JOURNAL OF PHYSICAL CHEMISTRY, 2025, Vol. 99, No. 1, c. 133–141

PHYSICAL CHEMISTRY OF SEPARATION PROCESSES. CHROMATOGRAPHY

APPLYING MOLECULAR SIMILARITY TO ASSESS THE PREDICTION ACCURACY OF GAS CHROMATOGRAPHIC RETENTION INDICES USING DEEP LEARNING

D. D. Matyushin[a], A. Yu. Sholokhova[a, *], M. D. Khrisanfov[a,b], and S. A. Borovikova[a]

[a] *Frumkin Institute of Physical Chemistry and Electrochemistry of the Russian Academy of Sciences, Moscow, 119071 Russia*

[b] *Lomonosov Moscow State University, Department of Chemistry, Moscow, 119991 Russia*

**e-mail: shonastya@yandex.ru*

Received March 25, 2024
Revised May 22, 2024
Accepted May 24, 2024

Abstract. When predicting retention indices using deep learning, there is usually no way to assess the reliability of the prediction for a particular molecule. In this work, using stationary phases based on polyethylene glycol and the NIST 17 database as an example, it is shown that, on average, the closer the molecule in the training data set is to the compound being predicted, the more accurate the prediction. Tanimoto similarity of "molecular fingerprints" ECFP is the most appropriate molecular similarity calculation algorithm for this problem among the four considered. It is shown that for a number of transformation products of unsymmetrical dimethylhydrazine, whose structure was established using this prediction, it could be very unreliable.

Keywords: *gas chromatography, retention indices, machine learning, deep learning, molecular similarity*

DOI: 10.31857/S00444537250114e7

INTRODUCTION

The retention time in gas chromatography depends on the flow rate of the carrier gas, the geometric parameters of the chromatographic column, the temperature program and other factors. At the same time, the retention index [1] characterizing the retention time of the substance relative to the retention times of *n*-alkanes depends mainly on the structure of the retained compound and the chemical nature of the stationary phase [1–3]. Thus, the task of predicting the retention index for the given molecule and the given stationary phase is the task of predicting one single number by the structure of the molecule.

In chromatography-mass spectrometric analysis of a complex mixture containing unknown components, the assumption of the structure of an unknown compound is made on the basis of the mass spectrum, most often by library search [4, 5]. However, library search often yields an incorrect result, even if the compound in question is contained in the database [6]. When there are no compounds to be identified in the databases, the task becomes even more complicated [7]. However, the comparison of the observed retention index with that predicted by machine learning allows to discard incorrect candidates [6, 8, 9] and confirm preliminary identification [9–12]. The use of retention indices significantly increases the identification reliability [6, 9]. Experimental data on retention indices are only available for about one hundred thousand molecules [13], which is several times less than the number of molecules for which experimental mass spectra are available and several orders of magnitude less than the total number of known molecules. Thus, prediction of retention indices is an important task for modern chemistry.

Deep learning, i.e., a totality of statistical methods based on deep neural networks, has revolutionized many areas of science and technology in recent years. Deep neural networks are used for a variety of tasks from analytical chemistry [14] to machine vision and machine translation tasks [15]. In particular, deep learning is used to predict gas chromatographic retention indices [13, 16–20] by the molecule structure. In recent years, a number of models of this type have been developed [18]. Deep learning is significantly superior in accuracy to previously used models [16, 17]. In a number of works [9–12], such predicted retention indices are used to clarify identification.

Estimation of accuracy of models that predict retention indices is carried out using large data sets

and "average" accuracy metric is calculated for the entire data set [16–20] (for instance, standard or average absolute deviation). However, this makes it completely impossible to assess whether the prediction for a particular individual molecule is accurate. In some works, accuracy is calculated for individual classes of compounds [18, 19]; however, in this case, the classes (for instance, "aromatic compounds", "trimethylsilyl derivatives") are also quite wide and include a variety of molecules. In this regard, it is relevant to develop methods that can help assess whether the prediction of the retention index for a given molecule is reliable, i.e., methods to assess whether the prediction is trustworthy. The use of predicted retention indices can lead to incorrect results if it is for the molecules in question that prediction is highly unreliable. Recently, an approach has been developed for this task that compares predictions made using several independent models [21].

There are various methods to quantify how close the structures of the two molecules are, i.e., to estimate molecular similarity [22–25]. In particular, the similarity of so-called "molecular fingerprints" [25] (binary vectors, each bit of which shows whether a fragment is contained in the molecule) can be used, as well as finding a common subgraph between two molecules [22].

The purpose of this work is to study how molecular similarity between the molecule for which the retention index prediction is performed by deep learning and the molecules contained in the training dataset used to train the model affects the accuracy of the retention index prediction. This study is performed on the example of retention indices for polar stationary phases (Standard Polar type in the NIST database; polyethylene glycol and approximately chromatographically equivalent polymers based on it) and a previously published deep learning model embedded in the SVEKLA software [9, 16]. In addition, the purpose of this work is a preliminary assessment of whether the predictions of retention indices made in [9] and used to construct the structure of new transformation products of unsymmetric dimethylhydrazine are reliable.

METHODS

Dataset and Deep Learning Model

NIST 17 was used as the dataset. The data processing and preparation procedure is described in the previous work [16]. The dataset was divided into five sets randomly. Model training was performed five times, each time four sets were used as training ones and the fifth set was used as a test one. The prediction results for the test sets (each time, the training set lacks the compound for which prediction is performed) were combined and used for further work (5-fold cross-validation).

Two models were trained, viz. a one-dimensional convolutional neural network and a deep multilayer perceptron. Detailed descriptions of models are given in [13, 16]. Transfer training was used, viz. first, neural networks were trained to predict retention indices for non-polar stationary phases, and then the obtained weights of neural networks were used as initial values to train the model to predict retention indices for polar stationary phases. The molecules included in the test set were each time removed from the retention index dataset for the polar stationary phases used for training. Thus, there was no "data leakage", i.e., the molecules used for testing were not used in training at any stage. The training procedure is described in detail in the previous work [16].

The NIST 17 database contains several data records for each of the molecules. All these records were used in training and testing (they differ in which chromatographic column was used, as well as in measurement conditions). After the cross-validation procedure is performed, there is a pair of values for each record, viz. the experimental retention index and that predicted using a model that "did not see" this molecule during training. The initial database was divided into five sets so that all records for each of the molecules were placed in one of the sets selected randomly. Geometric isomers and stereoisomers were considered as one molecule. A more detailed description of procedures and algorithms is contained in previously published works [13, 16–17].

Calculating Molecular Similarity

The original dataset contained 89,086 records, each containing a molecule structure, a reference, and a predicted retention index. For each structure, the median value of the reference retention index was found. Thus, a data set containing 9,408 records consisting of molecule structure, reference and predicted value was obtained. Each molecule occurs exactly once in the given set.

For each molecule, "molecular fingerprints" (vectors showing the presence of certain fragments) were calculated using the ECFP algorithm [25] (radius 3, vector length 1024). For each pair of molecules, the Tanimoto similarity of "molecular fingerprints" is calculated

$$S = \frac{N_A + N_B}{N_A + N_B - N_{AB}}, \quad (1)$$

where N_A and N_B are the numbers of non-zero bits in the "molecular fingerprints" of each molecule, and N_{AB} is the number of bits that are non-zero in each of the two molecular fingerprints at the same time. For each molecule, 100 closest structures (having the highest value of molecular similarity S) included in the training dataset were selected when training the model used to predict the retention index for the molecule in question. Then, four methods for calculating molecular similarity were considered. For each of the methods, a molecular similarity value is obtained for the molecule included in the training dataset when training the model used to predict the retention index for the molecule in question and having the highest molecular similarity value with the molecule in question. This value is denoted as $S_{\max}$. Since these methods are more resource intensive, the search for the molecule with the highest molecular similarity value was performed for only 100 pre-selected candidates.

The first molecular similarity calculation method designated by MCS was to calculate the largest common fragment using the RDKit library, rdFMCS.FindMCS method. After this fragment was found, similarity was calculated using a formula similar to Eq. (1)

$$S = \frac{M_A + M_B}{M_A + M_B - M_{AB}}, \qquad (2)$$

where M_A and M_B are the numbers of atoms of each molecule, and M_{AB} is the number of atoms in the largest common fragment. Note that only the type of atoms and the structure of the molecular graph are taken into account. Hydrogen atoms are not considered.

The second method was the Rascal similarity calculation. This also calculates the largest common fragment by the Rascal algorithm [22] and the number of bonds and atoms in this fragment. The similarity is calculated by the following equation

$$S = \frac{\left(M_{AB} + B_{AB}\right)^2}{\left(M_A + B_A\right)\left(M_B + B_B\right)}, \qquad (3)$$

where M_A and M_B are the numbers of atoms of each molecule, B_A and B_B are the numbers of bonds in each molecule, M_{AB} and B_{AB} are the numbers of atoms and bonds in the largest common fragment, respectively. This method used the rdRascalMCES module of the RDKit library.

The third and fourth methods were designated by RDKitFP and ECFP. They calculated the similarity of "molecular fingerprints" by formula (1). Molecular descriptors calculated using the GetRDKitFPGenerator and GetMorganGenerator classes, respectively, were used. The length of the vector was considered to be 4096, and the radius (for ECFP) was taken to be 6. The ECFP method corresponds to "circular molecular fingerprints" [25]. The maxPath parameter for RDKitFP was also set to 6.

DISCUSSION OF RESULTS

Molecular Similarity and Accuracy of Prediction of Retention Indices

During cross validation, the original dataset (the NIST 17 database) was divided into five subsets. Each molecule from the NIST 17 database, for which the experimental value of the polar stationary phase retention index is available, had the molecule closest to it found (in four ways), i.e., the one having the highest value of the molecular similarity measure and included in another subset of the dataset. The hypothesis tested in this work is that the molecular similarity $S_{\max}$ between the molecule for which prediction is performed and its closest molecule from the training set is related to the prediction accuracy.

Figure 1 shows the distribution of molecules (the number of molecules in the respective range (bin) is designated by N) from the dataset involved by the value $S_{\max}$ for four molecular similarity calculation methods. Light grey shows molecules, for which the absolute prediction error using the algorithm [16] in question is not more than 100, and dark grey shows those for which the absolute prediction error is greater than 100. In what follows, we call such molecules "poorly predicted". The value 100 was used as a threshold since such a value was used in previous work [9] to discard false candidates in the analysis of the complex mixture. Thus, if the candidate structure in question is "poorly predicted", it may be falsely discarded (or vice versa not discarded) based on a comparison of the observed and predicted retention indices for the polar stationary phase.

As we can see from Fig. 1, when using the ECFP molecular similarity calculation method, the value $S_{\max}$ of the largest number of molecules is about 0.5. The median value $S_{\max}$ for all molecules is 0.53 in this case. For molecules with the values $S_{\max}$ less than ~0.5, the proportion of "poorly predicted" molecules is significantly higher than for others. For the RDKitFP molecular similarity calculation method, the median value $S_{\max}$ for all molecules is significantly higher and equals 0.89. Although most molecules have quite high values $S_{\max}$, in this case there is a similar trend as well — the number of "poorly predicted" molecules decreases significantly slower as $S_{\max}$ decreases as compared to the total number of molecules. For the MCS and Rascal molecular

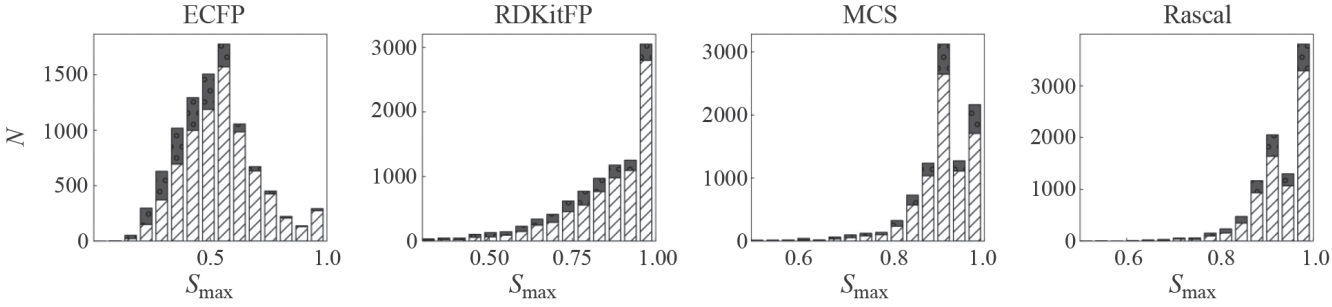

Fig. 1. Distribution of the number of molecules N in the NIST 17 retention index database (polar stationary phases) according to the values S_{max} (maximum molecular similarity value for all pairs including the molecule in question and the molecules from the training set) for the four molecular similarity calculation methods. Dark grey indicates "poorly predicted molecules" (the absolute prediction error is greater than 100), and light grey indicates the remaining molecules.

similarity calculation methods based on comparison of molecular graphs rather than "molecular fingerprints", the trend is less pronounced. For all molecular similarity calculation methods in the region of the smallest values S_{max}, most molecules are classified as "poorly predicted".

We can see that for all methods but for RDKitFP, the distribution of molecules by S_{max} has a pronounced bimodal character. For all methods, there are a significant number of molecules that have a very similar molecule in the training set, such as a homologue. In the case of the MCS algorithm, the molecular similarity between, for instance, cyclohexene and cyclohexane is 1.0: one double bond in the cycle is ignored since a common subgraph including all bonds between carbon atoms except this one includes all non-hydrogen atoms. This and other features of the algorithm result in a number of very different chemical molecules having a molecular similarity of 1.0. For the Rascal algorithm, a very high molecular similarity value is also possible for highly different molecules. For instance, 1-eicosanol and eicosanoic acid have the molecular similarity value 0.95 while this value is 0.52 when the RDKitFP method is used and 0.39 when ECFP is used. At the same time, ECFP gives a similarity of 1.0 for homologues containing a long sequence of carbon atoms, for instance for eicosanol and docosanol.

Figure 2 clearly shows how the proportion of "poorly predicted" (the average absolute error is greater than 100) molecules depends on S_{max}. For all methods but for MCS, this proportion grows rapidly with the decreasing S_{max}. Thus, small values S_{max} indicate that it is likely that the prediction for the molecule in question is very inaccurate. For all methods but for ECFP, the total number of molecules (also shown in Fig. 2 for convenience) in the respective range drops rapidly with a drop in the value S_{max}. In general, Figs. 1 and 2 show that ECFP is the best algorithm for calculating molecular similarity for this task.

In Figs. 1 and 2 and in the following sections, the proportion of "poorly predicted" compounds, i.e., compounds for which the absolute prediction error is greater than 100, is mainly discussed. Nevertheless, it is interesting to consider the error distribution for different ranges of S_{max}. Such absolute error distributions are shown in Fig. 3 for the ECFP and RDKitFP algorithms. In the case of ECFP, we can see that if $S_{max} > 0.9$, the vast majority of absolute error values do not exceed 50 while for absolute error values greater than 100, molecules with $S_{max} < 0.5$ begin to dominate. There are similar patterns for the RDKitFP algorithm.

Quantitative Comparison of Molecular Similarity Calculation Methods

If some molecular similarity value is used as a threshold, molecular similarity can be used as the simplest predictor of whether a given molecule is "poorly predicted". If the threshold value changes from 0 to 1, the prediction sensitivity (the proportion of identified "poorly predicted" molecules among all "poorly predicted" molecules) will increase and the specificity will decrease. Thus, it is possible to construct a receiver operator characteristic (ROC) curve [26, 27] characterizing the reliability of a given molecular similarity metric when used as a predictor. The area under this curve is [27] a metric of the accuracy of such a predictor.

Table 1 shows the area under the curve for various molecular similarity algorithms. At the same time, unlike Figs. 1 and 2, the RDKitFP and ECFP algorithms with different values of the maxPath and radius parameters were considered in this case. Table 1 gives the area under the curve for different values of these parameters.

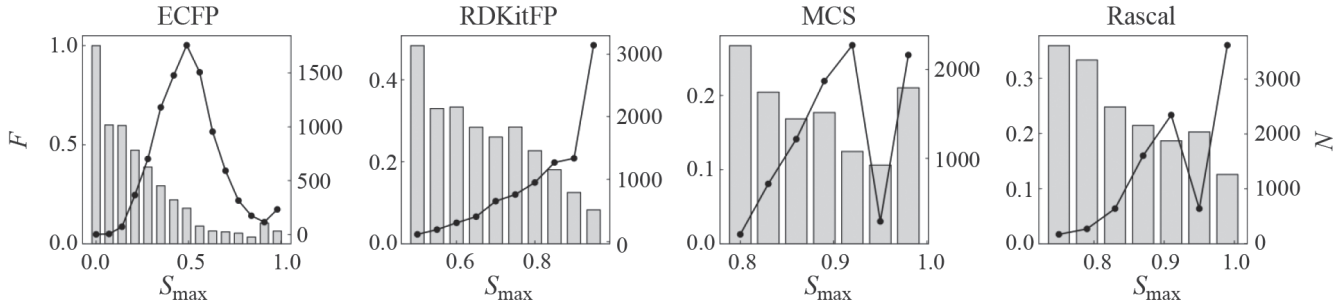

Fig. 2. Dependence of the total number of molecules N (solid circles and lines) and the fraction of "poorly predicted molecules" (the absolute prediction error is greater than 100) F (rectangles) on the value S_{max} (maximum molecular similarity value for all pairs including the molecule involved and molecules from the training set).

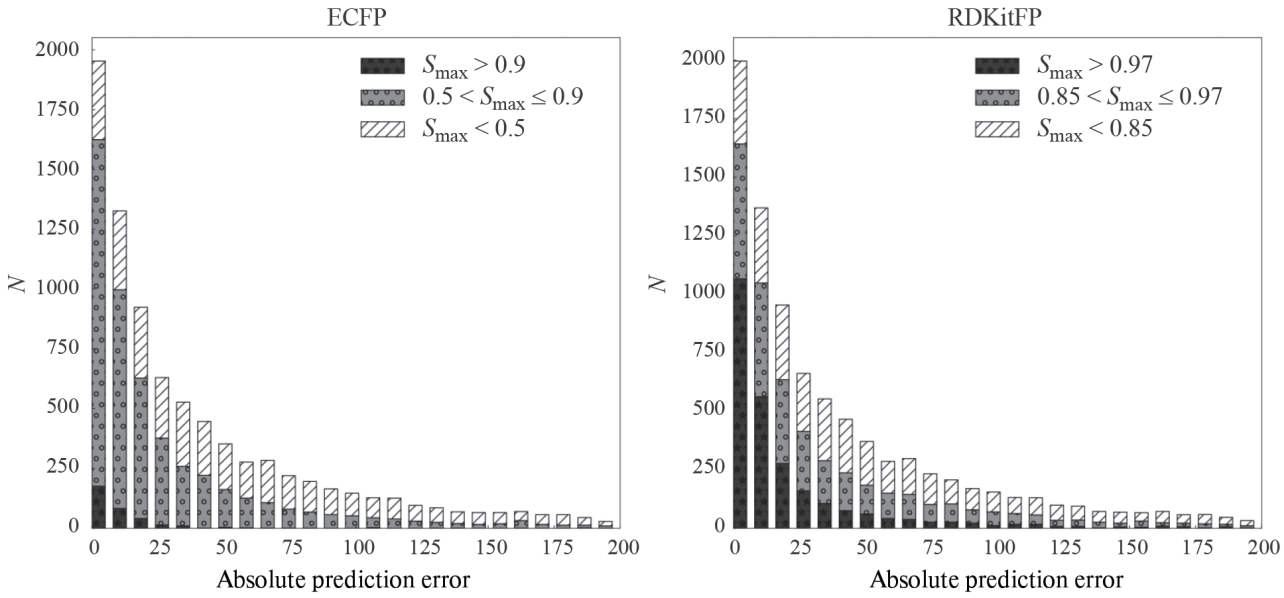

Fig. 3. Distribution of the number of molecules N over the absolute prediction error for different values S_{max} (the maximum molecular similarity value for all pairs that include the molecule involved and the molecules from the training set) for the two molecular similarity calculation methods.

The maxPath (RDKitFP "molecular fingerprints") and radius (ECFP "molecular fingerprints") parameters characterize the size of the substructures to which the "molecular fingerprint" bits correspond. The higher the values of these parameters, the larger substructures are considered.

Table 1 shows that the ECFP algorithm is best suited for this purpose, and there is practically no dependence on the radius parameter. The RDKitFP algorithm (when maxPath is 6 and higher) yields worse results. Other algorithms produce unreliable results. Note that the value of the area under the curve 0.5 corresponds to a random classifier, and the value 1 corresponds to the ideal classifier [26]. A value of 0.7 is sometimes considered the lowest acceptable [27]. ROC curves for molecular similarity calculation algorithms based on "molecular fingerprints" are given in Fig. 4. One can see that RDKitFP with maxPath = 3 fails to give any satisfactory accuracy, and ECFP exceeds RDKitFP.

Reliability of Identification of a Number of Nitrogen-Containing Compounds

Unsymmetrical dimethylhydrazine (UDMH) is a toxic compound used as a rocket fuel. When stored in an uncontrolled way and released into the environment, this compound forms a variety of transformation products [9, 12, 28], many of which are no less toxic than UDMH itself [12]. Studying UDMH transformation

Table 1. Area under the ROC curve when using different molecular similarity metrics as a predictor of whether a molecule is "poorly predicted" or not

Method	Area under the curve
RDKitFP (maxPath = 3)	0.62
RDKitFP (maxPath = 6)	0.69
RDKitFP (maxPath = 12)	0.70
RDKitFP (maxPath = 15)	0.69
ECFP (radius = 3)	0.72
ECFP (radius = 6)	0.72
ECFP (radius = 12)	0.72
MCS	0.55
Rascal	0.61

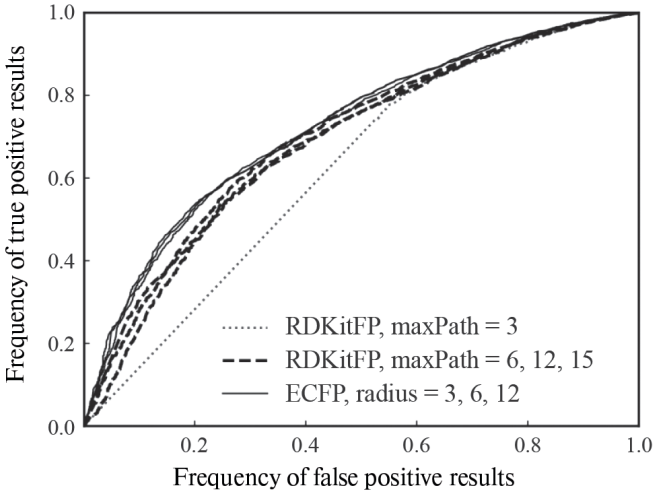

Fig. 4. ROC curves (specificity-sensitivity curves) for predicting whether a molecule is "poorly predicted" (the absolute prediction error is greater than 100) using different molecular similarity calculation algorithms. Curves for algorithms, for which the area under the curve differs by no more than 0.02, are labeled with a single line type for readability.

products is an important task. The structures of most transformation products are still unknown [9]. Various chromatography-mass spectrometry techniques are used to pre-determine the structures of the UDMH transformation products in complex mixtures. Recently, work [9] has been published that used prediction of retention indices in the polar stationary phase to confirm the structures of unknown UDMH transformation products. If the difference between the observed and predicted index exceeded 100, the candidate structure was discarded.

A total of 1,754 of 9,408 (19%) molecules in the dataset are "poorly predicted". However, among molecules with $S_{max} < 0.5$ (the ECFP algorithm), 31% are "poorly predicted". Table 2 shows the number of "poorly predicted" molecules for different ranges of S_{max} and the mean and median absolute error values for these ranges. The error values in Table 2 differ from those in the previous work [16], where exactly the same model was used, due to the fact that errors in [16] were calculated for all records of the NIST database, and the data in this work was previously averaged over all records for each compound. Thus, the contribution of those compounds that have many records in the NIST database to the average absolute error has been significantly reduced since if there are many records for the compound, the error modulus is calculated for each record and all these values (for each of the records for all compounds) are averaged. The average absolute error is the quotient of the sum of error modules and the number of records or molecules. In this work, each compound corresponds to one sum of error modules, as opposed to [16]. We noted that most compounds, for each of which the NIST database contains multiple records, have a relatively simple structure and, on average, predictions for such compounds are more accurate than for all compounds. In the NIST database, most structures have exactly one record, but for some structures (for instance, molecules such as benzene and ethanol), the database contains many records. The difference in the approach to calculating the average absolute error leads to the difference in the values given in [16] and in Table 2.

Figure 5 shows the structures proposed in [9] as structures of the UDMH transformation products using the prediction of the polar stationary phase retention indices. For each of the structures, molecular similarity (the ECFP algorithm) with the closest structure from the NIST database is shown. Two structures are contained in the NIST database — this value is 1.0 for them. For other structures, this value does not exceed 0.5. For a number of structures, this value is even less than 0.2. Thus, one cannot be sure that such predictions lead to correct results, and the results presented should be treated with caution. Nevertheless, structures confirmed by several methods of chromatography-mass spectrometry (gas and liquid) and several machine learning methods can be considered as sufficiently reliable [9, 12].

Table 2. Number of "poorly predicted" molecules and accuracy metrics for different ranges of S_{max} (the ECFP algorithm)

Range	"Poorly predicted" molecules	Molecules in total	Proportion of "poorly predicted" molecules, %	Mean absolute error	Median absolute error
All molecules	1754	9408	18.6	70.6	28.4
$S_{max} > 0.7$	83	1419	5.8	25.8	8.3
$S_{max} > 0.5$	512	5239	9.8	41.7	15.9
$S_{max} < 0.5$	1168	3820	30.6	109.9	57.0
$S_{max} < 0.3$	280	583	48.0	173.6	95.4
$S_{max} < 0.2$	50	30	60	311.0	181.7

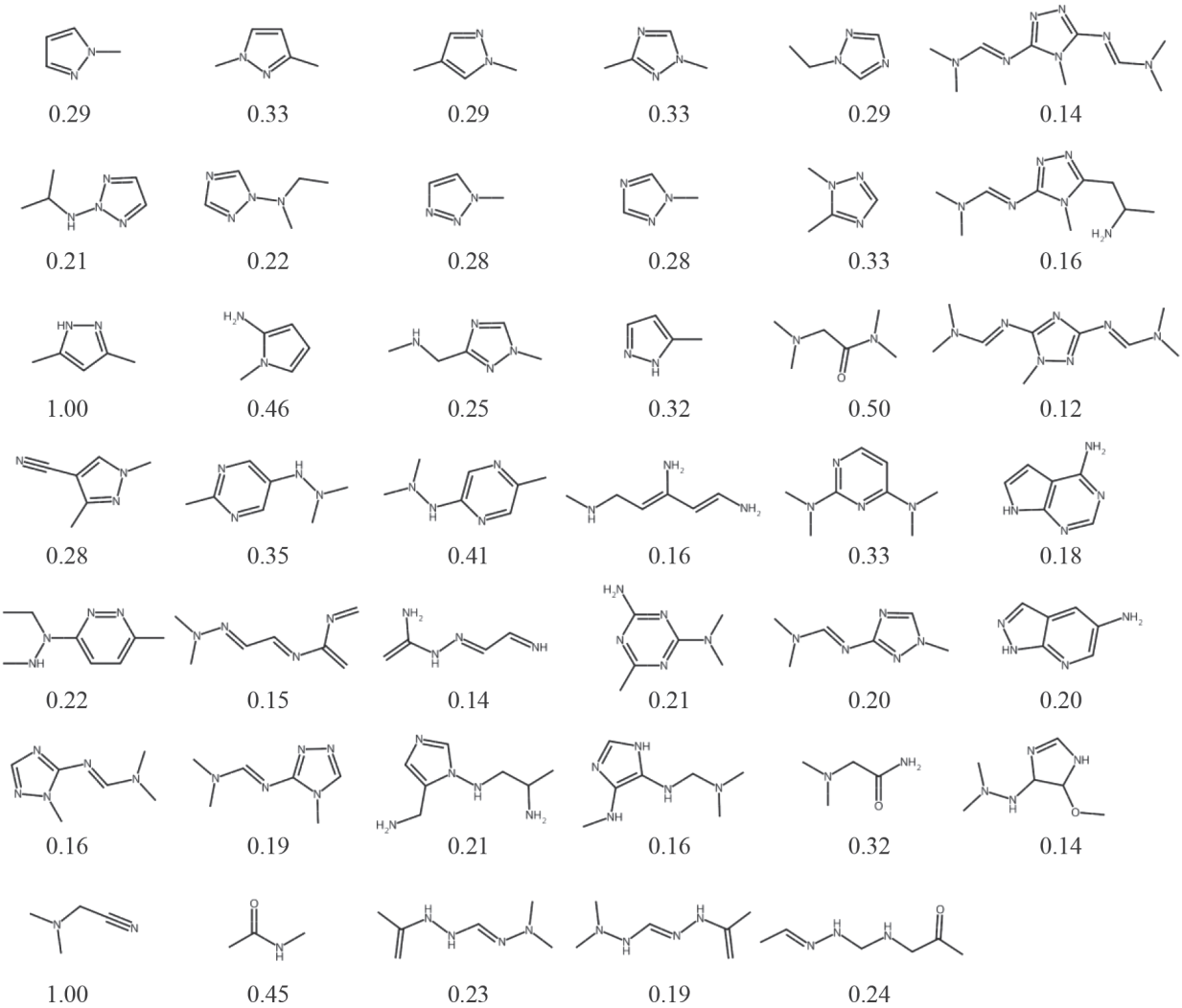

Fig. 5. Structures of the transformation products of unsymmetrical dimethylhydrazine proposed in [9] and the values S_{max} (the value of molecular similarity between the molecule involved and its closest molecule from the training set) for each of them. The molecular similarity calculation method is ECFP.

CONCLUSIONS

The accuracy of models predicting gas chromatographic retention indices is estimated using metrics such as the mean absolute error, which, however, do not allow estimating the accuracy for particular molecules. In this work, we showed that the fact that there are molecules in the training set that are close in structure to the molecule whose retention index is to be predicted greatly increases the probability that the prediction for this molecule will be accurate. The most suitable way to assess molecular similarity for this task is to use ECFP "molecular fingerprints". When the prediction of retention indices is used to construct structures of unknown chemical compounds, one needs to estimate the accuracy of prediction in one way or another. Thus, for instance, in one of the works that study transformation products of unsymmetrical dimethylhydrazine [9], the training dataset lacked molecules with high values of molecular similarity measure for most of the considered structures. Therefore, the conclusions made by predicting retention indices for these structures may not be entirely reliable. The source code of the scripts used to perform this work is available online: https://github.com/mtshn/molsimwax

FUNDING

The work was carried out with the support of the Russian Science Foundation (project no. 22-73-10053, https://rscf.ru/project/22-73-10053/).

REFERENCES

1. *Tarján G., Nyiredy S., Györ M. et al.* // J. of Chromatography A. 1989. Vol. 472. P. 1. https://doi.org/10.1016/S0021-9673(00)94099-8
2. *Franke J.-P., Wijsbeek J., De Zeeuw R.A.* // J. of Forensic Sciences. 1990. Vol. 35. No. 4. P. 813. https://doi.org/10.1520/JFS12893J
3. *Zellner B.A., Bicchi C., Dugo P. et al.* // Flavour and Fragrance J. 2008. Vol. 23. No. 5. Pp. 297–314. https://doi.org/10.1002/ffj.1887
4. *Milman B.L., Zhurkovich I.K.* // TrAC Trends in Analytical Chemistry. 2016. Vol. 80. Pp. 636–640. https://doi.org/10.1016/j.trac.2016.04.024
5. *Vinaixa M., Schymanski E.L., Neumann S. et al.* // TrAC Trends in Analytical Chemistry. 2016. Vol. 78. P. 23. https://doi.org/10.1016/j.trac.2015.09.005
6. *Matyushin D.D., Sholokhova A.Yu., Karnaeva A.E. et al.* // Chemometrics and Intelligent Laboratory Systems. 2020. Vol. 202. P. 104042. https://doi.org/10.1016/j.chemolab.2020.104042
7. *Schymanski E.L., Meringer M., Brack W.* // Analytical Chemistry. 2011. Vol. 83. No. 3. P. 903. https://doi.org/10.1021/ac102574h
8. *Dossin E., Martin E., Diana P. et al.* // Analytical Chemistry. 2016. Vol. 88. No. 15. Pp. 7539–7547. https://doi.org/10.1021/acs.analchem.6b00868
9. *Sholokhova A.Yu., Matyushin D.D., Grinevich O.I. et al.* // Molecules. 2023. Vol. 28. No. 8. P. 3409. https://doi.org/10.3390/molecules28083409
10. *Su Q.-Z., Vera P., Salafranca J. et al.* // Resources, Conservation and Recycling. 2021. Vol. 171. P. 105640. https://doi.org/10.1016/j.resconrec.2021.105640
11. *Su Q.-Z., Vera P., Nerín C. et al.* // Resources, Conservation and Recycling. 2021. Vol. 167. P. 105365. https://doi.org/10.1016/j.resconrec.2020.105365
12. *Sholokhova A.Yu., Grinevich O.I., Matyushin D.D. et al.* // Chemosphere. 2022. Vol. 307. P. 135764. https://doi.org/10.1016/j.chemosphere.2022.135764
13. *Matyushin D.D., Buryak A.K.* // IEEE Access. 2020. Vol. 8. P. 223140. https://doi.org/10.1109/ACCESS.2020.3045047
14. *Debus B., Parastar H., Harrington P. et al.* // TrAC Trends in Analytical Chemistry. 2021. Vol. 145. P. 116459. https://doi.org/10.1016/j.trac.2021.116459
15. *Dong S., Wang P., Abbas K.* // Computer Science Review. 2021. Vol. 40. P. 100379. https://doi.org/10.1016/j.cosrev.2021.100379
16. *Matyushin D.D., Sholokhova A.Yu., Buryak A.K.* // Intern. J. of Molecular Sciences. 2021. Vol. 22. No. 17. P. 9194. https://doi.org/10.3390/ijms22179194
17. *Matyushin D.D., Sholokhova A.Yu., Buryak A.K.* // J. of Chromatography A. 2019. Vol. 1607. P. 460395. https://doi.org/10.1016/j.chroma.2019.460395
18. *Anjum A., Liigand J., Milford R. et al.* // J. of Chromatography A. 2023. Vol. 1705. P. 464176. https://doi.org/10.1016/j.chroma.2023.464176
19. *Qu C., Schneider B.I., Kearsley A.J. et al.* // J. of Chromatography A. 2021. Vol. 1646. P. 462100. https://doi.org/10.1016/j.chroma.2021.462100
20. *Vrzal T., Malečková M., Olšovská J.* // Analytica Chimica Acta. 2021. Vol. 1147. P. 64. https://doi.org/10.1016/j.aca.2020.12.043
21. *Geer L.Y., Stein S.E., Mallard W.G. et al.* // J. of Chemical Information and Modeling. 2024. Vol. 64. No. 3. Pp. 690–696. https://doi.org/10.1021/acs.jcim.3c01758
22. *Raymond J.W., Gardiner E.J., Willett P.* // The Computer J. 2002. Vol. 45. No. 6. Pp. 631–644. https://doi.org/10.1093/comjnl/45.6.631

23. *Bender A., Glen R.C.* // Organic & Biomolecular Chemistry. 2004. Vol. 2. No. 22. P. 3204. https://doi.org/10.1039/B409813G
24. *Morehouse N.J., Clark T.N., McMann E.J. et al.* // Nature Communications. 2023. Vol. 14. No. 1. P. 308. https://doi.org/10.1038/s41467-022-35734-z
25. *Rogers D., Hahn M.* // J. of Chem. Inform. and Modeling. 2010. Vol. 50. No. 5. P. 742. https://doi.org/10.1021/ci100050t
26. *Hoo Z.H., Candlish J., Teare D.* // Emergency Medicine J. 2017. Vol. 34. No. 6. P. 357. https://doi.org/10.1136/emermed-2017-206735
27. *Polo T.C.F., Miot H.A.* // J. Vascular Brasileiro. 2020. Vol. 19. P. e20200186. https://doi.org/10.1590/1677-5449.200186
28. *Popov M.S., Ul'yanovskii N.V., Kosyakov D.S.* // Microchemical J. 2024. Vol. 197. P. 109833. https://doi.org/10.1016/j.microc.2023.109833