Сорбционные и хроматографические процессы. 2024. Т. 24, № 5. С. 711-722.
Sorbtsionnye i khromatograficheskie protsessy. 2024. Vol. 24, No 5. pp. 711-722.

ISSN 1680-0613

ОРИГИНАЛЬНЫЕ СТАТЬИ

Original article

A comparison of published in 2018-2024 general-purpose models for predicting gas chromatographic retention indices

Anastasia Yu. Sholokhova✉, Dmitriy D. Matyushin
A.N. Frumkin Institute of Physical Chemistry and Electrochemistry of Russian Academy of Sciences, Moscow, Russian Federation, shonastya@yandex.ru✉

Abstract. Retention indices are widely used in gas chromatography and chromatography-mass spectrometry as an additional factor in tentative identification (along with the mass spectrum). Reference data on retention indices are available only for a limited number of molecules; in other cases, retention indices predicted by mathematical models can be used. Models for predicting retention indices developed prior to 2018 mostly have either very low accuracy or a very narrow domain of applicability. However, in recent years, starting from 2018, the situation has begun to change: the use of deep neural networks and large training sets (mainly different versions of the NIST database) made it possible to build both accurate and general-purpose models for predicting gas chromatographic retention indices, with the accuracy increasing over time. In recent years, at least 7 deep learning-based models for predicting gas chromatographic retention indices have been released in the public domain. The authors always declare that their model is more accurate than previous models, however, in all cases, there are no independent measurements of accuracy. This work aimed to objectively and critically compare retention index prediction models and corresponding software using the same retention data set that was guaranteed not to intersect with the training sets used by the authors of the models. Seven models and corresponding software were considered, including MetExpert (2018), DeepReI (2021), SVEKLA (2021), and AIRI (2024). It was shown that for the non-polar stationary phase (ZB-5MS), the accuracy of the newest models gradually approaches the accuracy of the reference libraries and is quite high. The newer models are indeed more accurate than the older ones. At the same time, for the polar stationary phase (SH-Stabilwax), the accuracy (independent data set) is very low and significantly lower than that stated in the original papers devoted to the predictive models. For users with limited experience, the process of compiling and running software can be challenging, particularly when attempting to do so several years after publication. This is often due to incompatibility issues between model files and newer versions of the frameworks. It is not uncommon for software authors to discontinue any support of the software after an article has been published in a journal.

Keywords: gas chromatography, retention index, neural networks, machine learning
Acknowledgments: This research was supported by the Ministry of Science and Higher Education of the Russian Federation (№ 124041900012-4).
For citation: Sholokhova A.Yu., Matyushin D.D. A comparison of published in 2018-2024 general-purpose models for predicting gas chromatographic retention indices. *Sorbtsionnye i khromatograficheskie protsessy*. 2024. 24(5): 711-722. https://doi.org/10.17308/sorpchrom.2024.24/12510

Научная статья
УДК 544.03:543.544.3:004.8
doi: 10.17308/sorpchrom.2024.24/12510

Сравнение опубликованных в 2018-2024 гг. универсальных моделей для предсказания газохроматографических индексов удерживания

Анастасия Юрьевна Шолохова✉, Дмитрий Дмитриевич Матюшин
Институт физической химии и электрохимии им. А.Н. Фрумкина РАН, Москва, Россия, shonastya@yandex.ru✉

Аннотация. Индексы удерживания широко используются в газовой хроматографии и хромато-масс-спектрометрии в качестве дополнительного фактора при предварительной идентификации (наряду с масс-спектром). Справочные данные об индексах удерживания доступны лишь для ограниченного числа молекул, в остальных случаях можно использовать предсказанные с помощью математических моделей индексы удерживания. Модели для предсказания индексов удерживания, разработанные до 2018 года, в основном имеют или очень низкую точность, или очень узкую сферу применимости. Однако в последние годы, начиная с 2018 года, ситуация начала меняться: применение глубоких нейронных сетей и больших обучающих наборов (в основном разные версии базы данных NIST) позволило построить одновременно точные и универсальные модели для предсказания газохроматографических индексов удерживания, причем точность повышается с течением времени. За последние годы было опубликовано в открытом доступе как минимум 7 моделей, основанных на глубоком обучении, для предсказания газохроматографических индексов удерживания. Во всех случаях авторы декларируют, что точность их модели выше, чем точность предыдущих моделей, однако какие-либо независимые измерения точности во всех случаях отсутствуют. Целью данной работы было объективное критическое сравнение моделей для предсказания индексов удерживания и соответствующего программного обеспечения с использованием одного и того же набора данных об удерживании, заведомо не пересекающегося с обучающими наборами, использованными авторами моделей. Было рассмотрено 7 моделей и соответствующих компьютерных программ, в том числе модели MetExpert (2018), DeepReI (2021), SVEKLA (2021), AIRI (2024). Показано, что для неполярной неподвижной фазы (ZB-5MS) точность новейших моделей постепенно приближается к точности референсных библиотек и является чрезвычайно высокой. Более новые модели действительно являются более точными, чем более старые. В то же время для полярной неподвижной фазы (SH-Stabilwax) точность (независимый набор данных) очень низкая и значительно ниже, чем заявлено в оригинальных статьях, посвященных моделям для предсказания индексов удерживания. Отдельной проблемой для неопытного пользователя является компиляция и запуск программного обеспечения спустя несколько лет после публикации из-за несовместимости файлов моделей с новыми версиями фреймворков; авторы обычно не поддерживают никаким образом программное обеспечение после публикации статьи в журнале.

Ключевые слова: газовая хроматография, индекс удерживания, нейронные сети, машинное обучение.

Благодарности: данная работа выполнена при поддержке Министерства науки и высшего образования Российской Федерации в рамках госбюджетной темы № 124041900012-4.

Для цитирования: Sholokhova A.Yu., Matyushin D.D. A comparison of published in 2018-2024 general-purpose models for predicting gas chromatographic retention indices // *Сорбционные и хроматографические процессы*. 2024. Т. 24, № 5. С. 711-722. https://doi.org/10.17308/sorpchrom.2024.24/12510

Introduction

Retention indices (RI) based on *n*-alkanes, i.e., relative retention times, can be used in gas chromatography (GC) as an additional factor that increases the reliability of mass spectrometric (MS) identification [1-2]. Since a reference value for the RI is not available in databases for most of the available chemical compounds, the prediction of the RI based on the structure of the molecule is of great importance. Early studies [3] on RI prediction often considered very small data sets; all compounds, for which the model was built, belonging to one narrow class, and such models were difficult to use in practice. The first publicly published and truly versatile model appeared back in 2007 [4], but the prediction accuracy was very low and such RI were difficult to use for identification in practice [1].

Since 2018, there have been numerous publications devoted to the development of accurate, general-purpose (suitable for a wide variety of chemical compounds) models for predicting RI based on the structure of a molecule [2, 5-14], as well as the practical application of such models in the analysis of complex mixtures [15-16], in particular for the analysis of environmental objects [15] and in metabolomics [6]. In the majority of cases, such works use neural networks [2, 5-11] to predict RI. In 2019, our team was the first in the world to use deep learning to predict RI [2]. Since then, deep learning has become the main method for accurate and versatile prediction of gas chromatographic RI.

A variety of neural networks are used for RI prediction using deep learning: deep one-dimensional convolutional neural networks (1D CNN) using a string representation of the molecule structure (so-called SMILES strings [17]) as input [2, 9, 10], deep two-dimensional convolutional neural networks

(2D CNN) of various types [9, 11], multilayer perceptrons (MLP) using molecular descriptors (MD) or molecular fingerprints (MF) as features [6, 9, 10, 13], graph neural networks (GNN) processing the molecular graph directly [5, 7, 8]. MF and MD are numerical features characterizing the structure of the molecule. An overview of MD is given in many previous works [9, 18]. In addition to neural networks, other techniques such as gradient boosting (GB) [9, 12] and support vector regression [10, 13, 14] can also be used.

Unfortunately, the authors of many such works do not publish ready-to-use software and trained parameters of the models in the public domain [7, 12-14]. It is impossible to apply such models in practice otherwise than by reproducing the entire procedure for constructing the model as it was done by the authors. In 2018-2024, 7 works were published [2, 5, 6, 8-11] devoted to accurate and universal prediction of RI using models trained on large and diverse data sets, in which the resulting models and software are published in the public domain [2, 5, 6, 9-11] or available online [8]. The majority of these articles focus on standard and semi-standard non-polar phases (polydimethylsiloxane, 5%-phenyl-polydimethylsiloxane), only two of them [8, 10] also contain models for predicting RI for standard polar stationary phases (polyethylene glycol).

In most cases, the authors of studies devoted to the development of new models for predicting RI using machine learning provide a comparison of their model with previous ones in their publications. However, the comparison is performed using different data sets, and it is often difficult to be sure of the correctness of such a comparison. The published software in most cases [2, 6, 9, 11] is not updated and not supported after the publication of the corresponding article, and a compilation years after the initial publication may be difficult due to outdated versions of the frameworks and libraries used. There are often no works independent of the authors of the original model that use and critically evaluate the accuracy. In other cases [5, 8], on the contrary, the current version of the corresponding software may differ from that described in the journal publication.

The aim of the present study was to evaluate the accuracy and usability of a current range of general-purpose models (and corresponding software) for predicting gas chromatographic RI using the same independent data set. For this purpose, we used a recently published data set [19] of the RI of various organic compounds for ZB-5MS and SH-Stabilwax stationary phases.

Methods

Data set and accuracy evaluation. The data set for the ZB-5MS stationary phase was taken from the corresponding repository [19]. The data set was divided into two subsets. The first subset contained molecules for which RI data for standard or semi-standard stationary phases were available in the NIST 20 database. The second subset consisted of molecules for which RI data for standard or semi-standard stationary phases were absent in the NIST 20 database. The first subset was used to assess the accuracy of the RI values reported in the NIST 20 database. The second subset was used to evaluate the prediction accuracy of machine learning models. Since 5 of the 7 machine learning models considered were trained using the NIST database of different legacy versions (from NIST 08 to NIST 20), it was thus ensured for these models that the molecules used to assess the accuracy of the models were not part of the training data sets used to train these models.

The SH-Stabilwax stationary phase data set was divided similarly. In this case, the criterion for assigning a molecule to one of the subsets was the presence of data for this molecule in the NIST 20 database for standard polar stationary phases. The SMILES strings, which encode the structure of the molecule, were used without alteration as they were provided in the repository [19].

Table 1. Publicly available accurate and general-purpose retention index prediction models
Таблица 1. Общедоступные точные и универсальные модели для предсказания индексов удерживания

Designation	Year	NIST version	Model description	Reference
MetExpert	2018	-	Two-layer perceptron, uses MD as input features	[6]
JCA19	2019	NIST 08	Deep 1D CNN using SMILES strings as input	[2]
Access	2020	NIST 17	Four models that form the ensemble: 1D and 2D CNN, deep MLP, GB; SMILES strings, 2D molecule sketches, MD, and MF are used as inputs	[9]
DeepReI	2021	NIST 14	Deep 2D CNN using SMILES strings as input	[11]
SVEKLA	2021	NIST 17	Two models that form the ensemble: 1D CNN and deep MLP; SMILES strings, MD, and MF are used as inputs	[10]
GCMS-ID	2023	NIST 20	Deep attention-based GNN	[8]
AIRI	2024	NIST 23	Eight attention-based GNN (graph transformers) that form the ensemble	[5]

The parameters of the chromatographic modes, the description of the experiment, and the structural formulas of the molecules are given in the repository [19].

When determining whether a molecule is present in the NIST 20 database, stereoisomers were considered to be the same compound. The accuracy measures were root mean square error (RMSE), mean absolute error (MAE), median absolute error (MdAE), mean percentage error (MPE), median percentage error (MdPE), and coefficient of determination (R^2).

Models and software considered. The considered machine learning models and the designations are presented in table 1. For the MetExpert model [6] (version v1), the archive was downloaded from the corresponding repository [20]. The ANN folder contains the neural network weights and all other data necessary for reproducing the model. The equations, by which the calculation should be performed, are contained in the MetExpert_Pipeline.xlsb file (in the source code of the script). We calculated the MD using the command contained in the MetExpert_Pipeline.xlsb file, and we implemented further calculations ourselves using the neural network parameters given in the ANN folder.

The source code for the JCA19 model was taken from the Supplementary Material of the corresponding article [2]; for the Access model [9], the source code was taken from the repository [21]. The source code was compiled and executed in accordance with the instructions provided with the source code. The Java Development Kit (version 11.0.23) and Maven (version 3.6.3) were used. The SVEKLA [10] software (version 0.0.2-alpha1) was downloaded from the repository [22] (ready-to-use binaries). The graphical user interface (GUI) was not used, but command line options were used to evaluate the accuracy. The corresponding command line options are described in the information.pdf file in the repository [23]. For the Access and SVEKLA models [10], a value of 16 was used as the value of the stationary phase type for both polar and non-polar stationary phases. Detailed information on the stationary phase codes can be found in the Supplementary Materials to the corresponding articles [9, 10].

The GCMS-ID [8] model is available on the website [24], but the website address has changed over the last year and there is no

Table 2. Comparison of retention index prediction models and corresponding software
Таблица 2. Сравнение моделей для предсказания индексов удерживания и соответствующее программное обеспечение

	MetExpert	JCA19	Access	DeepReI	SVEKLA	GCMS-ID	AIRI
Graphical user interface	Yes	No	No	Yes	Yes	Yes	No
Source code and model are publicly available for download	Yes	Yes	Yes	Yes	Yes	No	Yes
Batch prediction	Possible*	Yes	Yes	Yes	Yes	No	Yes
Difficulty of installation and use	Unclear*	Medium	Medium	Hard	Easy	Easy	Medium
Ready-to-use binaries or website available	Unclear*	No	No	No	Yes	Yes	No
Polar stationary phases support	No	No	No	No	Yes	Yes	No
Non-standard stationary phases support	No	No	No	No	Yes	No	No
Persistent versions	Yes	Yes	Yes	No	Yes	No	In part
Accuracy**	Low	Low	Medium	Low	Medium	High	Highest

* – The MetExpert package contains a GUI (based on Microsoft Excel), ready-to-use binaries are publicly available. However, we were unable to run them and achieve retention index prediction. The predictive model can be very easily reimplemented independently, the parameters are published in a convenient form.
** – Quantitative comparison is given below

guarantee that it will continue to be available. The stationary phase type was selected as either "semi-standard non-polar" or "standard polar". The AIRI model [5] is implemented in the masskit_ai package (version 1.2.2, installed together with masskit, version 1.2.2) [25]. The SMILES strings were converted to .sdf format using the Open Babel utility (version 3.1.1), and then the instructions from the NIST website [26] were applied.

The DeepReI model [11] was installed according to instructions from the corresponding repository (version not specified). The following software versions were used: R 4.1.2, TensorFlow 2.0.0, Keras 2.3.1, and Python 3.7.16. A conda virtual environment was created with the appropriate versions of Python and TensorFlow. All web resources and repositories were accessed in July-August 2024. Unfortunately, in the future, the websites and repositories may be removed, and the instructions given may no longer work with newer versions of operating systems and software. A more detailed discussion of the persistence of predictive models is provided below. All calculations were performed using the Linux Mint operating system (version 21).

Results and discussion

Qualitative comparison of predictive models and related software. Table 2 presents a qualitative comparison of the predictive models and the corresponding software. Each model is accompanied by a computer program (script). Some of the models (MetExpert [6], SVEKLA [10], DeepReI [11], GCMS-ID [8]) are equipped with a

GUI, while others are run from the command line. However, only SVEKLA [10] and GCMS-ID [8] have a built-in molecule editor; for other models, the user is required to convert the structures to SMILES strings [17] prior to use. For all models except GCMS-ID, the weights (trainable parameters) of the neural networks and the source codes are available online. Thus, these software and models are available for full study and use in any way.

Not all software is equally easy to run and use. For example, DeepReI instructions [11] contain typos, and a user has to manually install many dependencies (not all of which are mentioned in the instructions) to run and use it. But the biggest difficulties for an inexperienced user are related to the fact that model files are not compatible with modern versions of Keras/TensorFlow, and the required versions of frameworks are not compatible with modern versions of Python, while DeepReI [11] itself is written in R, and Python dependencies are hidden behind R wrappers. In addition, when something goes wrong (e.g., the framework cannot load a model due to a version mismatch), the DeepReI GUI does not show any error messages, and RI prediction just does not work.

We were not able to achieve the prediction of RI directly using the MetExpert package [6, 20] as published. However, the model can be easily implemented independently by a user with minimal programming skills. Of all the software, only SVEKLA [10, 16, 22] has compiled and workable binaries that can be downloaded to a computer and directly run without compiling.

The GCMS-ID model [8], while easy to use and convenient, has important drawbacks when used in research. The model is not available for download, the prediction is server-side, and the user has no control over what happens and how well the model used matches what is described in the original publication [8]. There is no assurance that the model will work after a certain amount of time. Batch processing is not possible, only prediction of one molecule at a time is supported.

The persistence of models is an important issue. If a version of the software and model is available in an immutable repository (such as Figshare [21, 23]), the results will be reproducible even after a significant amount of time. Content from websites such as Github [22, 25] or a website owned by model creators [24, 26] can be removed at any time. Calculations made with such a model may not be reproducible at any point in time. In a situation where authors do not make releases with unambiguous version numbers, it may not be clear which version the calculation was made with. A significant challenge when attempting to reproduce results from articles published a considerable time ago is the obsolescence of dependencies and the necessity to utilize older versions. Nevertheless, at the time of writing this paper, we have successfully run all 7 models.

Quantitative comparison of accuracy of predictive models. In this section, we quantitatively compare the RI prediction accuracy of the 7 models listed in table 1 for the ZB-5MS stationary phase using the published data set [19]. This stationary phase is a semi-standard stationary phase (5%-phenyl-polymethylsiloxane). For 6 molecules (3-(2-methoxyethyl)octan-1-ol, 2-hydroxytyrosine, 6-methyl-2-pyridone, 3,6,9,12-tetraoxotridecanol, 3,6-dimethylphthalonitrile, indole-3-carbinole), all models give an error of more than 100 RI units. At the same time, the predictions of the models are close to each other. It is likely that the data set used contains errors, for example, due to mislabeling of samples. A simultaneous discrepancy between the predictions of a number of models and the experimental value may indicate an error in the data set [27].

However, we have no certainty that it is exactly an error in the data. An interesting example of how many models can go wrong simultaneously is 4-hydroxy-2-methoxybenzaldehyde. For this molecule, all but the two most recent and most accurate models (AIRI [5] and GCMS-ID [8]) give predictions that

Table 3. Accuracy of published general-purpose models for predicting retention indices based on the structure of a molecule (semi-standard non-polar stationary phase)
Таблица 3. Точность опубликованных универсальных моделей для предсказания индексов удерживания на основе структуры молекулы

Designation	RMSE	MAE	MdAE	MPE, %	MdPE, %	R^2
MetExpert	242.5	178.6	131.6	14.24	10.14	0.425
JCA19	101.5	76.8	57.2	5.35	4.42	0.941
Access (1D CNN)	64.3	50.0	41.2	3.70	2.96	0.968
Access (2D CNN)	58.7	44.7	32.3	3.38	2.48	0.965
Access (MLP)	55.4	36.3	21.1	2.59	1.64	0.970
Access (GB)	90.2	63.0	46.4	4.46	3.94	0.922
Access (Ensemble)	52.2	37.2	28.5	2.71	2.23	0.975
DeepReI	147.2	73.7	40.7	5.14	3.46	0.782
SVEKLA (1D CNN)	70.0	51.4	36.4	3.70	3.15	0.964
SVEKLA (MLP)	50.6	33.9	22.7	2.41	1.79	0.975
SVEKLA (Ensemble)	54.8	38.3	25.1	2.73	2.12	0.976
GCMS-ID	37.0	25.1	17.6	1.97	1.23	0.987
AIRI	30.9	17.0	10.4	1.35	0.72	0.991
NIST 20*	56.4	22.9	7.3	1.97	0.66	0.966
NIST 20 (distant** outliers removed)*	25.0	14.0	6.9	1.35	0.61	0.993

* - A different subset of experimental retention indices was used.
** - Discrepancy greater than 150 units.

are 100-200 units lower than the observed experimental value. The prediction of the two most accurate models coincides with the experimentally observed value. We believe that both an error in the data set and simultaneously equally incorrect predictions of a number of models at once are possible. This can be caused by an error in the training set, e.g., an incorrect RI value for a molecule close to the one for which the prediction is performed.

When calculating the accuracy measures, 6 molecules for which all models give an error of more than 100 RI units were excluded from the calculation. The contribution of these molecules to accuracy measures such as RMSE and MAE is too large and makes the comparison less clear. After excluding these 6 molecules from the data set, the accuracy measures were calculated. The results are summarized in table 3. In addition, accuracy measures are provided to compare our observed RI with the NIST database (using a different subset of the data).

The accuracy of the AIRI model [5] is impressively high. However, this model was trained using NIST 23, and some of the molecules from the set used to assess the accuracy may have been present in the training set. This makes such a comparison not entirely correct. The GCMS-ID model [8] is also highly accurate. The SVEKLA [10] model developed at the A.N. Frumkin Institute of Physical Chemistry and Electrochemistry of the Russian Academy of Sciences (IPCE RAS) ranks third in accuracy (an ensemble of MLP and 1D CNN). Fig. 1 shows how the accuracy of RI prediction increased in 2018-2024. In just 6 years, spectacular advances have been made in this field through the application of deep learning. Our team at

IPCE RAS was the first to apply deep learning to this task [2].

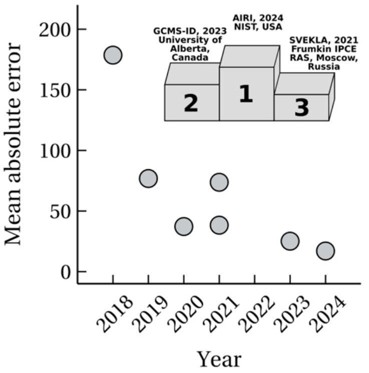

Fig. 1. Accuracy of general-purpose models for predicting retention indices published in 2018-2024, with an indication of the three most accurate models

Рис. 1. Точность универсальных моделей для предсказания индексов удерживания, опубликованных в 2018-2024 г. с указанием трех наиболее точных

The SVEKLA and Access models are ensembles of several models [9, 10]. Table 3 also provides a comparison of the different models included in the ensemble. Interestingly, for this data set, the 1D CNN gives much worse accuracy than the MLP, while for other data (essential oils, metabolites, and NIST subsets) the accuracy of these models is comparable [9, 10]. This shows that the accuracy and the ratio of the accuracies of different models strongly depend on the data used: there is no universally the most accurate predictive model.

Fig. 2 shows the cumulative distribution of prediction errors for different molecules. It is evident that the threshold value of 70 used in the previous work [16] for rejecting false candidates in tentative GC-MS identification is too low. Even for relatively accurate models, more than 10% of candidates will be erroneously rejected. In general, the cumulative distribution can be used to select a threshold value of the difference between predicted and observed RI for rejecting false candidates in tentative GC-MS identification.

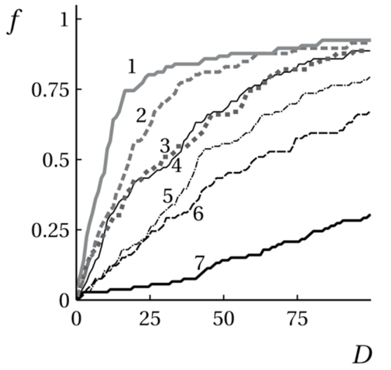

Fig. 2. Dependence of the fraction *f* of molecules, for which the absolute error is not higher than *D*, on the value of *D* for different predictive models: 1 – AIRI; 2 – GCMS-ID; 3 – SVEKLA; 4 – Access; 5 – DeepReI; 6 – JCA19; 7 – MetExpert

Рис. 2. Зависимость доли молекул *f*, для которых абсолютная ошибка не превышает *D*, от величины *D* для различных моделей

The authors of each subsequent work devoted to the prediction of RI declare that the achieved accuracy is higher than in previous works. Table 3 and fig. 1-2 show that this is generally true upon independent verification. The results are generally reproducible. Fig. 3 shows the correlations of predicted and experimental RI values for different predictive models. It also shows the correlation between the RI values from the repository [19] and the RI values from the NIST 20 database (another subset of molecules). The MetExpert model demonstrates relatively low accuracy when applied to these data sets. This is due to the fact that it was trained not on the NIST database, but on a small data set containing metabolites and essential oils [6]. For organofluorine compounds, it does not give satisfactory predictions; in fig. 3, the group of outliers (mainly organofluorine compounds) is shown by an ellipse. The accuracy of the model is very dependent on the presence of compounds close in structure to the predicted compounds in the training data set [28]. Fig. 2-3 show all compounds [19], including 6 molecules for which all models give an error greater than 100 units.

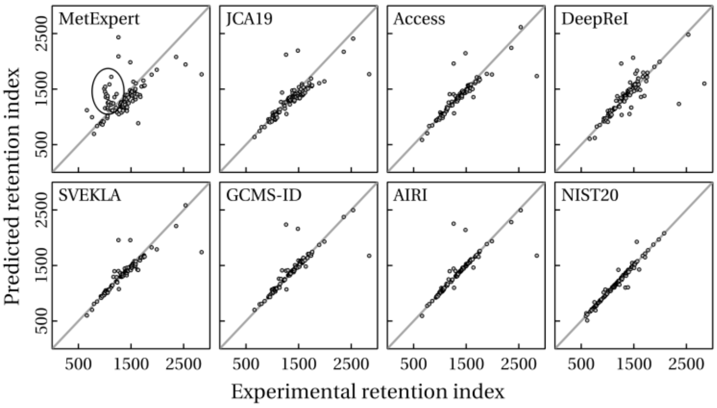

Fig. 3. Correlation between observed and predicted retention indices (semi-standard non-polar stationary phase) for different predictive models; a group of molecules, mainly polyfluoro-substituted compounds, for which MetExpert gives highly erroneous predictions is highlighted; in the case of NIST 20 (last subplot), library values are considered instead of predicted values; data for a different set of molecules are considered

Рис. 3. Корреляция между экспериментальными и предсказанными индексами удерживания для различных моделей; выделена группа молекул, в основном полифторзамещенных соединений, для которых MetExpert дает большие значения ошибки; в случае NIST 20 (последний график) вместо предсказанных значений рассматриваются библиотечные значения (данные для другого набора молекул)

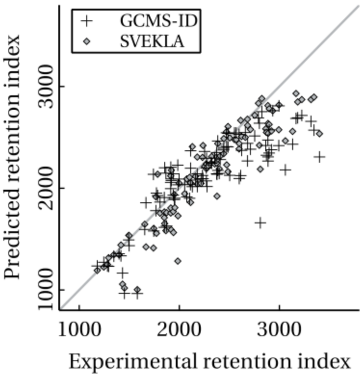

Fig. 4. Correlation between observed and predicted retention indices (standard polar stationary phase) for different predictive models

Рис. 4. Корреляция между наблюдаемыми и предсказанными индексами удерживания (стандартная полярная неподвижная фаза) для различных моделей

The accuracy of the AIRI model [5] (discrepancy values between observed and predicted RI) is comparable to the accuracy of the NIST 20 RI reference database itself, as shown in table 3. When comparing the observed RI from the repository [19] with the NIST 20 database, a perfect match is also not observed. The NIST database is known to contain a number of erroneous entries [27].

Prediction accuracy for polar stationary phases. The RI prediction accuracy for the polar stationary phase was estimated in a similar manner. Of the 7 models considered, only SVEKLA [10] and GCMS-ID [8] have the ability to predict RI for the polar stationary phase (polyethylene glycol). Fig. 4 shows the correlation between the RI predicted by the two models and the observed ones. The prediction accuracy is very low, and the discrepancy is hundreds of units for many molecules. Unfortunately, for the molecules considered (the structural formulas of all molecules are given in the repository [19]), none of the available models allow for achieving satisfactory accuracy in predicting RI for the polar stationary phase.

Table 4. Accuracy of published general-purpose models for predicting retention indices based on the structure of a molecule (standard polar stationary phase)
Таблица 4. Точность опубликованных универсальных моделей для предсказания индексов удерживания на основе структуры молекулы (стандартная полярная неподвижная фаза)

Designation	RMSE	MAE	MdAE	MPE, %	MdPE, %	R^2
SVEKLA (1D CNN)	284.3	211.9	166.9	9.33	7.40	0.791
SVEKLA (MLP)	240.7	172.0	117.7	7.34	5.75	0.864
SVEKLA (Ensemble)	250.6	185.4	133.7	8.05	5.89	0.847
GCMS-ID	329.5	235.5	161.6	9.68	7.53	0.708
NIST 20*	102.7	35.9	14.3	1.84	0.75	0.960
NIST 20 (distant** outliers removed)*	34.9	22.3	12.0	1.22	0.70	0.995

* - A different subset of experimental retention indices was used; ** - Discrepancy greater than 150 units.

At the same time, for those molecules for which the reference RI value is contained in the NIST database, there is a satisfactory agreement between the values from the repository [19] and the values from the NIST database. It is also evident (fig. 4) that the predictions of the two models differ greatly from one another. Thus, namely the low accuracy of RI prediction by published models for polar stationary phases is observed. The corresponding values of the accuracy measures are given in table 4. Such low prediction accuracy compared to that stated in the publications devoted to the corresponding models is because [28] the molecules, for which we performed testing, differ significantly in structure from most molecules for which the NIST database contains RI data for polar stationary phases.

Conclusions

In many areas of science, there is currently [29-30] a so-called “reproducibility crisis”: when trying to repeat scientific results from publications, researchers are faced with the fact that the results are not reproducible. In each case, it is difficult to establish the reason why this happened: it could be a mistake by the one trying to repeat, it could be a mistake in the original work, or it could be the result of dishonest actions by the author of the original work. At the same time, this study shows that the accuracy of models for predicting gas chromatographic retention indices really behaves exactly as the authors of the relevant papers claim: each subsequent model is indeed more accurate than the previous ones. While the 2018-2021 models had much lower accuracy compared to library retention indices (the average absolute error is several times higher), the accuracy of the latest models approaches the accuracy of experimental reference retention indices. Most likely, in the coming years, it will be possible to use the predicted retention indices as reference ones in most cases, and the growth of the size of retention index libraries will be of interest only from the point of view of the growth of training sets. At the same time, these optimistic remarks apply only to non-polar stationary phases. The accuracy of the prediction of retention indices for various chemical compounds for the polar stationary phase is very low, significantly lower than that claimed by the authors of the predictive models. We believe that the main reason for this discrepancy is that the training set is not representative and not diverse enough. However, it is to be hoped that in the near future, accurate and free software for predicting retention indices will be available for all stationary phases. “Raw” predictions of retention indices using all models considered have been added to the repository with experimental data [19].

Conflict of Interest

The authors declare that they have no known financial conflicts of interest or personal relationships that could have influenced the work reported in this article.

References

1 Zhang J., Koo I., Wang B., Gao Q. W., Zheng C. H., Zhang X., A large scale test dataset to determine optimal retention index threshold based on three mass spectral similarity measures, *Journal of Chromatography A.* 2012; 1251: 188-193. doi.org/10.1016/j.chroma.2012.06.036

2 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

3 Héberger K., Quantitative structure–(chromatographic) retention relationships, *Journal of Chromatography A.* 2007; 1158(1-2): 273-305. https://doi.org/10.1016/j.chroma.2007.03.108

4 Stein S. E., Babushok V. I., Brown R. L., Linstrom P. J., Estimation of Kováts Retention Indices Using Group Contributions, *Journal of Chemical Information and Modeling.* 2007; 47 (3): 975-980. https://doi.org/10.1021/ci600548y

5 Geer L. Y., Stein S. E., Mallard W. G., Slotta D. J., AIRI: Predicting Retention Indices and Their Uncertainties Using Artificial Intelligence, *Journal of Chemical Information and Modeling.* 2024; 64(3): 690-696. https://doi.org/10.1021/acs.jcim.3c01758

6 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

7 Qu C., Schneider B. I., Kearsley A. J., Keyrouz W., Allison T. C., Predicting Kováts Retention Indices Using Graph Neural Networks, *Journal of Chromatography A.* 2021; 1646: 462100. https://doi.org/10.1016/j.chroma.2021.462100

8 Anjum A., Liigand J., Milford R., Gautam V., Wishart D. S., Accurate prediction of isothermal gas chromatographic Kováts retention indices, *Journal of Chromatography A.* 2023; 1705: 464176. https://doi.org/10.1016/j.chroma.2023.464176

9 Matyushin D.D., Buryak A.K., Gas Chromatographic Retention Index Prediction Using Multimodal Machine Learning, *IEEE Access.* 2020; 8: 223140-223155. https://doi.org/10.1109/ACCESS.2020.3045047

10 Matyushin D.D., Sholokhova A.Yu., Buryak A.K., Deep Learning Based Prediction of Gas Chromatographic Retention Indices for a Wide Variety of Polar and Mid-Polar Liquid Stationary Phases, *International Journal of Molecular Sciences.* 2021; 22 (17): 9194. https://doi.org/10.3390/ijms22179194

11 Vrzal T., Malečková M., Olšovská J., DeepReI: Deep learning-based gas chromatographic retention index predictor, *Analytica Chimica Acta.* 2021; 1147: 64–71. doi.org/10.1016/j.aca.2020.12.043

12 Matyushin D.D., Sholokhova A.Yu., Buryak A.K., Gradient boosting for the prediction of gas chromatographic retention indices, *Sorbtsionnye I khromatograficheskie protsessy.* 2019; 19(6): 630-635. https://doi.org/10.17308/sorpchrom.2019.19/2223

13 de Cripan S. M., Cereto-Massagué A., Herrero P., Barcaru A., Canela N., Domingo-Almenara X., Machine Learning-Based Retention Time Prediction of Trimethylsilyl Derivatives of Metabolites, *Biomedicines.* 2022; 10(4): 879. https://doi.org/10.3390/biomedicines10040879

14 Matyushin D.D., Buryak A.K., Application of regression learning for gas chromatographic analysis and prediction of toxicity of organic molecules, *Russian Chemical Bulletin.* 2023; 72(2): 482-492. https://doi.org/10.1007/s11172-023-3811-2

15 Su Q. Z., Vera P., Nerín C., Lin Q. B., Zhong H. N. Safety concerns of recycling postconsumer polyolefins for food contact uses: Regarding (semi-)volatile migrants untargetedly screened, *Resources, Conservation and Recycling.* 2021; 167: 105365. https://doi.org/10.1016/j.resconrec.2020.105365

16 Sholokhova A. Yu., 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

17 Weininger D., SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, *Journal of Chemical Information and Computer Sciences.* 1988; 28(1): 31-36. https://doi.org/10.1021/ci00057a005

18 Zhokhov A.K., Loskutov A.Yu., Rybal'chenko I.V. Methodological Approaches to the Calculation and Prediction of Retention Indices in Capillary Gas Chromatography, *Journal of analytical chemistry.* 2018; 73(3): 207-220. https://doi.org/10.1134/S1061934818030127

19 Matyushin D.; Sholokhova A.Yu. (2024). A data set of retention indices and retention times for 200+ molecules and two stationary phases (gas chromatography). figshare. Dataset. https://doi.org/10.6084/m9.figshare.26119558.v2

20 https://sourceforge.net/projects/metexpert/ (accessed: 24.08.2024)

21 Matyushin D. (2020). Supplementary data and code for the article "Gas chromatographic retention index prediction using multimodal machine learning". figshare. Software. https://doi.org/10.6084/m9.figshare.12651680.v2 (accessed: 24.08.2024)

22 https://github.com/mtshn/svekla (accessed: 24.08.2024)

23 Matyushin D. (2021). Supplementary materials for the article "Deep learning based prediction of gas chromatographic retention indices for a wide variety of polar and mid-polar liquid stationary phases": source code of software and parameters of pre-trained models. figshare. Software. https://doi.org/10.6084/m9.figshare.14602317.v1 (accessed: 24.08.2024)

24 https://gcms-id.ca (accessed: 24.08.2024)

25 https://github.com/usnistgov/masskit_ai/ (accessed: 24.08.2024)

26 https://pages.nist.gov/masskit_ai/ (accessed: 24.08.2024)

27 Khrisanfov M.D., Matyushin D.D., Samokhin A.S. A general procedure for finding potentially erroneous entries in the database of retention indices. *Analytica Chimica Acta.* 2024; 1297: 342375. https://doi.org/10.1016/j.aca.2024.342375

28 https://github.com/mtshn/molsimwax (accessed: 24.08.2024)

29 Baker M. 1,500 scientists lift the lid on reproducibility, *Nature.* 2016; 533(7604): 452-454. https://doi.org/10.1038/533452a

30 Fanelli D., Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences.* 2018; 115(11): 2628-2631. https://doi.org/10.1073/pnas.1708272114

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 – в.н.с. лаборатории «умных» методов химического анализа, Институт физической химии и электрохимии имени А.Н. Фрумкина РАН, Москва, Россия

D.D. Matyushin – researcher, laboratory of physico-chemical principles of chromatography and chromatography – mass spectrometry; A.N. Frumkin Institute of Physical Chemistry and Electrochemistry, RAS, Moscow, Russian Federation, email: dm.matiushin@mail.ru; ORCID: 0000-0003-0978-7666

A.Yu. Sholokhova – leading researcher, laboratory of "smart" methods of chemical analysis; A.N. Frumkin Institute of Physical Chemistry and Electrochemistry, RAS, Moscow, Russian Federation, email: shonastya@yandex.ru; ORCID: 0000-0003-4192-1677

Статья поступила в редакцию 27.08.2024; одобрена после рецензирования 20.09.2024; принята к публикации 02.10.2024.
The article was submitted 27.08.2024; approved after reviewing 20.09.2024; accepted for publication 02.10.2024.