Heliyon 9 (2023) e16671

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Predicting the retention time of Synthetic Cannabinoids using a combinatorial QSAR approach☆

Lina Wu[a,b,1], Fu Xiao[c,d,1], Xiaomin Luo[c,d], Keming Yun[b], Di Wen[e], Jiaman Lin[a,b], Shuo Yang[a], Tianle Li[b], Ping Xiang[a,**], Yan Shi[a,*]

[a] *Academy of Forensic Science, Shanghai Key Laboratory of Forensic Medicine, Shanghai 200063, PR China*
[b] *Shanxi Medical University, Jinzhong 030600, PR China*
[c] *School of Chinese Materia Medica, Nanjing University of Chinese Medicine, Nanjing 210023, PR China*
[d] *Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Science, 555 Zuchongzhi Road, Shanghai 201203, PR China*
[e] *Hebei Medical University, Shijiazhuang 050017, PR China*

ARTICLE INFO

Keywords:
Synthetic cannabinoids
Retention time
Machine learning
QSAR

ABSTRACT

Background: Abuse of Synthetic Cannabinoids (SCs) has become a serious threat to public health. Due to the various structural and chemical group modified by criminals, their detection is a major challenge in forensic toxicological identification. Therefore, rapid and efficient identification of SCs is important for forensic toxicology and drug bans. The prediction of an analyte's retention time in liquid chromatography is an important index for the qualitative analysis of compounds and can provide informatics solutions for the interpretation of chromatographic data.

Methods: In this study, experimental data from high-resolution mass spectrometry (HRMS) are used to construct a regression model for predicting the retention time of SCs using machine learning methods. The prediction ability of the model is improved by adopting a strategy that combines different descriptors in different independent machine-learning methods.

Results: The best model was obtained with a method that combined Substructure Fingerprint Count and Finger printer features and the support vector regression (SVR) method, as it exhibited an R^2 value of 0.81 for the validation set and 0.83 for the test set. In addition, 4 new SCs were predicted by the optimized model, with a prediction error within 3%.

Conclusions: Our study provides a model that can predict the retention time of compounds and it can be used as a filter to reduce false-positive candidates when used in combination with LC-HRMS, especially in the absence of reference standards. This can improve the confidence of identification in non-targeted analysis and the reliability of identifying unknown substances.

☆ **Introduction:** Our study provides a model that can predict the retention time of compounds and when used in combination with HRMS, can be used as a filter to reduce false-positive candidates, especially in the absence of reference standards. This can improve the confidence of identification in non-targeted analysis and the reliability of identifying unknown substances.

* Corresponding author. Academy of Forensic science, Shanghai Key Laboratory of Forensic Medicine, No. 1347 Guangfuxi Road, Shanghai 200063, PR China.

** Corresponding author. Academy of Forensic science, Shanghai Key Laboratory of Forensic Medicine, No. 1347 Guangfuxi Road, Shanghai 200063, PR China.

E-mail addresses: xiangping2630@163.com (P. Xiang), shiy@ssfjd.cn (Y. Shi).

[1] Lina Wu and Fu Xiao are the first authors.

https://doi.org/10.1016/j.heliyon.2023.e16671
Received 1 October 2022; Received in revised form 23 May 2023; Accepted 24 May 2023
Available online 25 May 2023

1. Introduction

Nomenclature

Abbreviation

NPS	New psychoactive substances
SCs	Synthetic Cannabinoids
Rt	retention time
ML	machine learning

New psychoactive substances (NPS) are new compounds with stronger psychoactive properties due to modification of the chemical groups of the parent controlled drugs [1]. During the peak period in 2015, NPS appeared at the rate of at least one new substance per week [2]. SCs are one of the new psychoactive substances and they are the family with the most kinds of substances and the most serious abuse in the NPS. Criminals usually evade the law by modifying their chemical structures, and the new SCs they have produced have similar or even stronger effects than Δ^9-THC [3]. The drug abuse problem due to SCs has caused a major threat to public health and created great social harm [4]. The most effective measure to supervise SCs abuse is to fully understand and master the structural characteristics of these compounds. However, forensic toxicology laboratories are constantly facing analytical challenges when dealing with these substances. The large number of potential compounds to be investigated, the lack of available chemical reference standards, and the changing nature of these substances are the major problem facing forensic toxicology. Therefore, a model that can predict the chromatographic properties of unknown compounds and quickly master their chemical characteristics based on known SCs could be an important complementary tool for forensic toxicological identification.

Liquid chromatography-mass spectrometry is a popular technique for high-throughput analysis of toxic molecules because they are separated according to their physicochemical properties. The retention time (Rt) refers to the interval between the time of injection of the sample into the column and the time at which the peak maximum arrives at the detector. The retention time is determined by the degree of interaction between the analyte and the stationary and mobile phases [5]; therefore, it is another important factor that is independent of mass spectrometry information. This is an important condition for qualitative determination in poison identification, and it can help to understand the chemical properties of compounds undergoing chromatographic separation, thereby providing a more comprehensive understanding of the compounds being analyzed.

Quantitative Structure–Activity Relationship (QSAR) can be used as a tool to establish the correlation between the compound structures and their physical and biological properties and predict the biological activity of compounds. So far, the application of machine learning model is no longer limited to classic problems such as computer vision and image segmentation, but has been widely infiltrated into data analysis in all walks of life. QSAR model Combined with Machine Learning technology has been widely used in the field of screening New Psychoactive Substances [6,7].

A number of models based on machine learning (ML) algorithms have been developed to predict retention time. Some studies have established retention time prediction models for SCs to promote the Identification of New/Unknown Compounds [8,9]. In addition, one study shows that developed models can be used to predict the retention time of all analytes on HighResNPS for each participating laboratory's LC system to further support suspect screening [10]. In view of the scarcity of QSAR studies on the prediction of the retention properties of SCs, and it is important to obtain more chromatographic information about SCs to provide help for the untargeted screening of unknown SCs. In the present study, we designed several regression models based on ML algorithms for predicting the retention time of SCs by using experimental data. The most reliable prediction model was selected to predict the retention time of four new SCs. The proposed model can be used as an effective tool to aid in the prediction of retention times of SCs.

2. Experimental section

2.1. Reagents

All reagents used in the experiment were HPLC grade or better. 232 SCs compounds were purchased from Glpbio (California, USA), Cerilliant (Texas, USA), and Cayman (Michigan, USA). Methanol (MeOH), acetonitrile (ACN), and formic acid (FA) were purchased from Thermo Fisher Scientific (Massachusetts, USA). Ultrapurified water was made in-house with a Barnstead GENPURE PRO water system (Thermo Fisher Scientific; Massachusetts, USA). Neat compounds were prepared as 1 μg/mL working solutions by diluting stock solution ($c = 1$ mg/mL methanol) in methanol (1:1000). Other concentrations of working solutions used in the test were prepared by diluting the stock solutions with methanol. All solutions were stored at −20 °C. In general, the validity period of the standard stock solution under the storage condition of −20 °C is one year and that of the working solution is three months.

2.2. Sample preparation

In total, 232 compounds were prepared as 100 ng/mL standard solutions and were injected into an Orbitrap Exploris 120 LC-HRMS system. We obtained information in the form of compound names and retention times.

2.3. Instrument conditions

The LC-HRMS system was a Thermo Scientific Vanquish Flex UHPLC system equipped with a Thermo Scientific Orbitrap Exploris 120 mass spectrometer (Thermo Fisher Scientific, USA) interfaced with a heated electrospray ion (HESI) source.

2.4. LC-HRMS conditions

The compounds were separated with a Waters Acquity UPLC BEH C18 (100 × 2.1 mm, 1.7 μm) column and an equivalent VanGuard pre-column (2.1 × 5 mm) at 35 °C. The mobile phases were 0.1% formic acid in water (Phase A) and 0.1% formic acid in methanol (Phase B), and the flow rate was 0.4 mL/min. The mobile phase gradient was maintained from the initial 5% B for 1 min, increased to 60% B at 2 min, and then increased from 60% B to 95% B at 12 min. It remained 95% B for 4 min and then restored to the initial mobile phase composition ratio and equilibrium in 2 min. The whole elution time was 20 min, and the gradient rise curve was 7. The mass spectrum was obtained in full-scan positive data dependent analysis (DDA) mode with a mass range of 100–1000 *m/z* and a resolution of 60000 FWHM. The + ESI source parameters were: ion spray voltage: 3500 V; sheath gas (Arb): 40; Aux gas (Arb): 10; sweep gas (Arb): 0; RF Lens (Arb): 70; ion transfer tube temp: 320 °C; and vaporizer temp: 300 °C. The collision energy (CE) was compound dependent, at 20, 25, 35, or 40 eV. The mass tolerance of the parent mass and the fragments compared to the theoretical mass was <5 ppm [11].

2.5. Molecular descriptors

Molecular fingerprint is a kind of descriptor for describing molecular structural bonds in the form of binary strings. In this study, 232 retention times of standards was used to build the prediction model, which was further divided into a training set and a test set. We established the ML model by first calculating the molecular descriptors related to the 232 SCs molecular structure using PaDEL-Descriptor [12,13] software. Our most commonly used descriptors [14–16]included topology, geometry, electrostatics, quantum chemistry, and various physicochemical parameters, which accounted for more than 1444 descriptors (1D and 2D descriptors) and 12 types of fingerprints (total 17536 bits) (see online Supplemental Table). The characterization of these descriptors affects the retention behavior of compounds to varying degrees and are therefore very important for their feature selection.

2.6. Model building

Based on the calculated descriptors and fingerprints, we used SVR [17], Random Forest RF [18], Gradient Boosting Regression (GBR) [19], AdaBoost Regressor (ABR) [20] and eXtreme Gradient Boosting (XGB) [21]to establish the model.

2.6.1. Support vector regression (SVR)

SVR is a nonlinear prediction model based on kernel function. Its learning theory is based on statistical theory, which emphasizes statistical learning in the case of fewer samples. The powerful theoretical foundation of the SVR model provides high generalization ability, avoids overfitting, and can also efficiently process high-dimensional input vectors [22]. This method is very valuable in many practical applications [23–25].

2.6.2. Random Forest (RF)

RF, developed by Breiman and Cutler, is an ensemble leaning method based on decision trees. In machine learning, RF, as a predictor, outputs the prediction of combining the outputs of individual trees and follows specific rules in tree growth, tree composition, self-test, and post-processing [26]. Compared to other ML algorithms, RF is considered more stable in the presence of outliers and in very high dimensional parameter spaces [27].

2.6.3. Gradient Boosting Regression (GBR)

GBR is a supervised ML algorithm that learns from its mistakes. It essentially produces a model in the form of an ensemble of weak learning algorithms. GB uses an iterative gradient technique to minimize the loss function by iteratively selecting a function that points to a negative gradient [28]. It has the characteristics of high precision, high flexibility, and fast execution and is useful for regression [29].

2.6.4. AdaBoost regressor (ABR)

The ABR algorithm was proposed by Freund and Schapire in 1995 [30]. The main idea behind the algorithm is to maintain a distribution or set of weights over the training set, and the most basic theoretical characteristic of AdaBoost is its ability to reduce training errors. AdaBoost is more adaptive than previous ensemble algorithms because it can effectively transform weak learning algorithms into strong learning algorithms.

2.6.5. eXtremeGradient Boosting (XGB)

The XGB algorithm was first proposed by Chen and Guestrin in 2011 [21]. The XGB has two advantages: it uses a variety of methods to prevent overfitting as much as possible, and it can automatically use the CPU's multi-threaded parallel computing to improve the running speed [31]. It is a ML model that has the advantages of flexibility and scalability.

2.7. Model performance

Our use of different modeling methods and different descriptors in the modeling process and general statistical indicators were needed to evaluate the model performance. We used a 10-fold cross-validation, the test set, and external validation sets for model evaluation. Although many accuracy metrics have been developed and used in different types of research, no fixed rules dictate the selection of efficiency metrics [13]. We used four evaluation metrics commonly utilized to evaluate model performance: the decisive coefficient R-squared (R^2) and the mean absolute error (MAE), the mean squared error (MSE) and median absolute error (MedAE). The MAE, MSE and MedAE can reflect the error level of the whole prediction sample, while the R^2 can reflect the regression performance of the model. Among them, MAE and MedAE are based on the absolute error, which are the mean and median of the error respectively. If the absolute error of the true value and the predicted value is concerned, MAE or MedAE can be selected. If we pay attention to the square of the difference between the true value and the predicted value, we can choose MSE, which is used to measure the deviation between the whole true value of samples and the predicted value of the model. The R^2 indicates that the model has poor performance when the value is close to 0, while a value close to 1 indicates that the model has good performance. The R^2, MAE, MSE and MedAE were calculated using the following equations:

$$R^2 = 1 - \frac{\sum_{i=0}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{m}(y_i - \bar{y})^2} \tag{1}$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{m}|y_i - \hat{y}_i|}{m} \tag{2}$$

$$R^2 = \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{m} \tag{3}$$

$$\mathrm{MAE} = \frac{median}{i = 1, \ldots, n}(|y_i - \hat{y}_i|) \tag{4}$$

In the formula, m is the number of samples, y_i is the true value of the sample, here is the retention time obtained from experiment, and $\hat{y}_i$ is the predicted value that is the predicted retention time of the compound by the model. The smaller the value of MAE, MSE and MedAE, the larger the value of R^2, indicating that the predictive performance of the model is stronger.

3. Results and discussion

3.1. Data set analysis

In this study, the abscissa was taken as the molecular weight of the compound, and the ordinate was taken as the retention time to construct a scatter diagram. The resulting diagram shows that the compounds in the training set and the test set share similar interval distribution, and it also demonstrates the distribution of the training set and the test set in the data set, in which the sample size of the test set accounts for 20% of the total data set and the training set accounts for 80% (Fig. 1). The dataset including name, SMILES and Rt

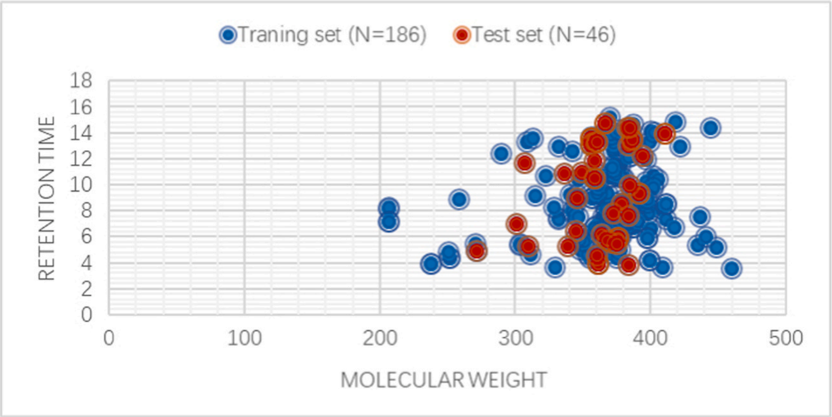


Fig. 1. Distribution of the training set and test set, where N represents the number of different data sets, and the data distribution is defined by molecular weight and retention time.

values for all 232 SCs in this study is shown in the supplementary material. This study is limited by the relatively small scale of the model set, and it cannot show the advantages of the applicability domain. However, when more data are available in the future, this key issue needs to be considered when developing enhanced models.

3.2. *Predictive performance of the different models*

The retention time prediction model was trained by a variety of ML algorithms, and significant performance differences exist between the different models. We illustrated this through R^2 and MAE values on the training set and test set.

We explored which combination of descriptor and machine learning algorithm has the best predictive performance by combining 13 feature descriptors with five different ML algorithms to train the Rt prediction models. Fig. 2A and B shows R^2 and MAE, respectively, in the test set under different descriptors corresponding to different machine learning algorithms.

Five machine learning algorithms (SVR, RF, GBR, ABR, and XGB) and 13 descriptors (1D&2D, FP, ExtFP, EStateFP, GraphFP, MACCSFP, PubchemFP, SubFP, SubFPC, KRFP, KRFPC, AP2D, and APC2D) were used to establish 65 prediction models. Comparison of the performance of 65 prediction models identified KRFPC and SubFPC as yielding the best results, followed by FP, KRFP, 1D&2D, MACCSFP, ExtFP, and APC2D. EStateFP, SubFP, and the remaining fingerprints performed the worst when the same algorithm was used, as shown in Fig. 2A.

We screened out better fingerprints that performed well in the experiment and combined them with different algorithms to build the models. The performance of different algorithms when the same fingerprint is used is shown in Fig. 3A. The figure confirms that the predictive performance was not particularly different for the various algorithms used in this study.

3.3. *Predictive performance of the model based on fingerprint combination*

The model results on the test set reveal that several algorithms, including SVR and GBR, have a better ability to fit the model when combined with some special descriptors, and that the features 1D&2D, SubFPC, and KRFPC perform the best. Based on the single variable model, the overall predictive performance was not as good as expected, and the R^2 value of the optimal model was only 0.78. However, this research is based on only a single variable model, whereas previous studies have tried to combine physicochemical descriptors and molecular fingerprints to form a combined fingerprint. That approach showed that the performance was better than the model based on single molecular fingerprints only [32,33]. Therefore, we chose a combination of good performance fingerprints to characterize the compounds as the model input to evaluate the predictive performance of the model. We combined the well-performing algorithms with its corresponding well-performing descriptors (FP, KRFPC, SubFPC, KRFP and 1D&2D were included in total) to generate 9 models. The top 6 models, based on their R^2 values, are listed in Table 1. The two best combined models were FP + KRFPC + SubFPC-SVR ($R^2_{test} = 0.821$) and FP + SubFPC-SVR ($R^2_{test} = 0.831$).

At the same time, we also compared the R^2 of the model based on the combined fingerprints with its corresponding single fingerprint model. We selected GBR and SVR algorithms which have good performance, and screened three molecular descriptors with good ability of characterization, then combined the screened molecular features into new features and applied the new features to the screened algorithms to generate new models. As shown in Fig. 3B, the R^2 is higher for the model based on the combined fingerprints than for the corresponding model based on a single fingerprint, indicating that the combined fingerprint model has achieved better predictive performance. For example, the combination of KRFP and 1D&2D features in the GBR algorithm model improves the prediction ability of the model. In the SVR model, the predictive performance of the model is significantly improved by combining the FP and SubFPC features than by utilizing the single feature model. In short, the performance of models based on different types of descriptors may depend on specific data sets, but we can still try to combine different types of descriptors to characterize compounds as a way to improve the model performance.

3.4. *Predictive performance of the optimal model*

Through the above research, we screened an SVR model based on FP and SubFPC fingerprints with good predictive performance. Table 2 shows the predictive performance parameters of the model validation set and test set. The developed model was available for free and has been uploaded to Github (https://github.com/RTPred/RT_Pred_for_SCs).

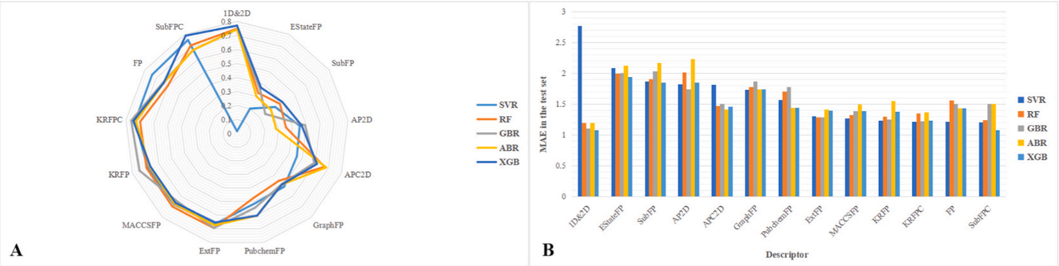


Fig. 2. R^2 (test set) in different descriptors under different algorithms(A) and MAE (test set) in different descriptors under different algorithms(B).

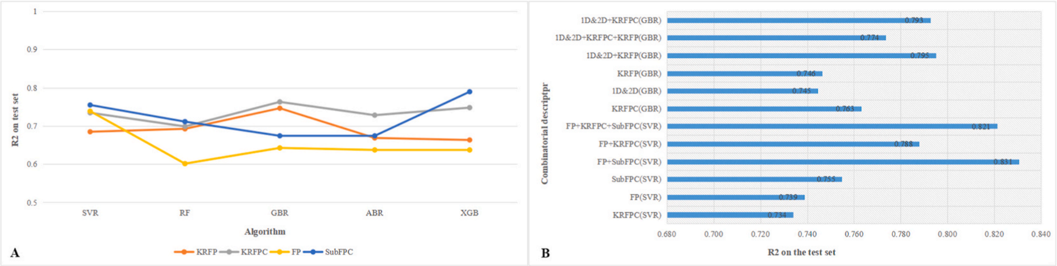


Fig. 3. Performance of five algorithms using the same fingerprint(A) and R^2 values of models based on single and combined descriptor(B).

Table 1
Predictive performance of two models based on combined fingerprint.

Algorithm	Descriptor	Train set		Validation set		Test set	
		MAE	R^2	MAE	R^2	MAE	R^2
SVR	FP + KRFPC	0.364	0.929	0.916	0.805	1.008	0.788
	FP + KRFPC + SubFPC	0.346	0.936	0.858	0.820	0.901	0.821
	FP + SubFPC	0.422	0.922	0.938	0.808	0.987	0.831
GBR	1D&2D + KRFPC + KRFP	0.214	0.988	1.269	0.725	1.153	0.774
	1D&2D + KRFPC	0.181	0.991	1.244	0.738	1.108	0.793
	1D&2D + KRFP	0.387	0.961	1.259	0.720	1.046	0.795

Table 2
Predictive performance of the optimal models (FP + SubFPC-SVR).

Dataset	R^2	MAE	MSE	MedAE
Validation set	0.808	0.938	2.016	0.546
Test set	0.831	0.987	1.728	0.660

3.5. Model predictive performance of the external test set

Traditional QSAR research mostly adopts a single modeling method and establishes a single model based on one type of descriptor. Compared to this traditional experimental approach, the present study focuses on the combination of multiple types of models using different machine learning algorithms and chemical descriptors to establish models for predicting the retention times of SCs. We used the best combined fingerprint model, FP + SubFPC-SVR, to predict the retention time of 236 SCs compounds including four new SCs

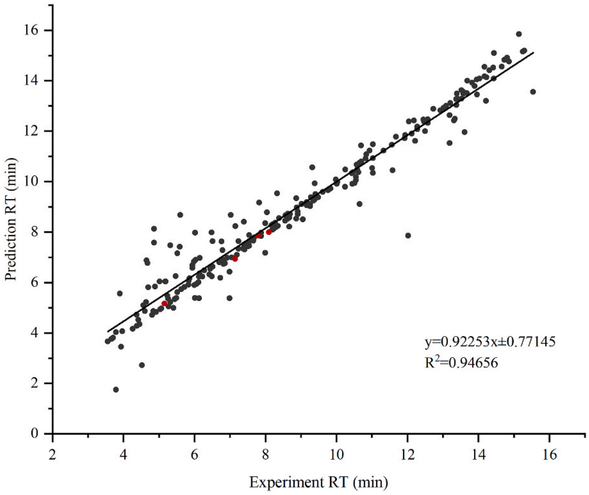


Fig. 4. Scatter plot for experimental versus predicted retention times of 236 SCs. The compounds in the external datasets are indicated by red dots; From left to right: 5-fluoro-AB-PINACA, 4-fluoro-MDMB-BICA, JWH-412-N-(5-hydroxypentyl)-metabolite and 5-fluoro-MDMB-PICA. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

compounds from external datasets. Fig. 4 shows the fit result between experimental and predicted retention time for the final model including all 236 compounds, and the four compounds in the external datasets are indicated by red dots. Table 3 shows the results of the experiments and the predicted retention times of these compounds. The relative error was within 3% and the R_t^p inaccuracy remained at <0.5min, indicating that the retention time of SCs predicted by the QSAR model had a good correlation with the chemical structure characteristics of the compounds.

3.6. Comparison with existing research

The retention time prediction model proposed by Polettini et al. [8] has been applied in the untargeted identification and isomer identification of SCs, while Polettini's modeling samples focus on the SCs parent compounds. On the basis of previous studies, we expanded the amount of modeling data and increase the number of metabolites including 173 SCs parent compounds and 59 metabolites. The SCs metabolites used for modeling is relatively large, which is helpful to improve the performance of the retention time prediction model with metabolites. And we used the same combination descriptors and dataset partition method to compare the predictive performance of 232 SCs between the multiple linear regression (MLR) algorithm from Polettini et al. and the SVR algorithm used in our study. A comparison result in Table 4 showed that the R^2 value of cross validation of SVR model in this work is 0.808 which is better than 0.205 of the MLR model of the reference. The MAE value of the validation set of SVR is 0.938 which is smaller than 1.903 of the MLR. And the test set has similar results. It demonstrates that the performance of SVR algorithm is better than that of MLR.

Our research and that of Polettini are both prediction models for the retention time of SCs in a kind of LC system. However, Pasin's research has proposed a retention time prediction model integrating multiple LC systems with different elution conditions using the retention time data from the online crowd-sourced database HighResNPS [10]. This research focused on the integration of different LC systems and the establishment of databases, so only labels (names) were used to distinguish different categories. In comparison, our model takes different molecular descriptors as independent variable and predict the retention times by inputting compound structures, and we focus on the impact of different descriptors and algorithms on the predictive performance. Therefore, the algorithms and descriptors which perform good can be used for reference by other laboratories. Simultaneously, our research data including 232 SCs ' retention times and related structural information will further expand the HighResNPS database (https://highresnps.forensic.ku.dk/) that can help to study the algorithms of integrating different LC systems.

However, our research still has several limitations. First, the specific regression model in our study can only be used to SCs, but the model can be retrained by inputting new categories. The modeling process is simple and fast, and some models can be completed in a few seconds. Besides, it will have the potential source of inaccuracies because the retention time may change slightly over time. This can be caused by small changes in temperature due to the viscous heat effect but also by column aging. These changes are not always systematic, therefore ensuring the column temperature constant is important, such as maintaining the temperature of the column incubator during the experiment. If the set of retention times used for modeling was accumulated over time, the confidence range can be increased by a certain correction. The accuracy of model prediction might be reduced if the "current" system is sufficiently different from the original retention times system. Moreover, the performance of prediction model can be further improved by input more LC system parameters or mass spectrum information to the model.

4. Conclusion

The universality and harmfulness of SCs emphasize the need for effective tools to predict the chemical properties of SCs and homologous substances to aid in the analysis and identification of SCs in forensic poisonings. The purpose of this study was to establish a QSAR model that can predict the retention time of SCs. We combined the limited experimental results as training data, used five regression algorithms, including SVR and GBR, and trained several regression models with good predictive performance by combining different types of descriptors. We selected an SVR model with good statistical performance to predict the retention time of four SCs from external sets, and the predictive performance was good. The current research provides an effective tool for predicting the chromatographic properties of SCs. The chromatographic information reflects the retention of chemicals on the chromatographic column and is a crucial screening tool for improving the reliability of identification by comparison with experimental retention times. Therefore, this approach will provide useful information for the analysis and identification of non-targeted substances. In the absence of reference standards, predicting the retention time according to the molecular structure can improve the reliability of structural analysis and the identification of unknown metabolites in non-targeted LC-HRMS analysis.

Research funding

None declared.

Author contributions

Lina Wu and Fu Xiao designed and performed the experiments, analyzed and interpreted the data and wrote the paper. Xiaomin Luo and Keming Yun contributed reagents, materials, analysis tools and critically revised important intellectual content of the paper. Di Wen and Jiaman Lin also analyzed the data. Shuo Yang and Tianle Li also performed the experiments. Ping Xiang and Yan Shi conceived and designed the experiments, contributed reagents, materials, analysis data and final approval of the version submitted. All

Table 3
Performance of MODEL for predictions of Rt for SCs compounds.

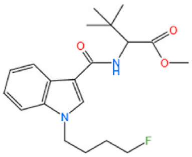
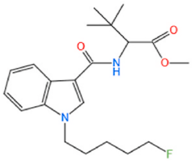
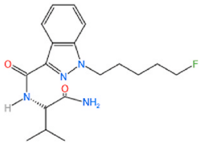
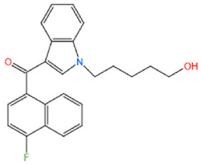
Compound	$[M+H]^+$	Molecular Formula	structure	R_t^E (min)	R_t^P (min)	Prediction error (%)
4-fluoro-MDMB-BICA	363.21	$C_{20}H_{27}FN_2O_3$		6.94	7.15	2.96
5-fluoro-MDMB-PICA	377.22	$C_{21}H_{29}FN_2O_3$		7.99	8.10	1.33
5-fluoro-AB-PINACA	349.20	$C_{18}H_{25}FN_4O_2$		5.17	5.16	0.20
JWH-412-N-(5-hydroxypentyl)-metabolite	376.17	$C_{24}H_{22}FNO_2$		7.86	7.80	0.81

Table 4
Comparison of SVR and MLR algorithms in predicting the RT of SCs compounds.

Reference	Algorithm	Descriptor	Validation set		Test set	
			MAE	R^2	MAE	R^2
This work	SVR	FP + SubFPC	0.938	0.808	0.988	0.831
Polettini et al. [8]	MLR	FP + SubFPC	1.903	0.205	1.670	0.611

authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Informed consent

Informed consent was obtained from all individuals included in this study.

Ethical approval

The local Institutional Review Board deemed the study exempt from review.

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.

Acknowledgments

This study was supported by grants from National Key Research and Development Program of China (2022YFC3302003), National Natural Science Foundation of China (81971789,82271932), Central Guidance on Local Science and Technology Development Fund of Hebei Province (226Z5601G), Shanghai Key Laboratory of Forensic Medicine (21DZ2270800) and Shanghai Forensic Service Platform (19DZ2292700).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2023.e16671.

References

[1] A. Pisarska, P. Deluca, Z. Demetrovics, et al., Novel psychoactive substances (NPS) - knowledge and experiences of drug users from Hungary, Poland, the UK and the USA, Neuropsychopharmacol Hung 21 (2019) 152–163.

[2] D.K. Tracy, D.M. Wood, D. Baumeister, Novel psychoactive substances: types, mechanisms of action, and effects, BMJ 356 (2017) i6848, https://doi.org/10.1136/bmj.i6848.

[3] K. Cohen, A. Weinstein, The effects of cannabinoids on executive functions: evidence from cannabis and synthetic cannabinoids-A systematic review, Brain Sci. 8 (2018), https://doi.org/10.3390/brainsci8030040.

[4] K. Cohen, A.M. Weinstein, Synthetic and non-synthetic cannabinoid drugs and their adverse effects-A review from public health prospective, Front. Public Health 6 (2018) 162, https://doi.org/10.3389/fpubh.2018.00162.

[5] P. Bonini, T. Kind, H. Tsugawa, et al., Retip: retention time prediction for compound annotation in untargeted metabolomics, Anal. Chem. 92 (2020) 7515–7522, https://doi.org/10.1021/acs.analchem.9b05765.

[6] U.W. Liebal, A.N.T. Phan, M. Sudhakar, et al., Machine learning applications for mass spectrometry-based metabolomics, Metabolites 10 (2020), https://doi.org/10.3390/metabo10060243.

[7] Z. Zhou, R.N. Zare, Personal information from latent fingerprints using desorption electrospray ionization mass spectrometry and machine learning, Anal. Chem. 89 (2017) 1369–1372, https://doi.org/10.1021/acs.analchem.6b04498.

[8] A.E. Polettini, J. Kutzler, C. Sauer, et al., LC-QTOF-MS presumptive identification of synthetic cannabinoids without reference chromatographic retention/mass spectral information. I. Reversed-phase retention time QSPR prediction as an aid to identification of new/unknown compounds, J. Anal. Toxicol. 45 (2021) 429–439, https://doi.org/10.1093/jat/bkaa126.

[9] A.E. Polettini, J. Kutzler, C. Sauer, et al., LC-QTOF-MS presumptive identification of synthetic cannabinoids without reference chromatographic retention/mass spectral information. II. Evaluation of a computational approach for predicting and identifying unknown high-resolution product ion mass spectra, J. Anal. Toxicol. 45 (2021) 440–461, https://doi.org/10.1093/jat/bkaa127.

[10] D. Pasin, C.B. Mollerup, B.S. Rasmussen, et al., Development of a single retention time prediction model integrating multiple liquid chromatography systems: application to new psychoactive substances, Anal. Chim. Acta 1184 (2021), 339035, https://doi.org/10.1016/j.aca.2021.339035.

[11] Y. Shi, M. Liu, X. Li, et al., Simultaneous screening of 239 synthetic cannabinoids and metabolites in blood and urine samples using liquid chromatography-high resolution mass spectrometry, J. Chromatogr. A 1663 (2022), 462743, https://doi.org/10.1016/j.chroma.2021.462743.

[12] C.W. Yap, PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints, J. Comput. Chem. 32 (2011) 1466–1474, https://doi.org/10.1002/jcc.21707.

[13] J. Wang, P. Du, T. Niu, et al., A novel hybrid system based on a new proposed algorithm—multi-Objective Whale Optimization Algorithm for wind speed forecasting, Appl. Energy 208 (2017) 344–360, https://doi.org/10.1016/j.apenergy.2017.10.031.

[14] C. Zhang, Y. Zhou, S. Gu, et al., In silico prediction of hERG potassium channel blockage by chemical category approaches, Toxicol. Res. 5 (2016) 570–582, https://doi.org/10.1039/c5tx00294j.

[15] E.A. Sosnina, D.I. Osolodkin, E.V. Radchenko, et al., Influence of descriptor implementation on compound ranking based on multiparameter assessment, J. Chem. Inf. Model. 58 (2018) 1083–1093, https://doi.org/10.1021/acs.jcim.7b00734.

[16] A. Gupta, V. Kumar, P. Aparoy, Role of topological, electronic, geometrical, constitutional and quantum chemical based descriptors in QSAR: mPGES-1 as a case study, Curr. Top. Med. Chem. 18 (2018) 1075–1090, https://doi.org/10.2174/1568026618666180719164149.

[17] J.Y. Hsia, C.J. Lin, Parameter selection for linear support vector regression, IEEE Transact. Neural Networks Learn. Syst. 31 (2020) 5639–5644, https://doi.org/10.1109/tnnls.2020.2967637.

[18] B.H. Menze, B.M. Kelm, R. Masuch, et al., A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data, BMC Bioinf. 10 (2009) 213, https://doi.org/10.1186/1471-2105-10-213.

[19] A. Mayr, H. Binder, O. Gefeller, et al., Extending statistical boosting. An overview of recent methodological developments, Methods Inf. Med. 53 (2014) 428–435, https://doi.org/10.3414/me13-01-0123.

[20] J.Q. Chen, H.Y. Chen, W.J. Dai, et al., Artificial intelligence approach to find lead compounds for treating tumors, J. Phys. Chem. Lett. 10 (2019) 4382–4400, https://doi.org/10.1021/acs.jpclett.9b01426.

[21] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, San Francisco, California, USA, 2016, pp. 785–794.

[22] T. Zou, Y. Dou, H. Mi, et al., Support vector regression for determination of component of compound oxytetracycline powder on near-infrared spectroscopy, Anal. Biochem. 355 (2006) 1–7, https://doi.org/10.1016/j.ab.2006.04.025.

[23] C.Y. Zhao, R.S. Zhang, H.X. Liu, et al., Diagnosing anorexia based on partial least squares, back propagation neural network, and support vector machines, J. Chem. Inf. Comput. Sci. 44 (2004) 2040–2046, https://doi.org/10.1021/ci049877y.

[24] Z. Yuan, T.L. Bailey, R.D. Teasdale, Prediction of protein B-factor profiles, Proteins 58 (2005) 905–912, https://doi.org/10.1002/prot.20375.

[25] M. Song, C.M. Breneman, J. Bi, et al., Prediction of protein retention times in anion-exchange chromatography systems using support vector regression, J. Chem. Inf. Comput. Sci. 42 (2002) 1347–1357, https://doi.org/10.1021/ci025580t.

[26] A. Sarica, A. Cerasa, A. Quattrone, Random forest algorithm for the classification of neuroimaging data in alzheimer's disease: a systematic review, Front. Aging Neurosci. 9 (2017) 329, https://doi.org/10.3389/fnagi.2017.00329.

[27] R. Caruana, A. Niculescu-Mizil, An empirical comparison of supervised learning algorithms, in: Proceedings of the 23rd International Conference on Machine Learning, Association for Computing Machinery, Pittsburgh, Pennsylvania, USA, 2006, pp. 161–168.

[28] A. Naemi, T. Schmidt, M. Mansourvar, et al., Machine learning techniques for mortality prediction in emergency departments: a systematic review, BMJ Open 11 (2021), e052663, https://doi.org/10.1136/bmjopen-2021-052663.

[29] A. Natekin, A. Knoll, Gradient boosting machines, a tutorial, Front. Neurorob. 7 (2013) 21, https://doi.org/10.3389/fnbot.2013.00021.

[30] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. Syst. Sci. 55 (1997) 119–139, https://doi.org/10.1006/jcss.1997.1504.

[31] Z. Zhao, W. Yang, Y. Zhai, et al., Identify DNA-binding proteins through the extreme gradient boosting algorithm, Front. Genet. 12 (2021), 821996, https://doi.org/10.3389/fgene.2021.821996.

[32] Y. Chen, H. Yang, Z. Wu, et al., Prediction of farnesoid X receptor disruptors with machine learning methods, Chem. Res. Toxicol. 31 (2018) 1128–1137, https://doi.org/10.1021/acs.chemrestox.8b00162.

[33] H. Du, Y. Cai, H. Yang, et al., In silico prediction of chemicals binding to aromatase with machine learning methods, Chem. Res. Toxicol. 30 (2017) 1209–1218, https://doi.org/10.1021/acs.chemrestox.7b00037.