive substances, Trac. Trends Anal. Chem. 57 (2014) 107–117.

[7] J.M. Colby, K.L. Thoren, K.L. Lynch, Suspect screening using LC–QqTOF is a useful tool for detecting drugs in biological samples, J. Anal. Toxicol. 42 (2018) 207–213.

[8] R. Bade, J.M. White, L. Nguyen, B.J. Tscharke, J.F. Mueller, J.W. O'Brien, K.V. Thomas, C. Gerber, Determining changes in new psychoactive substance use in Australia by wastewater analysis, Sci. Total Environ. (2020) 731.

[9] P. Vervliet, O. Mortelé, C. Gys, M. Degreef, K. Lanckmans, K. Maudens, A. Covaci, A.L.N. van Nuijs, F.Y. Lai, Suspect and non-target screening workflows to investigate the in vitro and in vivo metabolism of the synthetic cannabinoid 5Cl-THJ-018, Drug Test. Anal. 11 (2019) 479–491.

[10] N. Salgueiro-González, S. Castiglioni, E. Gracia-Lor, L. Bijlsma, A. Celma, R. Bagnati, F. Hernández, E. Zuccato, Flexible high resolution-mass spectrometry approach for screening new psychoactive substances in urban wastewater, Sci. Total Environ. 689 (2019) 679–690.

[11] C.B. Mollerup, B.S. Rasmussen, S.S. Johansen, M. Mardal, K. Linnet, P.W. Dalsgaard, Retrospective analysis for valproate screening targets with liquid chromatography–high resolution mass spectrometry with positive electrospray ionization: an omics-based approach, Drug Test. Anal. 11 (2019) 730–738.

[12] D.S. Wishart, Y.D. Feunang, A. Marcu, A.C. Guo, K. Liang, R. Vázquez-Fresno, T. Sajed, D. Johnson, C. Li, N. Karu, Z. Sayeeda, E. Lo, N. Assempour, M. Berjanskii, S. Singhal, D. Arndt, Y. Liang, H. Badran, J. Grant, A. Serra-Cayuela, Y. Liu, R. Mandal, V. Neveu, A. Pon, C. Knox, M. Wilson, C. Manach, A. Scalbert, Hmdb 4.0: the human metabolome database for 2018, Nucleic Acids Res. 46 (2018). D608-d17.

[13] H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M.Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito, T. Nishioka MassBank, A public repository for sharing mass spectral data for life sciences, J. Mass Spectrom. 45 (2010) 703–714.

[14] A. Urbas, T. Schoenberger, C. Corbett, K. Lippa, F. Rudolphi, W. Robien, NPS Data Hub: a web-based community driven analytical data repository for new psychoactive substances, Forensic Chem. 9 (2018) 76–81.

[15] M. Mardal, M.F. Andreasen, C.B. Mollerup, P. Stockham, R. Telving, N.S. Thomaidis, K.S. Diamanti, K. Linnet, P.W. Dalsgaard, HighResNPS.com: an online crowd-sourced HR-MS database for suspect and non-targeted screening of new psychoactive substances, J. Anal. Toxicol. 43 (2019) 520–527.

[16] M. Cüpper, P.W. Dalsgaard, K. Linnet, Identification of new psychoactive substances in seized material using UHPLC-QTOF-MS and an online mass spectral database, J. Anal. Toxicol. 44 (2020) 1047–1051.

[17] K. Diamanti, R. Aalizadeh, N. Alygizakis, A. Galani, M. Mardal, N.S. Thomaidis, Wide-scope target and suspect screening methodologies to investigate the occurrence of new psychoactive substances in influent wastewater from Athens, Sci. Total Environ. 685 (2019) 1058–1065.

[18] P.O.M. Gundersen, S. Broecker, L. Slørdal, O. Spigset, M. Josefsson, Retrospective screening of synthetic cannabinoids, synthetic opioids and designer benzodiazepines in data files from forensic post mortem samples analysed by UHPLC-QTOF-MS from 2014 to 2018, Forensic Sci. Int. (2020) 311.

[19] T.H. Miller, A. Musenga, D.A. Cowan, L.P. Barron, Prediction of chromatographic retention time in high-resolution anti-doping screening data using artificial neural networks, Anal. Chem. 85 (2013) 10330–10337.

[20] 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.

[21] R. Aalizadeh, N.S. Thomaidis, A.A. Bletsou, P. Gago-Ferrero, Quantitative structure–retention relationship models to Support nontarget high-resolution mass spectrometric screening of emerging contaminants in environmental samples, J. Chem. Inf. Model. 56 (2016) 1384–1398.

[22] 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.

[23] 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.

[24] B.C. Naylor, J.L. Catrow, J.A. Maschek, J.E. Cox, Q.S.R.R. Automator, A tool for automating retention time prediction in lipidomics and metabolomics, Metabolites 10 (2020) 237.

[25] J. Stanstrup, S. Neumann, U. Vrhovšek. PredRet: prediction of retention time by direct mapping between multiple chromatographic systems, Anal. Chem. 87 (2015) 9421–9428.

[26] R. Bouwmeester, L. Martens, S. Degroeve, Generalized calibration across liquid chromatography setups for generic prediction of small-molecule retention

times, Anal. Chem. 92 (2020) 6571–6578.

[27] M. Witting, S. Böcker, Current status of retention time prediction in metabolite identification, J. Separ. Sci. 43 (2020) 1746–1754.

[28] K. Potdar, T.S. Pardawala, C.D. Pai, A comparative study of categorical variable encoding techniques for neural network classifiers, Int. J. Comput. Appl. 175 (2017) 7–9.

[29] D.M. Lowe, P.T. Corbett, P. Murray-Rust, R.C. Glen, Chemical name to structure: OPSIN, an open source solution, J. Chem. Inf. Model. 51 (2011) 739–753.

[30] G. Klopman, J.-Y. Li, S. Wang, M. Dimayuga, Computer automated log P calculations based on an extended group contribution approach, J. Chem. Inf. Comput. Sci. 34 (1994) 752–781.

[31] 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 naturally occurring nucleoside antibiotics, J. Chem. Inf. Comput. Sci. 29 (1989) 163–172.

[32] ChemAxon, logP plugin, 2020. Accessed in 20-Nov-20, https://docs.chemaxon.com/display/docs/logp-plugin.

[33] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D.G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, X. Zheng, editors. TensorFlow: a system for large-scale machine learning, in: Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation, USENIX Association, Savannah, GA, USA, 2016.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[35] W. McKinney, editor Data structures for statistical computing in Python. Proceedings of the 9th Python in Science Conference 2010; Austin, TX, USA: SciPy.

[36] C.R. Harris, K.J. Millman, S.J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N.J. Smith, R. Kern, M. Picus, S. Hoyer, M.H. van Kerkwijk, M. Brett, A. Haldane, J.F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T.E. Oliphant, Array programming with NumPy, Nature 585 (2020) 357–362.

[37] J.D. Hunter Matplotlib, A 2D graphics environment, Comput. Sci. Eng. 9 (2007) 90–95.

[38] M. Waskom, seaborn: statistical data visualization, J. Open Source Softw. 60 (2021) 1–4.

[39] pandas, pandas.get_dummies, 2020. Accessed in 18-Jun-20, https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.get_dummies.html.

[40] scikit-learn, 2020. StandardScaler. Accessed in 18-Jun-20, https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.

[41] Y. Djoumbou Feunang, R. Eisner, C. Knox, L. Chepelev, J. Hastings, G. Owen, E. Fahy, C. Steinbeck, S. Subramanian, E. Bolton, R. Greiner, D.S. Wishart ClassyFire, Automated chemical classification with a comprehensive, computable taxonomy, J. Cheminf. 8 (2016) 61.

[42] Daylight Chemical Information Systems Inc, 4. Smarts - a Language for Describing Molecular Patterns, 2019. Accessed in 19-Aug-20, https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.

[43] A. Capecchi, D. Probst, J.-L. Reymond, One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome, J. Cheminf. 12 (2020) 43.