



ОРИГИНАЛЬНЫЕ СТАТЬИ

Original article

UDK 543.544

doi: 10.17308/sorpchrom.2023.23/11317

Extraction of information about the molecule structure directly from GC-MS data

Dmitry D. Matyushin, Anastasiya Yu. Sholokhova[✉]

Institute of Physical Chemistry and Electrochemistry RAS (IPCE RAS), Moscow, Russian Federation, shonastya@yandex.ru[✉]

Abstract. Gas chromatography – mass spectrometry (GC-MS) is a very important method of chemical analysis. GC-MS can be used for non-target chemical analysis and preliminary screening of completely unknown compounds. Electron ionization mass spectrometry is commonly used in GC-MS. Some information can be extracted directly from GC-MS data using machine learning methods. There are several previous works in which machine learning models extract information about the presence or absence of given substructures in a molecule directly from the electron ionization mass spectrum. Rarely, the additional data such as molecular weight and retention index are used together with the mass spectrum as input features of such models, however, no systematic comparison of how the use of such data increases the accuracy of the prediction was previously conducted. In this work, gradient boosting was used for prediction of the presence or absence of given substructures in a molecule. The following substructures were considered: aromatic ring, 5-membered aromatic ring, 6-membered aromatic ring without heteroatoms (benzene ring), nitrogen-containing aromatic ring, primary, secondary, and tertiary amino groups, nitrile, hydroxyl, carbonyl, methoxy, methyl, and carboxyl groups. Three types of additional features were used: molecular weight and neutral loss spectra (molecular weight also allows for the neutral loss spectra computation), retention index for the non-polar stationary phase, and retention index for the polar stationary phase. A total of 8 feature sets were considered. In most cases, the molecular weight and neutral loss spectrum considerably improve the accuracy. Retention indices also allow for further accuracy increase. For polar functional groups such as carbonyl and hydroxyl, the effect of using retention indices is maximal. The use of retention indices for two stationary phases allows for the achievement of the best accuracy. The best accuracy of prediction was achieved for the benzene ring and aromatic ring, the worst (but still high) accuracy was observed for the secondary amino group. The achieved accuracy was compared with the previous results. In addition to the classification tasks, the regression tasks were considered. The gradient boosting models that predict the number of aromatic atoms, methyl groups, and benzene rings were developed. It was observed that the use of additional features considerably improves the accuracy in this case. Finally, it should be noted that the regression models underestimate the number of occurrences when the number is high.

Keywords: mass spectrometry, gas chromatography, non-target analysis, machine learning, gradient boosting.

Acknowledgments: the research is supported by the Russian Science Foundation (project No. 22-73-10053), <https://rscf.ru/project/22-73-10053/>

For citation: Matyushin D.D., Sholokhova A.Yu. Extraction of information about the molecule structure directly from GC-MS data. *Sorbtsionnye i khromatograficheskie protsessy*. 2023. 23(3): 373-383. (In Russ.). <https://doi.org/10.17308/sorpchrom.2023.23/11317>

Научная статья

Извлечение информации о структуре молекул непосредственно из данных ГХ-МС

Дмитрий Дмитриевич Матюшин, Анастасия Юрьевна Шолохова[✉]

Институт физической химии и электрохимии им. А.Н. Фрумкина РАН, Москва, Россия, shonastya@yandex.ru[✉]

Аннотация. Газовая хромато-масс-спектрометрия (ГХ-МС) – очень важный метод химического анализа. ГХ-МС можно использовать для нецелевого химического анализа и предварительного скрининга совершенно неизвестных соединений. В ГХ-МС обычно используется масс-спектрометрия с ионизацией электронами. Возможно извлечение информации непосредственно из данных ГХ-МС с использованием методов машинного обучения. Есть несколько ранее опубликованных работ, в которых модели машинного обучения извлекают информацию о наличии или отсутствии заданных подструктур в молекуле непосредственно из масс-спектра электронной ионизации. Дополнительные данные, такие как молекулярная масса и индекс удерживания, изредка используются вместе с масс-спектром в качестве входных данных для таких моделей, однако ранее не проводилось систематического сравнения того, как использование таких данных повышает точность предсказания. В этой работе для предсказания наличия или отсутствия заданных подструктур в молекуле использовался градиентный бустинг. Были рассмотрены следующие подструктуры: ароматическое кольцо, 5-членное ароматическое кольцо, 6-членное ароматическое кольцо без гетероатомов (бензольное кольцо), азотсодержащее ароматическое кольцо, первичные, вторичные и третичные аминогруппы, нитрил, гидроксил, карбонил, метокси, метил и карбоксильные группы. Использовались три типа дополнительных входных данных: молекулярная масса и спектры нейтральных потерь (молекулярная масса позволяет вычислять спектры нейтральных потерь), индекс удерживания для неполярной неподвижной фазы и индекс удерживания для полярной неподвижной фазы. Всего было рассмотрено 8 наборов входных данных. В большинстве случаев молекулярная масса и спектр нейтральных потерь значительно улучшают точность. Индексы удерживания также позволяют дополнительно повысить точность. Для полярных функциональных групп, таких как карбонил и гидроксил, эффект от использования индексов удерживания максимален. Использование индексов удерживания для двух стационарных фаз позволяет добиться наилучшей точности. Наилучшая точность предсказания достигнута для бензольного и ароматического колец, наихудшая (но все же высокая) – для вторичной аминогруппы. Достигнутая точность была сравнена с результатами из предыдущей работы. Помимо задач классификации были рассмотрены регрессионные задачи. Были разработаны модели на основе градиентного бустинга, которые предсказывают количество ароматических атомов, метильных групп и бензольных колец. Было замечено, что использование дополнительных входных данных значительно повышает точность и в этом случае. Наконец, следует отметить, что регрессионные модели недооценивают количество вхождений, когда это число велико.

Ключевые слова: масс-спектрометрия, газовая хроматография, нецелевой анализ, машинное обучение, градиентный бустинг.

Благодарности: исследование выполнено за счет гранта Российского научного фонда (проект № 22-73-10053), предоставленного Институту физической химии и электрохимии имени А.Н. Фрумкина Российской академии наук, <https://rscf.ru/project/22-73-10053/>

Для цитирования: Матюшин Д.Д., Шолохова А.Ю. Извлечение информации о структуре молекул непосредственно из данных ГХ-МС // *Сорбционные и хроматографические процессы. 2023. Т. 23, № 3. С. 373-383.* <https://doi.org/10.17308/sorpchrom.2023.23/11317>

Introduction

Gas chromatography – mass spectrometry (GC-MS) is a widely used analytical method for both targeted and non-targeted analysis of complex mixtures of volatile compounds. It is widely used in industry, pharmaceuticals, environmental analysis [1-2], and in metabolomics studies [3-4]. The most widely used approach for non-target GC-MS analysis is the library search in MS databases [5], such as the NIST 17 database [6]. Unfortunately, the majority of organic molecules are absent in all MS databases, and standard samples also are not available. Qualitative screening of such molecules is a complex task [7-8]. Fortunately, there are

software and machine learning approaches that facilitate this task. There are many tools for prediction of mass spectra from the structure of a molecule [8-10]. There are also tools for prediction of the presence or absence of substructures (fragments of a molecule) in a molecule based on the mass spectra [4, 11-16]. There are many such tools [17-18]. Some of them predict the so-called “molecular fingerprint” (a long vector of bits; each bit corresponds to the presence or absence of a structural feature) [4, 11, 12]. Others predict the presence or absence of specific fragments and functional groups that are of most interest to the researcher.

The accuracy of prediction of “molecular fingerprint” even by the most modern tools

is not very high. For example, in the work [11], for approximately half of the bits constituting the fingerprint, the accuracy is less than 0.9. Such a molecular fingerprint can be used for automated search in the database of possible candidates. Many works are devoted to prediction of the presence or absence of common substructures and functional groups. For example, in the work [13], the presence or absence of several structural features such as a benzene ring, a dimethylamine group, etc., was predicted. The observed classification accuracy lies in the range of 76-95 for all considered fragments (except trimethylsilyl groups). Another similar work is the well-known work by Varma et al. [14]. Several classifying models of common substructures were developed. The majority of such works [13, 14], including the MOLGEN-MS software [18], use only the mass spectrum as a source of features (Figure 1A).

In the work [17], the model predicts the presence or absence for many common functional groups, such as $-\text{NH}_2$, $-\text{OH}$, $=\text{O}$. In that work, unlike the majority of other works, the molecular weight (determined in a mass spectral experiment) is used for creation of features. In another work [15], the Golm Metabolome Database is used for training and validation, and the authors predict the presence or absence of many functional groups with relatively high probability. The authors use information about the retention index as an additional feature. The vast majority of these works provide only binary classifications: predicting the presence or absence of fragments, while determining the

number of occurrences can be a valuable task.

The aims of this work are (I) to study how the additional use of information about retention (on different stationary phases) and molecular weight (Figure 1B) together with the mass spectrum affect the accuracy of prediction, and (II) to consider not only classification, but also regression models that predict not only the presence or absence of a given structural feature but also the number of such features in a molecule.

Methods

The NIST 17 database was used for training and validation of developed models. Molecules containing elements other than H, B, C, N, O, F, P, S, Cl, Br, I were excluded from the data sets, as well as molecules with a molecular weight (MW) of more than 300 and molecules for which the retention index cannot be predicted using a 1D convolutional neural network (CNN) [19]. Spectra with peaks with m/z more than 300 and spectra without peaks with m/z less than 50 were excluded from the data set. The absence of peaks with m/z less than 50 highly likely means that the scanning range starts from high m/z values. Such spectra are not suitable enough for the considered task. The data set contained 132489 spectra. The initial data set was split into training, validation, and test data sets containing 105871, 13367, and 13251 mass spectra, respectively.

Gradient boosting was used for prediction of the number and presence of substructures. The models were trained using the XGBoost [20] library (version 1.5.1) using our own

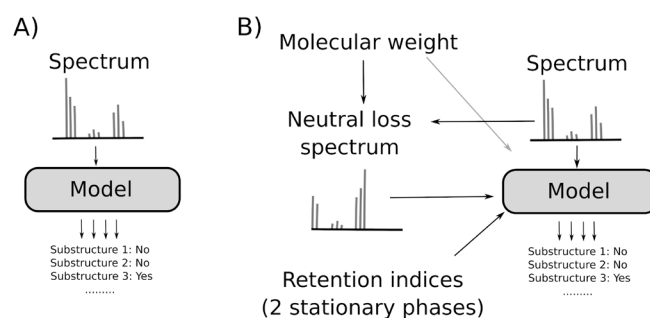


Fig. 1. Extraction of information about the structure from the electron ionization mass spectrum (A) and from all available GC-MS data (B)

Table 1. Hyperparameters of the XGBoost library used in this work

Parameter	Value
Learning rate (eta)	0.01
Minimum loss reduction required for a split (gamma)	0.01
L2 regularization term on weights (lambda)	7
Subsample ratio (subsample)	0.6
Maximum tree depth (maxDepth)	11
Minimum sum of weight needed (minChildWeight)	1
Objective function for classification tasks	Logistic regression for binary classification (binary:logitraw)
Objective function for regression tasks	Mean squared error (reg:squarederror)
Number of trees (classification tasks)	6000
Tree construction algorithm	Faster histogram algorithm (hist)

program (the Java Programming language was used). The used hyperparameters are given in Table 1. For regression tasks, the early stopping was used: if the accuracy for a validation set is not improved for 250 iterations, the training is stopped. The following features were used with the gradient boosting model: mass spectrum (scaled to a range of 0-1 intensities of peaks for each integer m/z in the range 1-300), neutral loss spectrum (see below), MW (divided by 1000), retention indices (RI) for polar (RI_{polar}) and non-polar ($RI_{\text{non-polar}}$) stationary phases (SP), as well as the difference between RI for polar and non-polar SP ($RI_{\text{polar}} - RI_{\text{non-polar}}$). The RI values were divided by 1000. The neutral loss spectrum is interrelated with the mass spectrum by the following equation:

$$N_n = \begin{cases} I_{M-n}, n < M \\ 0, n \geq M \end{cases} \quad (1)$$

where I_n – intensity of mass spectra corresponding to $m/z = n$; N_n – intensities of the neutral loss spectrum, M – MW of a molecule.

Because NIST 17 contains information about RI only for few molecules, RI predicted using 1D CNN were used as features. These RI values are close to the experimental ones [19]. For prediction of RI, 1D CNN with the following hyperparameters was used: 2 CNN layers with 300 output channels; 2 fully connected layers (kernel = 6) with 600 and 1 output nodes; rectified lin-

ear activation function was used for all layers except the linear output layer; early stopping using a validation set was used. More information about 1D CNN for retention index prediction is given in our previous work [19]. The mean and median absolute errors of prediction were 45.5 and 17.2, respectively, for non-polar SP. For polar SP, these values were 67.7 and 29.5, respectively. The error values are given for test sets. As the initial values of the trainable parameters (weights and biases) of the neural network for polar SP, the parameters obtained for non-polar SP were used.

The following accuracy measures were considered for binary classification tasks:

$$\text{True positive rate (TPR or recall)} = \frac{TP}{TP+FN} \quad (2)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3)$$

$$\text{False positive rate (FPR)} = \frac{FP}{FP+TN} \quad (4)$$

$$F_1 \text{ score } (F_1) = \frac{2 \cdot TP}{2 \cdot TP + FN + FP} \quad (5)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (6)$$

$$C = -\left(\frac{1}{N}\right) \sum_{i=1}^N Y_i \log(y_i) + (1 - Y_i) \log(1 - y_i) \quad (7)$$

where TN, FN, TP, FP – number of true negative, false negative, true positive, false positive predictions; y_i , Y_i – prediction and correct answer for the i -th sample; N – number of samples, C – binary cross-entropy (CE). For computation of TN, FN, TP, FP, the float predictions numbers were rounded up to 0

Table 2. Additional features that are used together with the raw mass spectrum

Additional features	Feature set							
	1	2	3	4	5	6	7	8
MW and neutral loss spectrum	-	+	-	+	-	+	-	+
RI _{non-polar}	-	-	+	+	-	-	+	+
RI _{polar}	-	-	-	-	+	+	+	+
RI _{polar} - RI _{non-polar}	-	-	-	-	-	-	+	+

Table 3. Area under the receiver operating characteristics curves (ROC-AUC) for various regression tasks and various sets of features (see Table 2)

Substructure	Feature set							
	1	2	3	4	5	6	7	8
-NH-	0.915	0.917	0.919	0.922	0.919	0.921	0.921	0.924
Aromatic 5-membered ring	0.909	0.913	0.914	0.917	0.913	0.916	0.917	0.920
-CH ₃	0.953	0.968	0.958	0.973	0.959	0.973	0.960	0.973
-NH ₂	0.932	0.952	0.937	0.958	0.938	0.956	0.940	0.957
Tertiary sp ³ nitrogen atom	0.967	0.968	0.968	0.969	0.968	0.969	0.969	0.970
Nitrile	0.929	0.957	0.928	0.956	0.934	0.960	0.940	0.963
Aromatic nitrogen	0.968	0.970	0.971	0.974	0.970	0.972	0.973	0.976
-O-CH ₃	0.972	0.982	0.972	0.982	0.973	0.982	0.974	0.982
-C(=O)-O- (carboxyl, ester, or anhydride)	0.957	0.978	0.959	0.980	0.961	0.980	0.965	0.983
Aromatic ring	0.993	0.994	0.995	0.995	0.995	0.995	0.995	0.995
Carbonyl	0.926	0.943	0.929	0.945	0.932	0.948	0.938	0.952
Hydroxyl	0.903	0.930	0.918	0.942	0.941	0.960	0.960	0.968
Benzene ring (6-membered aromatic ring without heteroatoms)	0.993	0.994	0.995	0.995	0.995	0.995	0.995	0.995

or 1. ROC (Receiver Operating Characteristics) curves were also considered, and the area under such curves (ROC-AUC) was used as an additional measure.

For regression tasks, the root mean square error (RMSE) was used. RMSE was used for early stopping for regression. The validation set was used for early stopping, all accuracy measures given below were calculated for the test set.

Results and discussion

In order to study how the use of various features (neutral loss spectrum, RI) affects the accuracy of prediction of the presence or absence of a given substructure in a molecule based on the electron ionization mass

spectrum, the series of computational experiments were conducted (see Table 2). For 13 substructures (see Table 3), the gradient boosting model was trained with 8 sets of features and the accuracy was evaluated. In all 8 cases, the intensities corresponding to m/z 1-300 in the raw mass spectrum were used as features with or without additional features. The area under the ROC curve [21] was considered as the primary accuracy measure. The XGBoost predictor predicts a float value in the range [0, 1] instead of a binary value. This value characterizes the probability of the presence or absence of the given fragment. The accuracy measures such as TPR (recall), precision, FPR (see equations (2)-(4)) depend on what is considered a threshold value above which the XGBoost

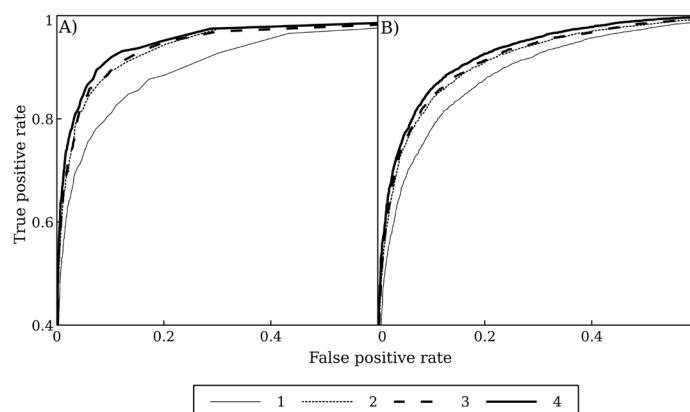


Fig. 2. ROC (Receiver Operating Characteristics) curves for classification models that predict the presence of nitrile (A) and carbonyl (B) groups. The curves denoted as 1-4 correspond to the feature sets (see Table 2) 1, 2, 4, 8, respectively

Table 4. Accuracy of prediction of the presence or absence of aromatic rings in a molecule from GC-MS data for different feature sets (see Table 2)

Accuracy measure	Feature set							
	1	2	3	4	5	6	7	8
Recall (TPR)	0.733	0.824	0.748	0.832	0.750	0.830	0.770	0.848
Precision	0.934	0.931	0.925	0.930	0.929	0.930	0.931	0.933
F ₁ score	0.821	0.874	0.827	0.878	0.830	0.877	0.843	0.889
Accuracy	0.908	0.931	0.909	0.933	0.911	0.933	0.917	0.938
Binary cross-entropy	0.235	0.172	0.229	0.167	0.226	0.167	0.213	0.154

prediction is considered as positive. The ROC-AUC measure [21] does not depend on the threshold value and characterizes such a predictor well. The perfect ROC curve passes through the points (0,0), (0,1), (1,1). The closer the ROC curve passes to the point (0,1), the closer it is to the perfect one.

As an example, Figure 2 shows ROC curves for prediction of the presence or absence of nitrile and carbonyl groups. Note that for better readability, the axis range in Figure 2 is not [0, 1]. Figure 2 clearly shows that the use of the neutral loss spectrum and molecular weight as additional features considerably improves the prediction accuracy. The use of RI_{non-polar} as an additional feature does not greatly improve the accuracy, but the use of information about RI for two SP causes an additional growth of the accuracy. Table 3 demonstrates that for all substructures, the use of MW and neutral loss spectrum considerably improves the accuracy, and the use of RI for two SP allows for the achievement of the best accuracy. In some

cases, the use of RI for only one SP gives the accuracy growth comparable to the use of RI for two SP (for example, -CH₃, -NH₂, benzene ring), in other cases, the use of RI for two SP gives considerably better accuracy (for example, -OH, aromatic nitrogen). The complete data are given in Table 3. For other than ROC-AUC accuracy measures, the situation is similar. For example, Table 4 shows various accuracy measures for prediction of the presence or absence of aromatic atoms in a molecule using various feature sets. The precision is almost constant for various feature sets, while the recall increases. It means that the overall accuracy improves. Note that for all accuracy measures except CE, the value 1.0 corresponds to the perfect model, and for CE, the value 0.0 corresponds to the perfect model.

In addition to the classification task: the prediction of the presence or absence of a given substructure, the regression task was also considered: the prediction of how many times a substructure is present in a molecule

Table 5. Accuracy of prediction of the number of various substructures in a molecule from GC-MS data using different feature sets (see Table 2)

Substructure	Feature set							
	1	2	3	4	5	6	7	8
Aromatic atoms	1.93	1.75	1.70	1.60	1.79	1.67	1.66	1.57
-CH ₃ groups	0.84	0.77	0.81	0.71	0.80	0.71	0.78	0.70
Benzene rings	0.28	0.27	0.26	0.26	0.26	0.26	0.26	0.25

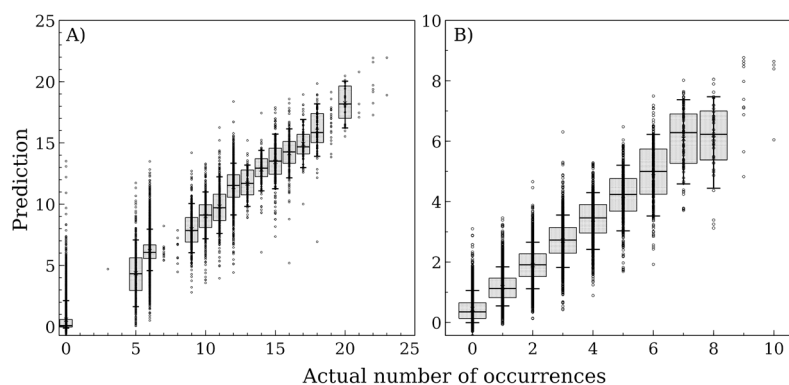


Fig. 3. Actual vs predicted plots for prediction of the number of aromatic atoms (A) and CH₃ groups (B) in a molecule. The boxes show the range containing the half of the predictions, and the whiskers demonstrate the ranges containing 90% of the predictions

for 3 substructures (see Table 5). In this case, the effect of additional features is even larger than for classification tasks. Figure 3 shows the actual vs predicted plots for prediction of the number of aromatic atoms and CH₃ groups in a molecule. The prediction accuracy is quite high for a low number of occurrences, however, Figure 3 clearly shows that in cases when the correct answer is high, this value is considerably underestimated. As an additional example, the ROC curve for prediction of the presence or absence of a benzene ring (6-membered ring containing only carbons) and the actual vs predicted plot for the number of benzene rings are shown in Figure 4. Figures 3-4 show the data for the full set of features.

We also tried to tune the XGBoost hyperparameters for each feature set and task separately and to make such comparisons using various sets of hyperparameters (other than given in Table 1). However, the same patterns are observed for various sets of hyperparameters, and the same hyperparameters are nearly optimal for various sets of features and various tasks. Taking this fact

into account, all comparisons were made using the same set of hyperparameters. The accuracy of classification was also compared with the accuracy given in the work of Stein et al. [17]. Because such measures as recall and precision depend on the probability threshold (and as the threshold increases, the precision increases and the recall decreases), the comparison was made at the fixed precision value of 0.9. Table 6 contains such a comparison. Since the work [17] does not use information about RI, we considered only feature sets 1 and 2 (see Table 2) in this comparison. It can be concluded that the considered classifiers in most cases have the same or better accuracy compared with the described in the work [17]. The worse accuracy was observed only for the worst predicted substructures: non-aromatic nitrogens. However, that work [17] uses a much older version of the NIST library, and this comparison is not completely correct. The correct comparison of the machine learning methods should be made using the same data set.

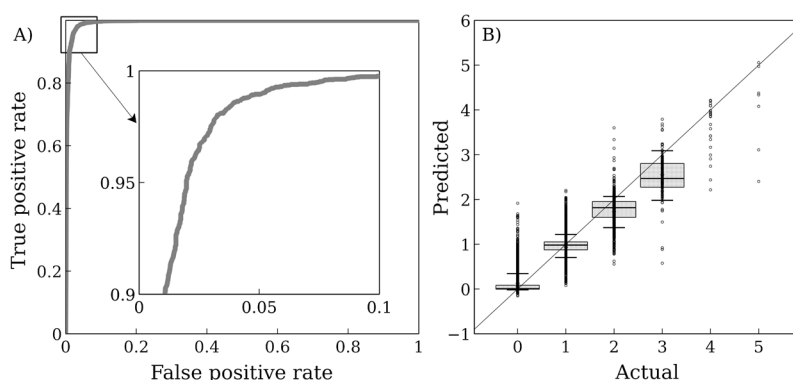
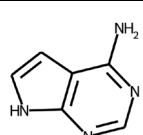
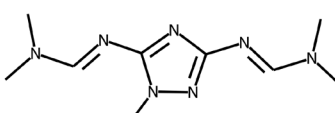
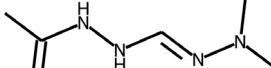


Fig. 4. ROC (Receiver Operating Characteristics) curve and the actual vs predicted plot for prediction of the presence (A) and number (B) of benzene rings in a molecule

Table 6. Comparison of the classification accuracy achieved in this work with the accuracy from the work [17]. The values of recall at precision = 0.9 are given

Substructure	Feature set (this work)		Previous results [17]	
	1 (without MW)	2 (with MW)	Without MW	With MW
Hydroxyl	0.42	0.56	0.40	0.58
Carbonyl	0.72	0.79	0.70	0.76
-C(=O)-O- (carboxyl, ester, or anhydride)	0.77	0.86	0.36	0.47
-NH-	0.13	0.14	0.28	0.28
-NH ₂	0.23	0.24	0.25	0.40
Aromatic ring	0.997	0.998	0.98	0.99
-O-CH ₃	0.76	0.84	0.49	0.74

Table 7. Application of models to UDMH transformation products

Structure	Methyl	Aromatic ring	Number of aromatic rings
	Absent	Present	2
	Present	Present	1
	Present	Absent	0

The models developed in this work were applied to the mass spectra and retention indices of the recently identified transformation products of unsymmetrical dimethylhydrazine (UDMH). These compounds are not available in mass spectral databases, and elucidating their structure using only

chromatography and mass spectrometry is a very difficult task [22]. Table 7 shows the structures of the three UDMH transformation products and model predictions for them. The first compound contains two conjugated aromatic rings. Previously, prior to the publication of work [22] by our team,



similar transformation products of UDMH were not known. The models developed in this work receive the retention indices for two stationary phases and mass spectra as an input and make it possible to obtain preliminary information about the structure and limit the number of possible candidates. This is an excellent starting point for further elucidation of the structure [22], for which the observed experimental data are consistent with the results of predicting mass spectra and retention indices. The prediction models developed in this work will be implemented in our previously published free and open-source software [22] that can be obtained by the following link: <https://github.com/mtshn/svekla>

Conclusions

A model that allows for the prediction of the presence and number of given substructures in a molecule based on GC-MS data was built using the XGBoost library. The use of additional data besides the electron ionization mass spectrum allows for the considerable improvement of the prediction accuracy. If the molecular weight is known in addition to the mass spectrum (for example, it

References

1. Ohoro C.R., Adeniji A.O., Okoh A.I., Okoh O.O., Distribution and Chemical Analysis of Pharmaceuticals and Personal Care Products (PPCPs) in the Environmental Systems: A Review, *International journal of environmental research and public health*. 2019; 16(17): 3026. <https://doi.org/10.3390/ijerph16173026>
2. Nika M. C., Alygizakis N., Arvaniti O. S., Thomaidis N. S., Non-target screening of emerging contaminants in landfills: A review, *Current Opinion in Environmental Science & Health*. 2023; 32: 100430. <https://doi.org/10.1016/j.coesh.2022.100430>
3. Beale D. J., Pinu F. R., Kouremenos K. A., Poojary M. M. et al., Review of recent developments in GC-MS approaches to metabolomics-based research, *Metabolomics*. 2018; 14(11): 152. <https://doi.org/10.1007/s11306-018-1449-2>

can be determined using a chemical ionization ion source), the prediction accuracy considerably improves. In this case, the neutral loss spectrum can be used as an additional set of features. The retention index is less important, but the use of the retention index allows for the considerable improvement of the prediction of the presence or absence of polar functional groups such as hydroxyl and carbonyl. The use of retention indices for two stationary phases allows for the greater improvement of the accuracy compared with a single retention index. For regression tasks, when models predict the number of occurrences rather than the presence or absence, the use of additional features considerably increases the accuracy.

Конфликт интересов

Авторы заявляют, что у них нет известных финансовых конфликтов интересов или личных отношений, которые могли бы повлиять на работу, представленную в этой статье.

4. Qiu F., Lei Z., Sumner L.W., MetExpert: An expert system to enhance gas chromatography–mass spectrometry-based metabolite identifications, *Analytica Chimica Acta*. 2018; 1037: 316-326. <https://doi.org/10.1016/j.aca.2018.03.052>
5. Vinaixa M., Schymanski E. L., Neumann S., Navarro M. et al., Mass spectral databases for LC/MS- and GC/MS-based metabolomics: State of the field and future prospects, *TrAC Trends in Analytical Chemistry*. 2016; 78: 23-35. <https://doi.org/10.1016/j.trac.2015.09.005>
6. Moorthy A.S., Wallace W.E., Kearsley A.J., Tchekhovskoi D.V., Stein S.E., Combining Fragment-Ion and Neutral-Loss Matching during Mass Spectral Library Searching: A New General Purpose Algorithm Applicable to Illicit Drug Identification, *Analytical chemistry*. 2017; 89(24): 13261-13268. <https://doi.org/10.1021/acs.analchem.7b03320>



7. Schymanski E.L., Meinert C., Meringer M., Brack W., The use of MS classifiers and structure generation to assist in the identification of unknowns in effect-directed analysis, *Analytica Chimica Acta*. 2008; 615(2): 136-147. <https://doi.org/10.1016/j.aca.2008.03.060>
8. Allen F., Pon A., Greiner R., Wishart D., Computational Prediction of Electron Ionization Mass Spectra to Assist in GC/MS Compound Identification, *Analytical chemistry*. 2016; 88 (15): 7689-7697. <https://doi.org/10.1021/acs.anal-chem.6b01622>
9. Wei J.N., Belanger D., Adams R.P., Sculley D., Rapid Prediction of Electron-Ionization Mass Spectrometry Using Neural Networks, *ACS central science*. 2019; 5(4): 700-708. <https://doi.org/10.1021/acscentsci.9b00085>
10. Zhu R.L., Jonas E., Rapid Approximate Subset-Based Spectra Prediction for Electron Ionization–Mass Spectrometry, *Analytical chemistry*. 2023; 95 (5): 2653-2663. <https://doi.org/10.1021/acs.anal-chem.2c02093>
11. Ji H., Deng H., Lu H., Zhang Z., Predicting a Molecular Fingerprint from an Electron Ionization Mass Spectrum with Deep Neural Networks, *Analytical chemistry*. 2020; 92(13): 8649-8653. <https://doi.org/10.1021/acs.anal-chem.0c01450>
12. Ljoncheva M., Stepišnik T., Kosjek T., Džeroski S., Machine learning for identification of silylated derivatives from mass spectra, *Journal of Cheminformatics*. 2022; 14(1): 62. <https://doi.org/10.1186/s13321-022-00636-1>
13. Yoshida H., Leardi R., Funatsu K., Varmuza K., Feature selection by genetic algorithms for mass spectral classifiers, *Analytica Chimica Acta*. 2001; 446(1-2): 483-492. [https://doi.org/10.1016/S0003-2670\(01\)00910-2](https://doi.org/10.1016/S0003-2670(01)00910-2)
14. Varmuza K., Werther W., Mass Spectral Classifiers for Supporting Systematic Structure Elucidation, *Journal of Chemical Information and Computer Sciences*. 1996; 36(2): 323-333. <https://doi.org/10.1021/ci9501406>
15. Hummel J., Strehmel N., Selbig J., Walther D., Kopka J., Decision tree supported substructure prediction of metabolites from GC-MS profiles, *Metabolomics*. 2010; 6(2): 322-333. <https://doi.org/10.1007/s11306-010-0198-7>
16. Xiong Q., Zhang Y., Li M., Computer-assisted prediction of pesticide substructure using mass spectra, *Analytica Chimica Acta*. 2007; 593(2): 199-206. <https://doi.org/10.1016/j.aca.2007.04.060>
17. Stein S.E., Chemical substructure identification by mass spectral library searching, *Journal of the American Society for Mass Spectrometry*. 1995; 6(8): 644-655. [https://doi.org/10.1016/S1044-0305\(05\)80054-6](https://doi.org/10.1016/S1044-0305(05)80054-6)
18. Meringer M., Schymanski E., Small Molecule Identification with MOLGEN and Mass Spectrometry, *Metabolites*. 2013; 3(2): 440-462. <https://doi.org/10.3390/metabo3020440>
19. Matyushin D.D., Sholokhova A.Yu., Buryak A.K., A deep convolutional neural network for the estimation of gas chromatographic retention indices, *Journal of Chromatography A*. 2019; 1607: 460395. <https://doi.org/10.1016/j.chroma.2019.460395>
20. Chen T., Guestrin C., XGBoost: A Scalable Tree Boosting System, 2016, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco California USA: ACM, pp. 785-794. <https://doi.org/10.1145/2939672.2939785>
21. Jin Huang, Ling C.X., Using AUC and accuracy in evaluating learning algorithms, *IEEE Transactions on Knowledge and Data Engineering*. 2005; 17(3): 299-310. <https://doi.org/10.1109/TKDE.2005.50>
22. Sholokhova A.Y., Matyushin D.D., Grinevich O.I., Borovikova S.A., Buryak A.K., Intelligent Workflow and Software for Non-Target Analysis of Complex Samples Using a Mixture of Toxic Transformation Products of Unsymmetrical Dimethylhydrazine as an Example, *Molecules*.



2023; 28(8): 3409.
<https://doi.org/10.3390/molecules28083409>

Информация об авторах / Information about the authors

Д.Д. Матюшин – н.с. лаборатории физико-химических основ хроматографии и хромато-масс-спектрометрии, Институт физической химии и электрохимии имени А.Н. Фрумкина РАН, Москва, Россия

А.Ю. Шолохова – с.н.с. лаборатории физико-химических основ хроматографии и хромато-масс-спектрометрии, Институт физической химии и электрохимии имени А.Н. Фрумкина РАН, Москва, Россия

D.D. Matyushin – researcher, laboratory of physicochemical principles of chromatography and chromatography – mass spectrometry; Institute of Physical chemistry and electrochemistry, Moscow, Russian Federation, e-mail: dm.matiushin@mail.ru

A.Yu. Sholokhova – researcher, laboratory of physicochemical principles of chromatography and chromatography – mass spectrometry; Institute of Physical chemistry and electrochemistry, Moscow, Russian Federation, e-mail: shonastya@yandex.ru

Статья поступила в редакцию 27.03.2023; одобрена после рецензирования 20.04.2023; принята к публикации 25.04.2023.

The article was submitted 27.03.2023; approved after reviewing 20.04.2023; accepted for publication 25.04.2023.