QSRR model for identification and screening of emerging pollutants based on artificial intelligence algorithms

Qi He, Hua Li, Binyan Jin, Wei Li, Bing Shao and Li Zhang

ABSTRACT
It is urgent to identify and screen emerging pollutants (EPs), which have caused great harm to human health and the environment. In their detection of liquid chromatography-mass spectrometry (LC-MS), the quantitative structure–retention relationship (QSRR) model is simple and efficient to predict the retention behavior of compounds. In the present work, we collected more data with the relative retention time (RRT) of 490 compounds, and filtered the molecular descriptors with lasso regression and multiple linear regression analysis. Then ten important molecular descriptors were screened and applied the QSRR models with deep neural network (DNN), multiple linear regression (MLR), and support vector machine. The DNN model had the best accuracy which the correlation coefficient R2 reached 0.913. Finally, we determined the applicability of the DNN model through a descriptor value range to assist in the identification and screening of EPs.

1. Introduction
In recent years, the focus of environmental analysis has gradually shifted from classic contaminants to emerging pollutants (EPs). EPs refer to pollutants newly discovered in the environment or known for a long time but only recently attracting attention. EPs include pharmaceuticals, personal care products (PPCPs), pesticides, cleaning products, cosmetics, fragrances, and hormones [1,2]. EPs have now been widely observed in wastewaters and wastewater treatment plants that are not designed to remove these kinds of pollutants and their degradation-resistant metabolites; hence, they are disposed of directly into water bodies [2,3]. As EPs have caused great harm to human health and the environment [4], it is important to screen and identify them. Liquid chromatography-mass spectrometry (LC-MS) plays an important role in the analysis of water environmental pollutants, especially in the screening of a large number of pollutants. At present, the most widely used analytical method in the identification of EPs is LC-MS, which has the characteristics of high resolution, high sensitivity, and good mass accuracy [5]. However, considering the diverse behaviors and production sources of EPs, it is challenging to screen and identify EPs by LC-MS [6] only depending on precise molecular weight and fragment ion without the retention time (RT). Thus, prediction of the retention time will be useful to screen and identify EPs.

Quantitative structure–retention relationships (QSRRs) are statistically derived models relating the quantities characterizing the molecular structure of the analytes (molecular descriptors and fingerprints) with their chromatographic retention parameters [7–9]. So, RT of a compound can be predicted by only its molecular structure, greatly reducing the time and cost required for identification. QSRRs are widely used in predicting
the RT for a new analyte and identifying unknown analytes. Miller [10] constructed a QSRR for anti-doping screening using 86 compounds included in the London 2012 Olympic and Paralympic Games drug testing schedule by artificial neural network (ANN) methods. For 93% of the doping-related compounds, the prediction error of the RT was less than 0.5 min. Jiao [11] modeled the relationship between the molecular distance-edge vector (MEDV) index and the relative retention time (RRT) of polybrominated diphenyl ethers (PBDEs) by using a radial basis function artificial neural network (RBF-ANN). There were 128 PBDEs in the dataset, with 88 PBDEs in the training set and 38 PBDEs in the test set. The root mean square relative error (RMSRE) of the test set prediction was 8.21. Bade [12] used the ANN method to construct a QSRR model for a dataset containing 444 Rts of emerging pollutants; the model used 100 randomly selected compounds and exhibited a 95% prediction accuracy within 2 min. Norelddeen [13] constructed a QSRR for 95 veterinary drugs using multiple linear regression (MLR); the model between the drug RT and the selected molecular descriptors had a good correlation coefficient ($R^2 = 0.966$). Domingo-Almenara [14] constructed a deep neural network (DNN) model for the METLIN small molecule retention time database (containing 80,038 small molecule RT data, 75% as training set and 25% as test set) with RT as response value and extended connectivity fingerprints (ECFP) as variable. The model was able to achieve accurate RT prediction with median relative error and absolute errors were 6% and 35s, respectively. However, there have been few QSRR models for EPs, and the relationship between molecular descriptors and RRT has not been analyzed. In our work, 490 compounds (391 compounds after optimization) including EPs with RRTs were obtained from the MassBank. The molecular descriptors were calculated by the software Molecular Operating Environment (MOE 2020.09) [15] and PaDEL-descriptor [16] as the variables. Taking the RRT as the response value, three methods – deep neural network (DNN), MLR, and a support vector machine (SVM) – were used to build the prediction models. This study aims to establish a simple and accurate QSRR model to predict the RRT of structurally diverse EPs.

2. Materials and Methods
2.1. Dataset and optimization
The RRTs data of 490 compounds from MassBank (https://massbank.eu/) was collected and covered a large range (From 0.04 to 1.858), which included pharmaceuticals, pesticides, hormones, biochemical reagents, and others (Table S1). Their structural diversity was confirmed using a Tanimoto heat map (Figure S1). The training set and the test set were divided according to fingerprint cluster analysis and RRT classification. Then, there were 343 compounds in the training set and 147 compounds in the test set as the first classification.

Further, we analyzed and optimized the dataset by removing non-retained molecules and outliers. The non-retained molecules were defined as molecules that will not retain in the column and will elute before the start of the gradient, typically within the first minute [14]. Then, we removed these non-retained molecules with RRTs less than 0.085. The outliers were defined as molecules with a Cook’s distance > [4/(number of samples- The number of features – 1)]. The Cook’s distance of the molecule was calculated by the ‘car’ package [17] in R. Finally, 275 compounds make up the training set and 116 compounds form the test set for the following models.

2.2. Calculation and screening of molecular descriptors
We used the software MOE and the free software PaDEL to calculate the molecular descriptors. A total of 2311 molecular descriptors were calculated and screened before the model was constructed to reduce data redundancy. Firstly, highly correlated descriptors (the correlation coefficient > 0.9), descriptors not available for all compounds, and descriptors with a nearly constant value were removed. In order to ensure the distinguishability of the descriptors for different samples, the descriptors were screened according to the criteria of mean > 0.01, standard deviation > 0.1, and relative standard deviation > 0.1. So, 306 descriptors were removed after the first screening. Secondly, we used the ‘lars’ package [18] in R to perform lasso regression analysis to filter the over-fitting descriptors. The Cp-statistic value of the model changing with steps is presented in Figure S2. When the curve tends to be smooth, the step number here is chosen as the best descriptor value, and 29 molecular descriptors were retained. Thirdly, MLR analysis was performed in R and the descriptors with high significance (p-value < 0.001) were selected. Finally, the ten descriptors were screened and used to build models, including $E_{nb}$, XLogP, $P_V_{FPOL}$, Q_V_{FPNEG}, G_SLOGP_0, SpMAD_Dzp, $P_V_{FNEG}$, SpMaxB_Bhm, MIC0, and TDB4v. The correlation coefficient and definitions of these descriptors are shown in Figure S3 and Table S2, respectively.

2.3. QSRR model building
All QSRR models were built using DNNs, MLR, and SVM methods (using the ‘e1071’ package [19] through R. The SVM type is ‘EPs-regression’, and the SVM kernel is ‘linear’. The DNN model was constructed using
‘keras’ package for R [20]. It consisted of two fully connected hidden layers of 256 and 128 units activated via a non-linearity function (ReLU) connected to an output layer consisting of one unit with no activation. The optimizer was ‘Adam’. Mean squared error (MSE) and mean absolute deviation (MAE) were used as a loss function and metrics function, respectively. A total of 250 epochs were used to train the model, and the batch size was 1.

2.4. Model validation

The validation of the QSRR model was conducted by the internal and external validation. The internal validation of the training set was performed using K-fold cross-validation (CV) at K = 10. The MAE and MSE were used to adjust and evaluate the accuracy, quality, and determine the error between the experimental and predicted RRTs in the QSRR model. For the external validation of the test set, we calculated MAE (Eq. 1a), MSE (Eq. 1b), and the correlation coefficients $R^2$ between the experimental and predicted RRTs to assess the ability of the model to predict new compounds.

$$MSE = \frac{\sum_{i=1}^{n}(y_a - y_b)^2}{n} \quad (1a)$$

$$MAE = \frac{\sum_{i=1}^{n}|y_a - y_b|}{n} \quad (1b)$$

where $y_a$ is the experimental value of the relative retention time and $y_b$ is the predicted value of the relative retention time.

3. Results and Discussion

3.1. Optimization of dataset

We calculated the FP molecular fingerprint using the PaDEL-Descriptor software [16] and plotted the Tanimoto similarity heat map to evaluate the diversity of the 490 molecular structures (Figure S1). The Tanimoto similarity index ranges from 0 to 1, and their colors are changed from blue to red. The smaller the similarity index, the bluer the color. The color of the heat map was almost blue as a whole, and the average Tanimoto similarity index of the dataset was 0.26 (Figure S1). This indicates that the molecules in the dataset have a low similarity and good structural diversity.

The dataset was divided into the training set and the test set. To ensure the structural diversity of the training set, we used MOE’s Fingerprint Cluster function to perform cluster analysis (according to 50% similarity and 50% overlap) on the molecules, which is based on the MACCS molecular fingerprint. 490 molecules were divided into 136 classes and each class existed in the training set. Then, we obtained a training set with 343 molecules and a test set with 147 molecules.

The first model was constructed using the MLR method to predict the RRTs of the training set. However, we found there were two kinds of molecules with a large error (Figure 1).

One was non-retained molecules (RRT < 0.085) and the other was outliers determined by the Cook’s distance (Figure S4). Thus, we removed these compounds (60 non-retained molecules and 30 outliers) to reconstruct the MLR model. The correlation coefficient ($R^2$) between the experimental and predicted RRTs increased from 0.8888 to 0.9216, while the largest relative error percent decreased from 77% to 33%. So,
we would use the optimized dataset (training set with 275 molecules and test set with 116 molecules) to build and validate the following models.

3.2. Analysis of molecular descriptors

After three steps of descriptor screening, we obtained ten molecular descriptors with low two-by-two correlation and high significance (p-value ≤ 0.001). As we know, the chromatographic retention behavior of compounds is caused by the interactions between molecules, stationary phases, and mobile phases. The molecular structures and their properties have the important correlation. In our model, the ten molecular descriptors could represent different molecular properties (Table S2). So, we further analyze the mathematical relationship between RRT and these ten molecular descriptors.

The molecules were divided into two types: those with RRTs ≤ 1 and those with RRTs ≥ 1. A box diagram was used to analyze the value differences between the two types of molecules using ten molecular descriptors (Figure 2).

Which showed that seven descriptors were quite different. The values of XLogP, G_S_LOGP_0, P_V_FNEG, SpMax8_Bhm, and TDB4v when RRTs ≥ 1 are greater than those when RRTs < 1, while P_V_FPOL, and MICO are the opposite.

In reversed phase chromatography, the polarity of the mobile phase is greater than that of the stationary phase. So, the more polar substances flow out earlier, while the less polar substances flow out later. As shown in Figure 2, XLogP is related to the polarity of the substance. The larger the Xlogp, the less polar the substance is, and therefore the longer the RRT.

P_V_FPOL is also related to the polarity of the substance. The larger the P_V_FPOL, the more polar the substance is, so the shorter the RRT. P_V_FNEG is related to the electronegativity of the substance. The larger the P_V_FNEG, the stronger the electronegativity of the substance, which caused the RRT longer through a stationary phase with opposite electrical properties.

3.3. Model comparison

We used three methods DNN, MLR, and SVM to build QSRR models of the training set, respectively. The predicted RRTs of the training set and test set were listed in Tables S3 and S4. The correlation curve between the experimental RRTs and predicted RRTs of the test set was plotted and the relative error was calculated (Figure 3).

The correlation coefficients R² of MLR and SVM models were 0.894 and 0.892, respectively, which had the similar accuracy. The MSE of the MLR model was 1.7% less than that of the SVM model, and the MAE of these two models was almost same. This probably because the parameter kernel of SVM model we chosen is ‘linear’ rather than polynomial, radial basis, or sigmoid. Also, we found that the MAE and MSE were the smallest, and the accuracy was highest when setting linear as the kernel to train the model by 10-fold cross-validation (Table S5). Therefore, we speculate these screened molecular descriptors can well describe the properties related to molecular retention behavior, and there is an obvious linear relationship between them. The molecular descriptors and RRTs can be quantified well by the simplest MLR model. The regression equation is shown as Eq. 2.

Figure 2. Relationship analysis between descriptor and RRT. The molecules were divided into two types: with RRT < 1 (yellow) and ≥ 1 (blue).
Figure 3. The correlation curve between experimental RRT and predicted RRT of (a) MLR, (b) SVM, and (c) DNN. (d) The relative error between experimental RRT and predicted RRT of three models.

\[
RRT_{\text{pre}} = -0.006185 + 0.001096 \times E_{\text{pb}} + 0.07526 \times X\text{LogP} - \\
1.246 \times \text{PEOE_VSA FPOL} + 0.633 \times \text{Q VSA FPNEG} + 0.2333 \times \\
\text{GCUT SLOGP} + 0.01789 \times \text{SpMAD Dzp} + 0.6737 \times \text{PEOE_VSA FNEG} + \\
0.1549 \times \text{SpMax Bhm} - 0.01117 \times \text{MIC0} + 0.007161 \times \text{TDB4v}.
\]

\[N = 275, R^2 = 0.9216, R^2_{\text{Adj}} = 0.9186, F = 310.3\]  \hspace{1cm} (2)

Although the linear models have been able to quantify RRTs and descriptors well, there may be a better method to construct a model between the RRTs and the ten descriptors. Therefore, we tried to build DNN model based on deep learning and figured out this model performed best. The ten descriptors were used as an input layer, and RRT was treated as output layer. In the training procedure, we used 10-fold cross-validation to continuously tune the model parameters. Finally, we obtained the optimal parameters (shown in 2.3) with minimum MAE and MSE values (Table S5). For the test set, compared with the MLR and SVM linear regression models, the DNN regression model had better accuracy. The correlation coefficient between the experimental RRTs and predicted RRTs of the DNN was 0.02 higher than that of the linear model. The MSE and MAE decreased by 17.3% and 13.7% respectively, compared with the linear model, and the average relative error of the DNN was 14%, which was 2% lower than the linear model. Compared with the RBF-ANN model of Jiao et al., the correlation coefficient between the experimental and predicted RRTs of the test set was lower by 0.076. However, the number of compounds in our test set was three times larger than theirs, and their model was specific to PBDEs, while our model was applicable to EPs with diverse structures and had universal advantage.

3.4. Model applicability analysis

When using consistent methods, the applicability of the model depends on the diversity of the sample structure of the training set. However, when the structure of the training set was diverse, due to the abstraction of the structure, we cannot define which structure the model is applicable to. We know that molecular descriptors can be used to describe the properties of molecules, which is molecular structure information. Therefore, we speculate that we can judge whether the RRTs of the molecule can be accurately predicted by the model through the value of molecular descriptors.
4. Conclusions

We show the distribution of the values of all molecules in the optimization dataset over the ten descriptors by a kernel density function diagram (Figure 4a). Each line represents a molecular descriptor, and the horizontal coordinate is the normalized descriptor value. The vertical coordinate is the distribution density of the compound. The redder the color, the more compounds that fall within the range of the descriptor. We also show the values of compounds that can be accurately predicted (relative error less than 10%) on each descriptor using a box diagram (Figure 4b). The vertical coordinate is the value of the descriptor after normalization. Comparing the two graphs, we found that the distribution of descriptor values of compounds that can be accurately predicted corresponded to the most concentrated distribution of descriptors of compounds in the optimization dataset. XLogP ranged from -0.871 to 4.574; TDB4v ranged from 520.36 to 1010.28; SpMax8_Bhm ranged from 2.27 to 3.18; SpMAD_Dzp ranged from 7.26 to 13.66; Q_V_FPNEG ranged from 0.041 to 0.37; P_V_FPOL ranged from 0.028 to 0.39; P_V_FNEG ranged from 0.18 to 0.61; MICO ranged from 9.64 to 15.36; G_SLOGP_0 distributed in three ranges: -2.26 to -2.25, -1.49 to -1.43 and -1.17 to -0.78; and E_nb ranged from -36.21 to 68.665. We speculate that molecules with descriptor values in this range will be accurately predicted by our DNN model.

Figure 4. (a) Kernel density function diagram indicating the distribution of the values of all molecules in the optimization dataset over the ten descriptors; (b) box diagram indicating the values of compounds that can be accurately predicted (relative error less than 10%) on the ten descriptors.

Acknowledgments

We thank LetPub (www.letpub.com) for its linguistic assistance during the preparation of this manuscript.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was funded by National Key R&D Program of China (grant number 2018YFC1602406); National Natural Science Foundation of China (grant number U173620).

Author contributions

Conceptualization, Li Zhang; methodology, Qi He; software, Hua Li; validation, Binyan Jin and Wei Li; formal analysis, Qi He, Hua Li and Li. Zhang; investigation, Qi He; writing – original draft, Qi He; writing – review & editing, Li Zhang; funding acquisition, Li Zhang. All authors have read and agreed to the published version of the manuscript.

References

[1] Mila E, Nika M-C, Thomaidis NS. Identification of first and second generation ozonation transformation products of niflumic acid by LC-QToF-MS. J Hazard Mater. 2019;365:804–812.
[2] Vargas-Berrones K, Bernal-Jácome L, Diaz de León-Martínez L, et al. Emerging pollutants (EPs) in Latin América: a critical review of under-studied EPs, case of study -Nonylphenol. SciTotal Environ. 2020;726:138493.

[3] Van Zijl MC, Aneck-Hahn NH, Swart P, et al. Estrogenic activity, chemical levels and health risk assessment of municipal distribution point water from Pretoria and Cape Town, South Africa. Chemosphere. 2017;186:305–313.

[4] Wiest L, Gosset A, Fildier A, et al. Occurrence and removal of emerging pollutants in urban sewage treatment plants using LC-QToF-MS suspect screening and quantification. SciTotal Environ. 2021;774:145779.

[5] Žuvela P, Skoczylas M, Jay Liu J, et al. Column Characterization and Selection Systems in Reversed-Phase High-Performance Liquid Chromatography. Chem Rev. 2019;119(6):3674–3729.

[6] Egbuna C, Amadi CN, Patrick-Iwuanyanwu KC, et al. Emerging pollutants in Nigeria: a systematic review. Environ Toxicol Pharmacol. 2021;85:103638.

[7] Hansch C, Fujita T. p-
-analysis. A Method for the Correlation of Biological Activity and Chemical Structure. J Am Chem Soc. 1964;86(8):1616–1626.

[8] Kaliszczan RQSRR. Quantitative Structure-(Chromatographic) Retention Relationships. Chem Rev. 2007;107(7):3212–3246.

[9] Talebi M, Schuster G, Shellie RA, et al. Performance comparison of partial least square-related variable selection methods for quantitative structure retention relationships modelling of retention times in reversed-phase liquid chromatography. J Chromatogr A. 2015;1424:69–76.

[10] Miller TH, Musenga A, Cowan DA, et al. Prediction of chromatographic retention time in high-resolution anti-doping screening data using artificial neural networks. Anal Chem. 2013;85(21):10330–10337.

[11] Jiao L, Xue Z, Wang G, et al. QSRR study on the relative retention time of polybrominated diphenyl ethers (PBDEs) by using molecular distance-edge vector index. Chemometr Intell Lab Syst. 2014;137:91–96.

[12] Bade R, Bijlsma L, Miller TH, et al. 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. SciTotal Environ. 2015;538:934–941.

[13] Noreldeen HAA, Liu X, Wang X, et al. Quantitative structure-retention relationships model for retention time prediction of veterinary drugs in food matrices. Int J Mass Spectrom. 2018;434:172–178.

[14] Domingo-Almenara X, Guijas C, Billings E, et al. The METLIN small molecule dataset for machine learning-based retention time prediction. Nat Commun. 2019;10(1):5811.

[15] Chemical Computing Group ULC, Molecular Operating Environment (MOE), Molecular Operating Environment (MOE), 2020.09 (Montreal, QC, Canada) 2020

[16] Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem. 2011;32(7):1466–1474.

[17] Fox J, Weisberg S. An R Companion to Applied Regression. Third. Thousand Oaks CA: Sage. URL. 2019. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

[18] Hastie T, Efron B. lars: least Angle Regression, Lasso and Forward Stagewise. R package version 1.2. https://CRAN.R-project.org/package=lars. 2013.

[19] Meyer D, Dimitriadou E, Hornik K, et al. e1071: misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-3. https://CRAN.R-project.org/package=e1071. 2019.

[20] Allaire JJ, Chollet F. keras: r Interface to ‘Keras’. R package. https://CRAN.R-project.org/package=keras 2018.