IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier

Gas Chromatographic Retention Index Prediction Using Multimodal Machine Learning

Dmitriy D. Matyushin[1], Aleksey K. Buryak[1]

[1]A.N. Frumkin Institute of Physical Chemistry and Electrochemistry, Russian Academy of Sciences, 31 Leninsky Prospect, Moscow, GSP-1, 119071, Russia

Corresponding author: Dmitriy D. Matyushin (e-mail: dm.matiushin@mail.ru).

Funding: The work was supported by the Ministry of Science and Higher Education of the Russian Federation (grant agreement №075-15-2020-782).

ABSTRACT Gas chromatography is a widely used method in analytical chemistry and metabolomics. Using gas chromatography, vaporizable compounds can be separated for their further identification. Retention indices are standardized values that depend only on a chemical structure of a compound and on a stationary phase and characterize the retention of a compound in a chromatographic system. Retention index prediction is an important task because databases contain experimental values for a small fraction of all possible molecules, while this information is usable for untargeted analysis. In this work, we consider four machine learning models for retention index prediction: 1D and 2D convolutional neural networks, deep residual multilayer perceptron, and gradient boosting. String representation of the molecule, 2D representation of the chemical structure, molecular descriptors and fingerprints, and molecular descriptors are used as inputs of these four models, respectively, along with information about the stationary phase. The first and third models show the best performance, while the other two perform slightly worse. The models predict retention index values for various standard and semi-standard non-polar stationary phases. Further improvement in performance was achieved using a linear model that uses the results of four previous models as inputs (model stacking). The models were tested using various diverse data sets: flavor compounds, essential oils, metabolomics-related compounds. Achieved accuracy: median absolute and percentage errors – 6-40 units and 0.8-2.2%. Accuracy depends on a test data set. The stacking model outperforms previously reported approaches for all test data sets. Parameters of a pre-trained model and some source code are provided.

INDEX TERMS Analytical chemistry, convolutional neural network, deep learning, gas chromatography, gradient boosting, residual neural network, retention index, untargeted chemical analysis.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3045047, IEEE Access

I. INTRODUCTION

Gas chromatography (GC) is an important method for separating compounds and chemical analysis and is widely used in metabolomics, environmental analysis and other fields. Using gas chromatography, mixtures of vaporizable compounds can be efficiently and rapidly separated for their further detection and identification using electron ionization mass spectrometry (MS) or other methods. A mixture of vapors of the compounds to be separated moves with a stream of gas (mobile phase) along the surface of a non-volatile liquid (stationary phase). Separation is achieved due to different volatility and affinity of different compounds to the stationary phase. This leads to the fact that different compounds are retained in the chromatographic system for a different periods of time. The retention time depends on all parameters of chromatographic separation (such as temperature and mobile phase flow) and is not transferable between different systems and conditions. Retention indices (RI) are dimensionless standardized values that depend only on a chemical structure of a compound and on a stationary phase and characterize the relative retention of compounds. There are many systems of retention indices: Kovats RI [1-3], Lee RI [4], RI based on fatty acid methyl esters (FAME) [5]. The most commonly used system is Kovats RI system, which is based on the relative retention time of a compound compared with the retention times of n-alkanes.

The combination of gas chromatography and mass spectrometry (GC-MS) is a common method of analysis of complex mixtures. Using mass spectrometry, it is possible to determine the molecular weight of the unknown and make a reasonable assumption about its structure basing on fragment ions. It is usually made using a search in mass spectral databases [6-8]. Since only a fraction of organic molecules is contained in such databases, methods that do not depend on experimental reference spectra are under development too [9-11]. In both cases (with or without a database of experimental mass spectra), the accuracy of identification can be improved using RI as an additional constraint [8-9, 12]. Many millions of organic molecules are described; experimental mass spectra are available for several hundred thousand of them. The number of compounds for which experimental RI is available does not exceed 200000. The largest RI database (will be released soon) – the NIST 20 database will contain RI for ~140 thousand compounds. The previous release – NIST 17 contains RI for 99400 compounds. Other databases [5-6, 13-15] are orders of magnitude smaller and significantly overlap [16] with NIST.

Accurate RI prediction and the use of predicted RI as a reference can significantly expand the application of RI for GC-MS identification. Current RI prediction methods that are intended to be near-universal (applicable to diverse organic compounds rather than to one narrow class of molecules) are much less accurate than experimental RI from databases [8, 16-17]. Mean absolute errors (MAE) and median absolute errors (MdAE) for the most accurate and versatile RI prediction methods are in the range 30-100 and 17-50 RI units, respectively [16-19]. For experimental RI from the NIST database, an error was previously reported in the range 11-13 RI units [8, 17]. This value strongly depends on the way how it was calculated: experimental values in a database are given for very different chromatographic systems, and such values are compared together for all “standard” and “semi-standard” non-polar stationary phases without distinction [8, 17]. For the majority of compounds, there is only one experimental value in a database, and it is difficult to really estimate how reliable it is. RI deviation between experiments with different column instances of exactly the same column type and the same experimental conditions is 1-4 RI units [1]. The deviation is up to 20 RI units for exactly the same column type but various experimental conditions (temperature, sample concentration) [1].

Predicted RI are inaccurate in comparison with experimental ones but can be used as reference for GC-MS library search [8, 12, 16, 20-22]. The use of RI prediction makes GC-MS identification more reliable both when a spectral database is used or not [9]. The dependence of the confidence of identification on the reference RI accuracy was recently discussed [8]. The development of more accurate RI prediction methods will improve GC-MS identification. RI is usually predicted with machine learning methods. Most publications devoted to this subject usually use quite small training and test sets (<200 compounds), which are not really diverse and cover only one narrow class of chemical compounds. Such works use molecular descriptors that are generated with various, often proprietary, software. It is not really possible to cover diverse categories of chemical compounds, such as metabolites, with several such models. Many such RI prediction models were extensively reviewed [16, 23-24]. The most notable works [9, 16-22, 25-30] about RI prediction, which claim to be more universal and use large and diverse data sets, are summarized in Table 1.

The RI prediction task is the prediction of one number (RI) basing on the structure of a molecule. There are many ways how a molecule can be represented. The most common input features for machine learning driven prediction of molecule properties are various molecular descriptors (MD) – relevant features that can be calculated basing on the structure and are interconnected with properties of compounds. There are many proprietary and free software packages for calculating MD [31-33]. Types of MD and their usage, in particular for RI prediction, were extensively reviewed in previous works [24, 33-36]. A typical diverse set of MD contains features that are very diverse in nature: integer and real numbers, categorical features with different meanings. MD-based RI prediction can be made using all variety of machine-learning

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3045047, IEEE Access

regression methods that work with tabular input. Linear regression [17, 19-22, 25, 28-30], neural networks [9, 26-27], *k*-nearest neighbors [20], support vectors regression [20-21], and gradient boosting [18] were used.

TABLE 1
RETENTION INDEX PREDICTION USING MACHINE LEARNING FOR LARGE AND DIVERSE DATA SETS

Compounds	N	Year	Ref.
Diverse set of toxicologically relevant compounds	846	2004	[25]
Terpenes	573	2007	[26]
NIST 05	25296	2007	[17]*
NIST 05	24509	2009	[21]
Diverse set of toxicologically relevant compounds	846	2009	[27]
Flavor-related compounds	656	2012	[28]*
Flavors and fragrances	1208	2015	[29]
Flavors and fragrances	1184	2015	[30]**
Diverse set of volatile compounds	560	2016	[20]**
Metabolomics-related compounds. Mostly trimethylsilyl- derivatives	337	2017	[22]
Metabolites, essential oils	2196	2018	[9]
Components of essential oils	791	2018	[19]
NIST 08, PubChem were used for training; metabolites, essential oils, flavors were used for testing	***	2019	[16]
NIST 17 was used for training; essential oils, flavors were used for testing	72976 (training)	2019	[18]

N – data set size; * – RI were predicted for both non-polar and polar stationary phases; ** – RI were predicted only for stationary phases other than standard and semi-standard non-polar; *** – complex training scheme with two training sets, multiple data sets for testing.

Another type of features that can be used as input for molecule properties prediction using a machine learning model is molecular fingerprints (MF). These features can be considered as a type of molecular descriptors. MF is a set of binary [37-38] or, rarely, integer [10, 39] features that contains a few hundred or even thousands of features of the same nature. MF are usually based on substructure counts or local topological features of a molecule [37-38].

Besides these features, there are many research works that use more raw representations of a molecule. A SMILES string representation of a molecule can be used as input for 1D convolutional [16, 40-42] or recurrent neural network [42-43]. This approach performs well in chromatography-related tasks [16, 44]. It was recently shown [16] that 1D convolutional neural network (CNN) outperforms all MD-based approaches for the RI prediction. A depiction of a molecular structure with some preprocessing can be used as input for 2D CNN [45-46]. This representation was never applied for RI prediction but gives good results for prediction of other molecular properties. Finally, there are more complex methods for using a deep neural network directly with a molecular graph [47-49], such as molecular graph convolutional networks. A molecule can be featurized in many ways, as shown above, and each of these ways can be used as input for a model, and the simultaneous use of various representations and machine learning models will give better results than using only one model and representation [42, 45, 50].

The term “multimodal machine learning” means that a machine learning model simultaneously uses different “modalities” of the input object, for example: for video classification, different “modalities” can be sound and visual components [51]; for biochemical activity of a small molecule, “modalities” are information related to the molecule and to the biological system [52]. When predicting the property of a single molecule, the use of different representations of the structure (MF, SMILES, 2D sketch) is often called multimodal machine learning [50, 53]. These representations give different information about the structure and can be considered as different “modalities”. The joint use of different representations of a molecule can significantly improve the performance of a model. Considering RI prediction task, information about the chromatographic stationary phase can be regarded as the additional “modality”. Usually, when a model is called multimodal, features from different “modalities” are processed by different input layers of the neural network rather than concatenated together at the input stage. Multimodal machine learning was not used for accurate RI prediction before.

Model stacking is a technique when predictions of multiple models are used as input for a second-level model (or meta-model) that makes the final prediction [54-55]. This improves the accuracy of prediction. Model stacking is often used in tasks related to chemistry and biology. For example, it was used for prediction of small molecule-protein interactions [56], for disease prediction [57], and for other biochemistry-related tasks [58]. Model stacking was successfully used for prediction of retention time in liquid chromatography [59-60]. To the best of our knowledge, model stacking of multiple models was not used for RI prediction. Works comparing multiple different machine learning methods for RI prediction [16, 21, 61] usually do not discuss their stacking or simultaneous usage. However, there are works that use the average of outputs of two RI prediction models for GC-MS library search [8, 20]. It should be noted that terms “model stacking” and “multimodal machine learning” are near-orthogonal. Model stacking can use the same “modality” for all base-level models, in this case it will be model stacking but not multimodal machine learning.

The aim of this work is development and comparison of several different machine-learning models that use different representations of a molecule to predict gas chromatographic RI, the joint use of these models with a linear meta-learner for model stacking for even more accurate prediction, and testing for various external test data sets to determine the domain of applicability. For external testing, several data sets with flavor compounds, essential oils, metabolites and metabolomics-related compounds were selected. Unlike previous works [16-18, 21], our models take into account information about a stationary phase instead of considering all non-polar

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3045047, IEEE Access

stationary phases as equal. Such model can be considered as multimodal machine learning since it uses different representations of the structure (those can be considered [50, 53] as separate "modalities") and considers both GC-related modalities: the molecule structure and the stationary phase. The purpose of this work is to create the most accurate RI prediction method at the moment that uses only free and open source libraries for MD computation and that can be directly used in analytical chemistry and metabolomics.

II. METHODS

A. DATA SETS

The NIST 17 database was used as a primary data source for training, validation, and testing. Initially, NIST 17 contains 404045 RI data records for 99400 compounds. Some of them are stereoisomers (*cis-trans* and optical). We excluded some of these data from our data set. For 210 compounds, Chemistry Development Kit (CDK), version 2.3 [62] encounters problems when processing their structures from a structure file. This number includes cases when InChI-key generated on the basis of the parsed structure is different from InChI-key given directly from the NIST database. This means that CDK processes the structure inadequately. 152 compounds were excluded because they contain unsupported symbols in their SMILES string. Only C, c, N, n, H, O, o, F, B, l, r, S, i, +, (,), [,], -, =, #, 1, 2, 3, 4, 5, 6, 7, 8, 9, s, P, %, I, s symbols are supported. This set covers all common organic elements. Compounds excluded due to this reason contain uncommon elements (such as selenium), metals or consist of several parts that are not bonded by covalent bonds. 18 compounds were excluded because they have SMILES string representation longer than 250 symbols, or have 2D representation (depiction) larger than 65*65 units (see below). Also for 1174 compounds, we encountered problems extracting explicit chemical structures from the NIST 17 database.

Stereoisomers (*cis-trans* and optical) were treated and counted as identical compounds. All RI data records that correspond to standard polar stationary phases were also excluded. Finally, we obtained a data set with 309756 RI data records for 88675 compounds that contain RI for standard non-polar and semi-standard non-polar stationary phases.

For external testing and establishing the applicability domain, we used 8 data sets from various sources. Table 2 summarizes these data sets and shows the designations of the data sets that are used in this work. GMD and FIEHNLIB data sets consist of metabolomics-related compounds. The structures of the compounds in these data sets are given in the non-derivatized form, while RI for most of the compounds are actually given for the derivatized form. Derivatization (for example, substitution of -OH groups with $-OSi(CH_3)_3$ groups) was made to increase volatility of polar compounds. We retained only those RI data records that are given for trimethylsilyl- derivatives or underivatized compounds. Other types of derivatization were not supported. Also, we retained only those derivatized forms for which the number of -OH groups in a molecule before derivatization is equal to the number of attached $-OSi(CH_3)_3$ groups. Other data records were excluded because it is not possible to determine the exact structure of the derivatized form for which RI was measured. We assumed that all $-OSi(CH_3)_3$ groups replace all -OH groups (not $-NH_2$ or other functional groups) if numbers of -OH groups in the non-derivatized molecule and $-OSi(CH_3)_3$ groups in the derivatized molecule match each other. For some molecules, this assumption may not be true, but there is no way to determine the exact structure of the derivatized form using the available data. Also, several compounds were excluded from GMD and FIEHNLIB data sets using the same criteria that were used for the NIST 17 data set.

Lee retention indices [4] and FAME-based retention indices [5] were converted to Kovats retention indices using previously reported polynomial equations [17, 5], see also the Supplementary material, section S1. Stereoisomers were treated as identical compounds. We stored structures of compounds in data sets as SMILES strings without symbols that designate *cis-trans* and geometric (e.g., optical) isomers. Our script ensures that identical SMILES strings are created for identical structures. The data sets used for testing are shown in Fig. 1.

TABLE 2
DATA SETS USED FOR EXTERNAL TESTING OF THE MODEL

Designation	Description	Stationary phase	RI data records*	Ref.**	Source
ESSOILS	Essential oils	DB-5	2073	[14]	
OUF	Metabolomics-related compounds. Mostly derivatives with high trimethylsilyl- groups content	CP-SIL 8 CB	337	[22]	
FLAVORS	Flavors and fragrances	OV-101	1208	[29]	[63]
FLNET1	Flavor-related compounds	OV-101	297	[28]	[15]
FLNET5	Flavor-related compounds	DB-5	405	[28]	[15]
SET184	Aliphatic hydrocarbons, alcohols, ethers, ketones, and esters	Squalane and OV-1	184	[64]	[65-67]
GMD	Metabolomics-related compounds	5%-Phenyl methylsiloxane	531	[13]	
FIEHNLIB	Metabolomics-related compounds	RTX-5Sil	601	[5]	

* – Number of data records that were actually used in this work after exclusion of unsupported compounds. For most (but not all) compounds, there is one RI data record in one data set. ** – This column contains references to the source from which the data used in this work were actually collected. In four cases, there is a secondary source that provides ready-to-use data with SMILES strings. A reference to the original source of the data is given in the next column for such cases.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3045047, IEEE Access

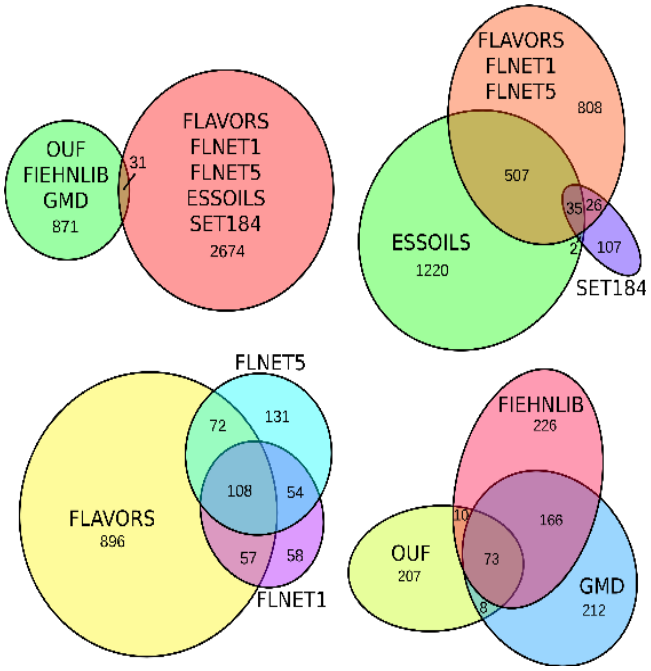

FIGURE 1. **Area-proportional Venn diagrams for the data sets used for testing in this work. Stereoisomers (*cis-trans* and optical) are counted as identical compounds.**

B. INPUT FEATURES FOR MACHINE LEARNING MODELS

For 1D CNN, we used one-hot encoded SMILES strings, as was previously published [16]. The number of possible symbols was the number of CNN channels (each possible symbol – one channel). For 2D CNN we created 2D coordinates of each atom and of the middle of each bond. 2D coordinates were created using CDK [62], and these coordinates correspond to the coordinates in 2D depiction (sketch) of the chemical structure. Molecules with depictions that do not fit into a square with dimensions 65*65 units were not considered. Other molecules were centered in this square, the square was split into cells with dimensions 0.5*0.5 units. This results in 130*130 cells. For each cell, 29 features were created. 26 one-hot features encode the type of an atom (if any atom is located in this cell), and 3 one-hot features encode the order of a bond if the center of any bond is located in this cell. If the cell does not contain bonds and atoms, all features are zero.

MD set includes all descriptors that are supported by CDK 2.3, except for 3D descriptors (i.e., descriptors that require pre-computed 3D coordinates) and two more descriptors: nAtomLAC and MolIP. These descriptors were not used because they are computed impractically slow with CDK 2.3 for some molecules. All other descriptors supported by CDK 2.3 were used. Some of them give NaN or throw exception for a small number of molecules. We use 0 as the value of descriptor in such rare cases. In addition to 243 descriptors computed by CDK, we used functional groups counters according to previous works [17, 68]. 84 features were created for each molecule, each feature means the number of occurrences of the respective fragment (functional group). It is implemented using 84 simple SMARTS queries. Fragments are not mutually exclusive. For example, a molecule that contains >N-N=O fragment also contains >N-, -N=O fragments. 243 MD created with CDK 2.3 and 84 functional groups features were concatenated together and are referred below as MD features or descriptors. We use additive extended-connectivity circular molecular fingerprints [10] with a diameter of 4 and a length of 1024. These MF are similar to usual extended-connectivity circular molecular fingerprints ECFP_4 [38] but consist of integer features instead of binary features.

Each RI data record in the NIST 17 database contains information about a stationary phase. There are 14 standard non-polar stationary phases and 20 semi-standard non-polar stationary phases for which there are at least 1000 data records. Stationary phases for which there are less than 1000 records are grouped into two types: “other standard non-polar” and “other semi-standard non-polar”. As a result, we consider 36 stationary phase types. Features containing information about the stationary phase consist of one-hot encoded stationary phase type (36 features) and information on whether the stationary phase is standard non-polar or semi-standard non-polar (an additional one-hot encoded feature).

Detailed information on input features for 2D CNN, supported and unsupported CDK descriptors, SMARTS patterns that correspond to the fragments used in this work, additive extended-connectivity circular molecular fingerprints, and types of stationary phases that are considered is given in the Supplementary material, section S2.

C. RETENTION INDEX PREDICTION USING MACHINE LEARNING

Three deep neural networks were used: 1D CNN, 2D CNN, and deep residual multi-layer perceptron with two inputs (MLP). Neural networks are shown in Fig. 2. Both CNN have a few convolutional layers followed by a global average pooling layer. Its output is concatenated with information about a stationary phase. MLP consists of two subnetworks with two separate inputs: the first one uses concatenated MD and information about a stationary phase, and the second one uses additive MF. Later there is a concatenation layer. The fingerprints-related subnetwork in MLP consists of an input dense layer followed by two residual blocks with two dense layers each. Four layers in two residual blocks use dropout with rate 5% (95% of connections are retained). Blocks are followed by an element-wise addition.

After a concatenation layer, in all three neural networks, there are two fully connected (dense) layers with 600 and 1 output nodes, respectively. The first fully-connected layer in the first neural network in MLP uses the hyperbolic tangent activation function. All three output

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3045047, IEEE Access

layers use the identity (linear) activation function. Everywhere else, the ReLU activation function is used.

Neural networks were trained using Eclipse Deeplearning4j framework [69], version 1.0.0-beta6. Optimization algorithm: Adam, weights initialization: ReLU, learning rate: 0.0003, objective loss function: mean absolute error, batch size: 16. All other hyperparameters are shown in Table 3. After running a given number of iterations, the parameters for that iteration for which there was the best MAE value for the validation set were saved and used for further testing. XGBoost [70], version 1.0.0 via XGBoost4j was also used. Concatenated MD and information about a stationary phase were used as input features. We made an extensive random search of hyperparameters. The actually used hyperparameters are shown in Table 3. Root mean square error was used as the objective function. 800 estimators were used without early stopping.

The outputs of all four base-level models are used as input for a linear meta-model (so called model stacking), see Fig. 2. Base-level models are listed in Table 3. The linear model was also trained using Deeplearning4j in order to minimize MAE.

Molecular descriptors were scaled in such a way that all descriptor values for compounds from the training set were in the range [0, 1]. However, it is possible that for compounds from the test set, MD will be slightly outside of this range. For training of the linear meta-model and for training of the neural networks, all RI values were divided by 1000.

For each of data sets used for testing, all compounds that are contained in the test set were excluded from the NIST 17 data set (all RI records for each compound). We ensure that there is no overlapping between the test set and the data sets that are used for training and validation. Then, the remaining training-validation data set

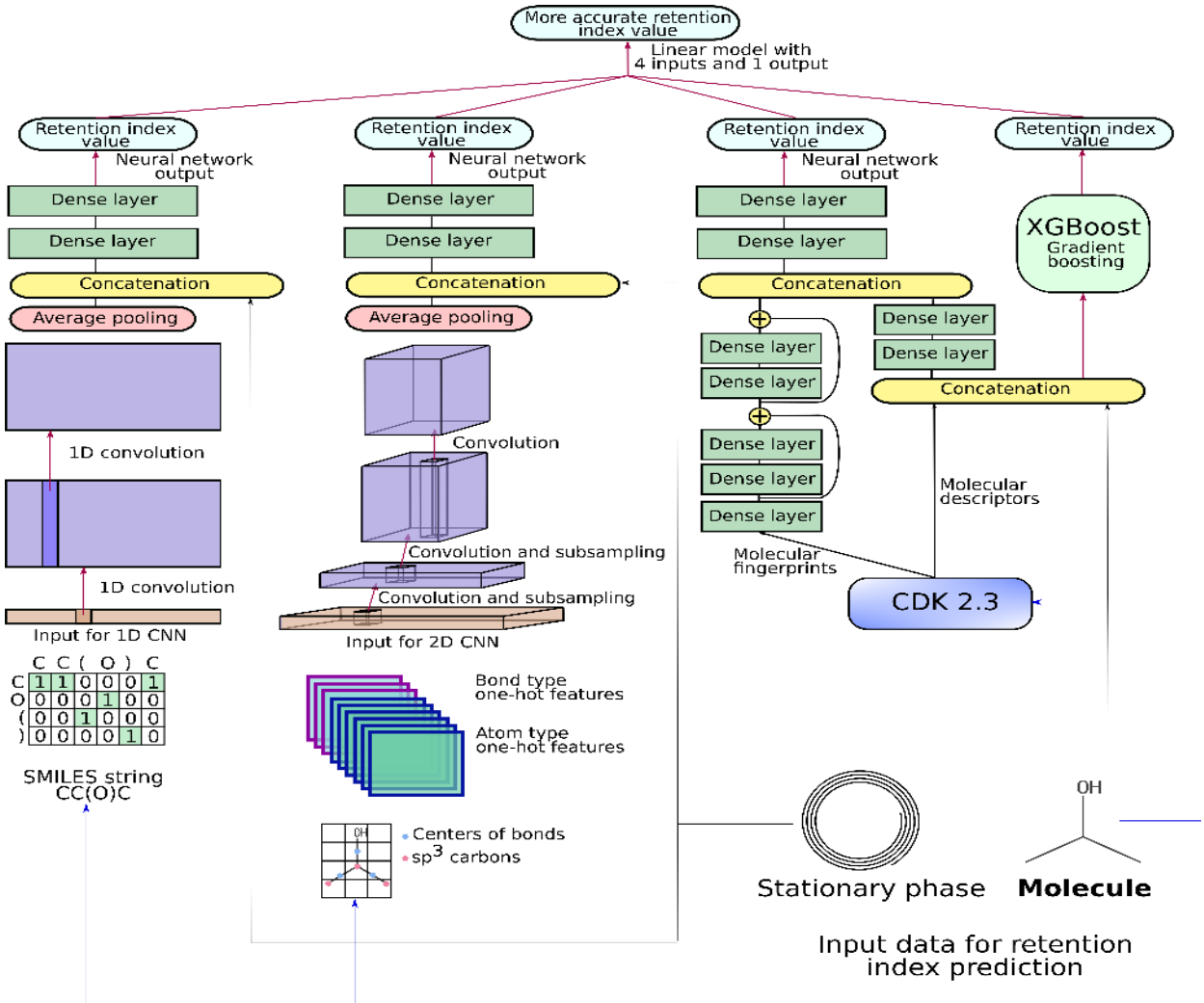

FIGURE 2. **Machine learning models used in this work. From left to right: 1D and 2D convolutional neural networks, deep residual multilayer perceptron, and gradient boosting (CNN1D, CNN2D, MLP, and XGBoost). Input data: one-hot encoded information about the stationary phase and various representations of the molecule structure. Outputs of four base-level models are used as inputs for a meta-learner.**

was split into a training set for the linear meta-model and a training-validation set for four first-level models (1:10, i.e., 10% of compounds were used for the linear meta-model). This training-validation set was split into a validation set (which is used for validation and monitoring of training) and a training set (1:20). Splits are shown in Fig. 3.

TABLE 3
HYPERPARAMETERS AND DESIGNATIONS OF BASE-LEVEL MODELS

Model	Hyperparameters
CNN1D	36 input channels; 2 1D convolutional layers. For both: kernel = 6, stride = 1, output channels = 300. Max number of iterations = 200000
CNN2D	29 input channels; 3 2D convolutional layers. For all: kernel = 4*4, stride = 1. Output channels: 50, 300, 300. First 2 2D convolutional layers are followed by MAX-pooling subsampling layers. Kernel and stride: 2*2 for both. Max number of iterations = 100000
MLP	In the descriptors-related subnetwork: 2 dense layers, first uses the TANH activation function. Both have 300 output nodes. In the fingerprints-related subnetwork: 5 dense layers with 1200 output nodes. Max number of iterations = 120000
XGBoost	eta = 0.05, gamma = 0.05, lambda = 0.05, max_depth = 21, min_child_weight = 21, subsample = 0.5, colsample_bytree = 0.5

All first-level models make more accurate predictions for compounds from the training set rather than for unseen compounds, but this effect is observed to varying degrees for different models. If the same training set is used for first-level models and for a meta-learner, the meta-learner will assign the largest weight to that first-level model which shows the best accuracy for the training set (i.e., to the most overfitted model) rather than to the actually most accurate model. To avoid this effect, we use separate training sets for first-level models and for a meta-learner.

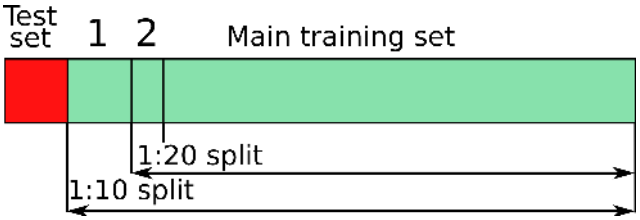

FIGURE 3. The data set split into the main training set, the meta-model training set (1), the validation set (2), and the test set. All splits are compounds-based, i.e., all data records for each compound are in one subset.

All splits are compounds-based. This means that all RI records for each compound (i.e., for each SMILES string) are contained in only one of the subsets. We used 10-fold cross-validation for the NIST 17 data set. It was split into 10 subsets, each of them was used for testing. All models were retrained from scratch for each of these subsets and for each of the external test sets. On average, for cross-validation splits, the training set, the meta-model training set, the validation set, and the test set contain 237512.7, 28345.2, 12922.5, 30975.6 data records and 68235.5, 7981, 3591, 8867.5 different compounds, respectively.

Supplementary material, section S3, contains further details about implementation and instructions about compiling and usage of pre-trained models. Some source code and pre-trained models are provided online: https://www.doi.org/10.6084/m9.figshare.12651680

III. RESULTS AND DISCUSSION

A. MODELS DEVELOPMENT

We tried multiple possible options during the development of the models. Using MD as an input for single-input MLP (without different subnetworks for MD and MF) does not allow achieving good accuracy [18]. We tried multiple setups for single-input MLP: we varied the number of layers in the range 2-5, nodes per layer (up to 2000), activation functions, regularization methods (L2, L1, dropout), residual connections. In all cases that we considered, single-input MLP performs worse than gradient boosting using the same data set and using the same feature set. Mean percentage error (MPE) is more than 3-3.2% for such models and subsets of NIST 17. For gradient boosting, MPE is about 2.7%. The error grows with the addition of more layers: 1-2 hidden layers work better for MD generated using CDK than deeper networks.

We noted that when MF (with multiple hundred to thousands of features of the similar nature) are used as input for MLP, deep neural networks with 5-10 layers and residual connections perform better than more shallow networks. The dual-input MLP that uses MF and MD together, with different network depth in these subnetworks, gives better accuracy (MPE is 2.0-2.3%) compared with single-input neural networks using the same input features.

It can be explained in the following way. MD and MF input feature vectors have a different nature. MD is a more relevant feature set, many MD are strongly correlated with RI value and characterize the molecule as a whole. MD are very heterogeneous, some of them are integer values, others are continual, and physical meanings of the features are very different. MF are much less relevant features – none of them are directly related to RI prediction. MF is a sparse vector of integers that count specific local structural features. This vector is less relevant but much more comprehensive and contains detailed information about a structure. For these feature vectors, very different configurations of a neural network are optimal: relatively shallow MLP with the TANH activation function (as proposed in previous works [9, 18]) after the input layer and much deeper residual MLP with the ReLU activation functions after all layers. The use of a neural network with two inputs allows using both MD and MF with near-optimal neural network depth and hyperparameters, and probably these features contain information that is complementary.

Finally, we made some random hyperparameters optimization. We tried other types of MF (PubChem fingerprints, MACCS fingerprints, binary extended-connectivity circular fingerprints) and tried different MF length (512-4096) and diameter. The advantage of additive

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3045047, IEEE Access

MF over binary MF can be explained by the regression nature of the task. Binary MF are more suitable for classification tasks. RI nearly linearly depends on the number of some substructures [17]. We varied the number of hidden nodes (500-2000), the number of layers in the residual block (2-4), the number of residual blocks (1-4), the activation functions (TANH and ReLU), and the dropout value.

Parameters of the 1D CNN model were very similar to previously reported [16]. We tried L2-regularization constant (l_2) values 0, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}. There is almost no difference in accuracy in the range 0-10^{-6}. The accuracy decreases with the growth of l_2. In our previous work [16], we made a random search of hyperparameters and obtained the best result for $l_2 = 10^{-5}$. This difference in behavior can be caused by the fact that we use a larger and less noisy data set in this work. This allowed us to increase the number of nodes and get rid of L2-regularization. The achieved MPE value is about 2.0%. More detailed comparison with previous works is given below. No additional methods to prevent overfitting (besides validation-based early stopping) were used. A more detailed study on regularization methods and hyperparameters of 1D CNN can be the subject of further research. We also tried to increase the number of layers but did not achieve considerable growth of the performance. Average pooling gives better accuracy than max pooling. This can also be explained by the regression, continual nature of the RI prediction task, similar to the advantage of additive MF over binary MF.

We tried multiple 2D CNN configurations and 2D representations of the molecule. We tried to use human-readable depiction of the molecule as neural network input, and a few simplified depictions, but all these models did not allow us to achieve MPE lesser than 3-3.5%. The important problem is the presence of structures with very long structural formulas in the NIST 17 library. There are many molecules with very long (20-50 atoms) linear chain fragments, so we should use a large (500*500 and more) input image or scale the image, or distort such “long” structures. Finally, we found that a multi-channel representation with a low spatial resolution containing some chemical information is more suitable for our task (MPE is about 2.9%). However, with the low spatial resolution, the rounding of atom coordinates significantly affects them. This “distorts” the 2D geometrical shape of functional groups and sometimes leads to the fact that different unconnected atoms have “identical” coordinates.

The relatively poor accuracy of 2D CNN compared with 1D CNN can be explained by the fact that a larger part of a molecule fits into the convolution kernel for 1D CNN compared with 2D CNN. Kernel = 6 for 1D CNN allows the neural network to extract at the input layer features with a size of 3-6 atoms. These are substructures the size of which is comparable with the size of substructures accounted by typical MF. This allows relatively shallow and easy-to-train 1D CNN without subsampling layers to extract enough relevant and coarse-grained features. 2D CNN requires a kernel with dimensions at least 4*4 to extract features the size of a functional group. The 2D CNN filter has many more parameters and extracts smaller features compared with the 1D CNN filter. The low spatial resolution, deeper network, and subsampling layers partially solve this problem. However, the low resolution leads to other problems, as described above. Detailed neural network interpretation, such as visualization of feature maps, search of inputs leading to maximal activation of certain CNN channels, was outside the scope of this work. However, it is possible for both 2D CNN and 1D CNN. Some recent works [71-72] are devoted to interpretable neural networks that predict molecular properties.

We tried to construct neural networks with 3 and 4 inputs, i.e., to combine our networks together the similar way as we did with MD and MF subnetworks in our MLP model. However, such combination does not allow us to achieve significantly better results than usage of model stacking and simple linear meta-model. As we noted above, single-input MLP does not give enough accuracy when MD are used as input representation. We tried other machine learning methods: random forest, regression tree, support vectors regression. The best results were achieved using gradient boosting. We also tried to use more input features (e.g., concatenate the input vector with MF), but there was no large performance gain.

MLP and CNN1D models are the most accurate and have close accuracy (MPE is about 2.0-2.1%). CNN2D and XGBoost are less accurate and also have similar accuracy (MPE is about 2.7-2.9%). Model stacking gives significant accuracy gain and allows achieving MPE about 1.8%. Detailed data on achieved accuracy of the finally developed models are given below in the following sections.

All base-level models are prone to overfitting but are prone to overfitting to varying degrees. We selected such values of hyperparameters that provide the best accuracy for the validation set. The accuracy for the training set is much better using these hyperparameters. The XGBoost model demonstrates the largest difference between accuracies for the training and test sets (MPE 1.0% and 2.4%, respectively). When we used the same training data set for base-level models and for a meta-learner, we observed that the meta-learner severely overweights XGBoost and assigns the largest weight to it. At the same time, for the unseen compounds, XGBoost is less accurate than CNN1D and MLP. As a result, model stacking gives almost no accuracy gain in this case, and only the use of separate training sets for base-level models and meta-learner allows achieving the best accuracy.

For successful using in model stacking, models must be accurate and their errors (differences between predicted and reference values) should not be strongly correlated. Fig. 4 shows correlation plots between errors obtained using CNN1D and errors obtained using other three

models. In general, the errors are strongly correlated. Three plots at the left in Fig. 4 show that there are many cases when all four models produce RI values that deviate from reference values very significantly: hundreds and even thousands of units. We call such cases "outliers". The most probable explanation is that in these cases the database contains wrong experimental data or wrong chemical structures, since for many of them all four models give a very similar prediction.

We did not use any "accuracy-based" exclusions of data records and included these values when computing accuracy. Preliminary experiments also show that removing outliers from the training set does not give significant accuracy gain for the test set (with outliers). We consider "accuracy-based" exclusions from the test sets as unfair. However, to consider correlations between errors, we excluded all data points for which all four models give an error of more than 100 units. The correlation plots for this case are shown in three plots at the right in Fig. 4. Errors are still moderately correlated.

TABLE 4
CORRELATION COEFFICIENTS BETWEEN RI PREDICTION ERRORS FOR DIFFERENT MODELS. ONLY THOSE DATA RECORDS WERE USED FOR WHICH AT LEAST ONE OF THE FOUR MODELS GIVES AN ERROR OF LESS THAN 100 UNITS

	CNN1D	CNN2D	MLP	XGBoost
CNN1D	1.00	0.45	0.53	0.31
CNN2D	0.45	1.00	0.34	0.30
MLP	0.53	0.34	1.00	0.28
XGBoost	0.31	0.30	0.28	1.00

The correlation coefficients are given in Table 4. XGBoost errors are less correlated with neural network errors than neural network errors with each other. The XGBoost model uses local structural peculiarities less than other models. 20 MD with the largest feature importance are: BCUTp-1l, BCUTc-1l, BCUTc-1h, ATSc4, ATSc5, ECCEN, ATSc3, AMR, BCUTp-1h, MDEC-23, ATSp1, ATSc2, ATSc1, WTPT-2, MDEC-22, MDEC-12, AlogP, XlogP, tpsaEfficiency, WPATH. The meaning of the descriptors is explained in CDK 2.3 documentation. Each of them characterizes the molecule as a whole rather than any spatially local features. Most of them are complex topological descriptors. The XGBoost model mostly relies on such features, while other models mostly rely on spatially local features. This is one of the possible causes why errors of the XGBoost model are less correlated with errors of other models. There are multiple previous works discussing the causes why RI strongly depends on certain descriptors [24, 28-29].

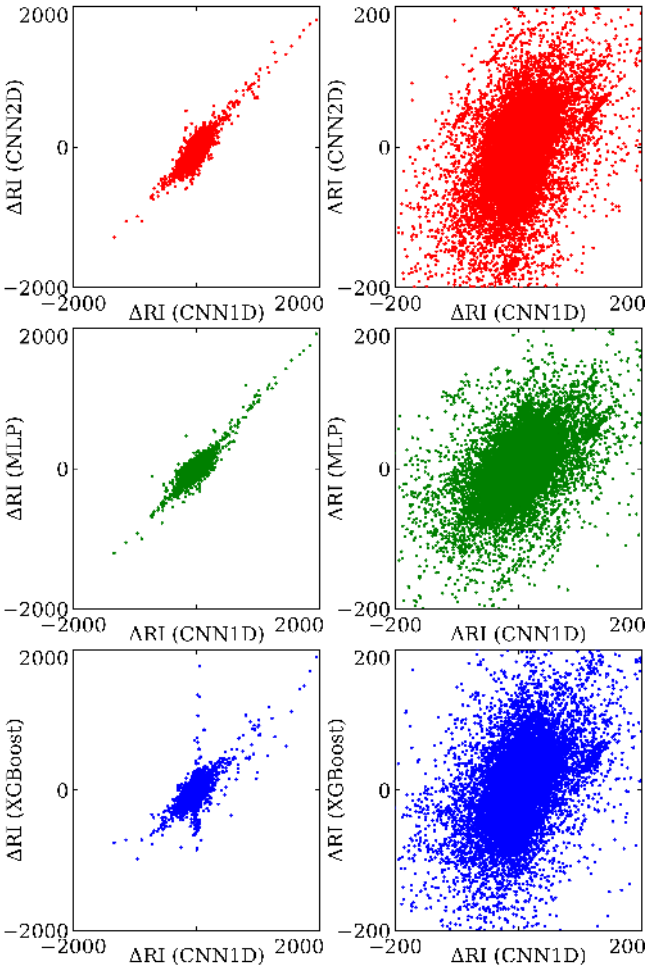

FIGURE 4. Correlation plots between prediction errors obtained using various RI prediction models. Three plots at the left show errors for all data records in the considered test set, and three plots at the right show errors only for those data records for which at least one model gives a prediction error of less than 100 units. Random 10% subset of the NIST 17 library is used as a test set.

B. ACCURACY OF THE FINALLY DEVELOPED RETENTION INDEX PREDICTION MODELS

Fig. 5 shows the distribution of MAE and MdAE for 10 test subsets of the NIST 17 data set that were used for cross-validation and for all finally developed models. For each of the subsets, the stacking model shows better results than the other four models. The accuracies of CNN1D and MLP models are very close to each other. The third most accurate model is XGBoost, and CNN2D is the last one. For each of 10 subsets, these two models are less accurate than the other two, but their accuracies are close to each other. Table 5 shows MAE and MdAE for all finally developed models and all data sets. For NIST 17, the overall result of cross-validation is shown. In this work, we use 8 external test data sets. These test sets differ from each other in the chemical nature of the compounds that these sets consist of.

Fig. 6 shows the distribution of data records in data sets by RI value, the principal components plot that shows the diversity of data used for external testing, and the correlation plot between molecular weight and RI for

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3045047, IEEE Access

TABLE 5
ABSOLUTE ERRORS (MEAN AND MEDIAN) FOR MODELS AND DATA SETS CONSIDERED IN THIS WORK

Data set	CNN1D		CNN2D		MLP		XGBoost		Stacking model	
	MAE	MdAE	MAE	MdAE	MAE	MdAE	MAE	MdAE	MAE	MdAE
NIST 17	31.5	16.5	44.5	25.5	30.8	17.1	41.3	24.2	27.7	14.4
Training set *	24.1	11.9	30.8	14.4	20.8	10.5	16.8	8.1	-	-
ESSOILS	36.1	22.0	44.5	28.7	35.4	22.2	41.1	28.6	31.3	18.2
OUF	56.8	39.2	63.8	49.8	54.2	40.6	70.7	57.8	51.7	39.5
FLAVORS	26.8	11.9	34.9	18.2	27.9	13.4	38.8	22.8	25.2	10.9
FLNET1	24.5	12.5	30.4	17.2	25.6	13.5	32.6	23.0	22.8	9.8
FLNET5	23.2	13.6	32.3	21.4	25.0	17.1	31.9	21.4	21.8	13.2
SET184	9.6	7.2	19.6	15.6	13.3	8.3	16.6	12.3	10.0	6.3
GMD	65.4	43.4	67.0	42.6	56.7	32.9	72.9	49.5	55.2	31.3
FIEHNLIB	98.6	38.4	103.5	49.9	99.7	39.9	107.3	52.3	92.5	34.7

* – An example of accuracy for a training set. Accuracy for NIST 17 is the result of cross-validation; other than NIST data sets are hold-out test sets.

TABLE 6
PERCENTAGE ERRORS (MEAN AND MEDIAN) FOR MODELS AND DATA SETS CONSIDERED IN THIS WORK

Data set	CNN1D		CNN2D		MLP		XGBoost		Stacking model	
	MPE	MdPE	MPE	MdPE	MPE	MdPE	MPE	MdPE	MPE	MdPE
NIST 17	2.05	1.18	2.86	1.85	2.04	1.25	2.73	1.76	1.80	1.04
ESSOILS	2.45	1.57	3.04	2.12	2.43	1.62	2.87	2.04	2.13	1.33
OUF	3.11	2.34	3.56	2.93	2.99	2.35	4.00	3.40	2.83	2.20
FLAVORS	2.24	1.03	2.95	1.61	2.36	1.18	3.35	1.94	2.12	0.97
FLNET1	2.28	1.14	2.78	1.61	2.34	1.26	3.07	1.99	2.10	1.03
FLNET5	2.16	1.31	3.01	2.07	2.30	1.46	3.00	2.06	2.00	1.17
SET184	1.40	0.94	2.87	2.00	1.96	1.19	2.49	1.62	1.50	0.84
GMD	3.43	2.27	3.47	2.46	2.92	1.71	3.83	2.71	2.87	1.71
FIEHNLIB	5.06	2.15	5.40	2.68	5.12	2.33	5.71	2.82	4.80	2.01

external test sets. Most of the compounds in FLNET5, FLNET1, FLAVORS, ESSOILS data sets are volatile compounds consisting of carbon, hydrogen, and oxygen. Oxygen content is low in most of the compounds. Only ~5% of the data records are given for nitrogen-containing compounds. None of the compounds contain silicon.

In OUF, GMD, and FIEHNLIB, on the contrary, most of the compounds are trimethylsilyl- derivatives of polar organic compounds in which -OH groups are replaced by $-OSi(CH_3)_3$ groups. ~44% of the data records are given for nitrogen-containing compounds in these data sets. The accuracy for these metabolomics-related highly polar derivatized compounds is much worse than for compounds from the other group of data sets. MAE values are strongly dominated by very few distant outliers. This is especially important for small data sets. The notable example is FIEHNLIB. For this data set, value of MAE is 92.5, that is much more than for OUF data set. MdAE for this data set is close to values of MdAE for other metabolomics-related data sets. One of the possible reasons for the low accuracy for metabolomics-related data sets is the incorrect elucidation of the structure of the derivatized form. As explained above, we consider only those compounds for which we can propose it with a high degree of confidence, but this procedure can still give wrong structures. However, for OUF data set, the accuracy is close to other data sets. Structures of derivatized forms were manually created by the authors [22] for this data set.

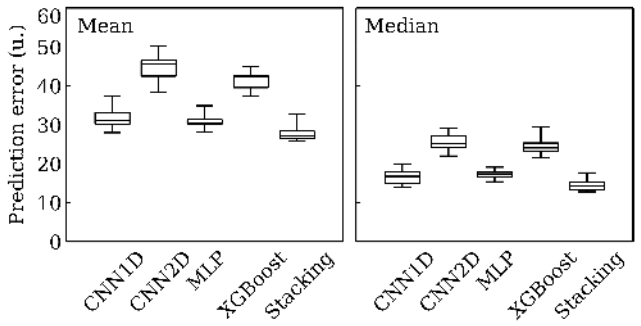

FIGURE 5. Mean and median absolute errors obtained using considered RI prediction models. These box-and-whiskers plots show the distribution of mean and median errors for 10 subsets of NIST 17 that were used for 10-fold cross-validation. Boxes and whiskers show the distribution through quartiles (full data range). The stacking model is more accurate compared with other models.

Examples of MAE and MdAE values for base-level models for the main training set are also given in Table 5. The accuracy for the compounds used for training is much better than for unseen compounds. This is typical for gradient boosting and deep neural networks. For the training set, XGBoost shows the best accuracy compared with neural networks. Coefficients of determination (R^2) for ESSOILS, OUF, FLAVORS, FLNET1, FLNET5, SET184, GMD, FIEHNLIB external test sets and for the stacking model are

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3045047, IEEE Access

0.98, 0.98, 0.95, 0.98, 0.99, 0.99, 0.98, 0.88, respectively; and root mean square errors (RMSE) are 52, 72, 72, 40, 35, 15, 92, 214, respectively.

Taking into account significant differences in RI values (see Fig. 6) from set to set, we also compared MPE and median percentage error (MdPE). These values are given in Table 6. All main trends are the same as for MAE and MdAE. All these measures can be used to compare the models. SET184 data set significantly differs from other data sets. It consists of compounds with lesser RI values and molecular weights (see Fig. 6). All of them are aliphatic and consist only of hydrogen, carbon, and oxygen. For this data set, all models predict retention more accurately than for other test sets in terms of both mean and median errors. For this data set, CNN1D performs much better than other models and even outperforms the stacking model in terms of MAE and MPE. In terms of median errors, the stacking model still performs better.

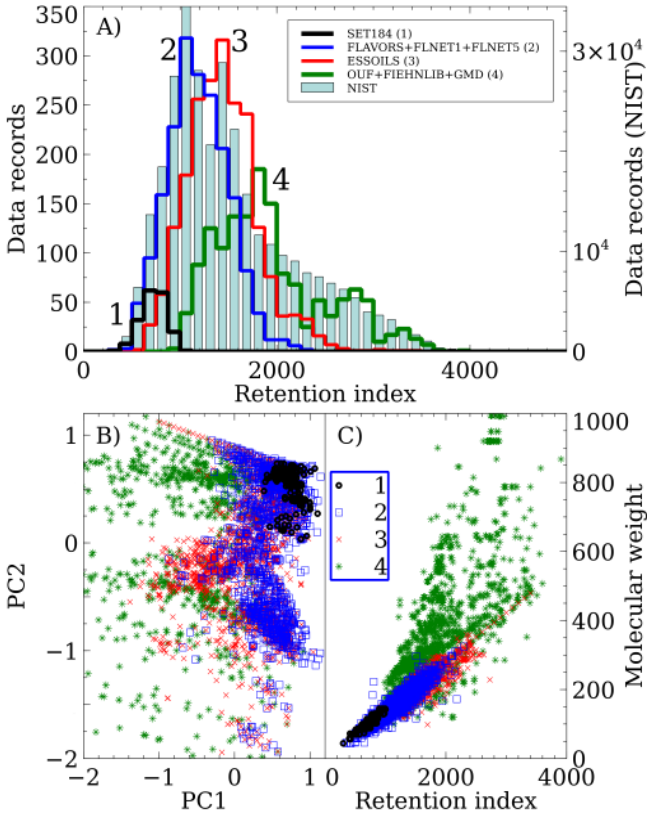

FIGURE 6. The diversity of data used in this work. (A) The distribution of RI data records in data sets by retention index value. The bars show the distribution for the NIST 17 data set, the curves for other data sets that were used for testing. (B) The principal components plot for external test data sets. Principal components were calculated using molecular descriptors used in this work (including functional groups counters). The first and the second principal components are shown. (C) The scatter plot that shows molecular weights and retention indices of compounds from external test data sets. Numbers denote data sets: 1 – SET184, 2 – flavor-related compounds (FLAVORS, FLNET1, FLNET5 data sets together), 3 – ESSOILS, 4 – metabolomics-related compounds (OUF, FIEHNLIB, GMD data sets together).

For all other test sets, the stacking model performs better than all other models in terms of all accuracy measures. Box-and-whiskers plots, similar to Fig. 5 for 8 external test sets, are shown in Fig. 7. Supplementary material, section S4, contains correlation plots between experimental and predicted values for various test sets. Fig. 8 shows the distribution of errors for various data sets for all models.

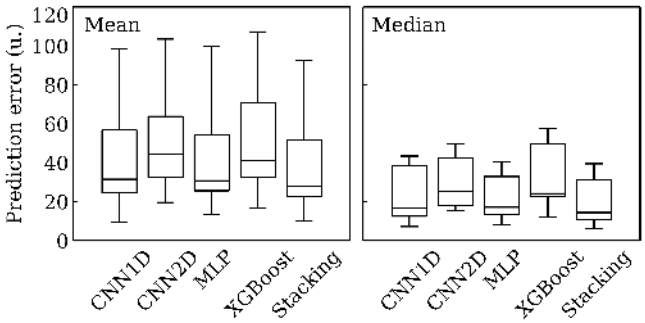

FIGURE 7. Mean and median absolute errors obtained using considered RI prediction models. These box-and-whiskers plots show the distribution of mean and median errors for 8 external test sets that were used in this work. Boxes and whiskers show the distribution through quartiles (full data range). The relatively large variation of the prediction errors over various test sets is caused by the diversity of used external test data sets.

C. COMPARISON OF PREDICTION ACCURACY WITH PREVIOUS RESULTS

Table 7 shows the accuracy values for the test sets that were reported in previous works. For ESSOILS, FLAVORS, GMD, we give both the result of our previous work [16] and the values that were reported previously. For OUF and FLAVORS data sets, only RMSE values were reported in previous works devoted to retention index prediction. We obtained RMSE values 80.1, 86.5, 74.8, 92.0, 71.9 using CNN1D, CNN2D, MLP, XGBoost, and the stacking model, respectively, for OUF data set. For FLAVORS data set, these five values are: 73.3, 76.9, 73.4, 79.6, 72.1. RMSE values are strongly dominated by distant outliers (i.e., cases when the prediction error is very huge), some of them are errors in the reference data.

All four of our models perform better or at approximately the same level as the previously reported models. CNN1D performs better than all previously reported models for all data sets except GMD. For GMD, it shows almost the identical accuracy with the previously reported accuracy [16]. The CNN1D model is very close to that model [16] but trained using a less noisy and larger data set.

The stacking model shows significantly better results than all previously reported models for all considered data sets. To our best knowledge, it seems to be the most accurate versatile RI prediction model at this moment for non-polar stationary phases. Probably, there are application specific models that show better accuracy for a limited narrow class of chemical compounds such as alkylbenzenes or FAME, but such models do not cover the entire chemical space and do not work well for diverse compounds. However, we used SET184 data set to demonstrate that our model performs well even for narrow data sets for which quite accurate models were previously developed.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3045047, IEEE Access

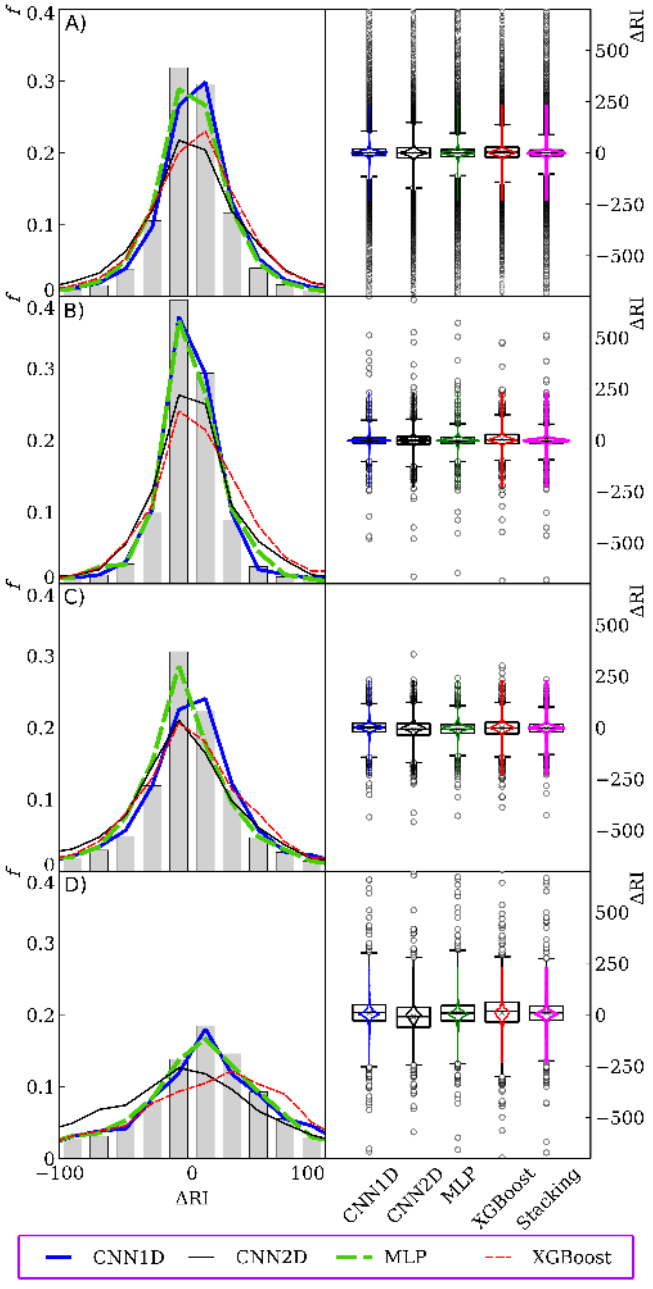

FIGURE 8. The distribution of RI prediction errors for 5 models and 4 data sets. Fraction of data records *f* in a bin is shown. A – NIST 17 (obtained using cross-validation), B – flavor-related compounds (FLAVORS, FLNET1, FLNET5 data sets together), C – ESSOILS, D – metabolomics-related compounds (OUF, FIEHNLIB, GMD data sets together). The bars show the distribution for the stacking model, the curves for other models. The boxes denote the median error, the whiskers denote the error range within which 98% of all data entries fall.

The direct face-to-face comparison of the accuracy of the NIST data set with previous works is complicated. To the best of our knowledge, all but one [18] of the comparable previous works [16-17, 21] used different versions of NIST. But even more important problem is that the previous works [16-18, 21] calculate accuracy using compounds-based data set and do not use information about the stationary phase. This means that for each compound all data records are grouped, and mean or median value is used as reference. Unlike the test sets considered above, the NIST 17 library contains multiple values (~4 on average) for each compound. This number varies from 1 (for most compounds) to more than 100. We calculate accuracy using individual RI data records. Using "averaged" data records for testing, one per compound, will decrease the accuracy of the model because we use information about the stationary phase. Using data records instead of averaged values for compounds, on the one hand, increases the number of distant outliers caused by incorrect reference data, on the another hand, it increases the role of well-studied compounds (for which there are many reference RI values). It is not clear how these factors affect accuracy.

TABLE 7
COMPARISON OF PREDICTION ACCURACY WITH PREVIOUS WORKS

Data set	Previously reported accuracy	ΔMAE*, %
NIST**	MAE = 58.4, MdAE = 34.3, MPE = 3.04%, MdPE = 1.68% [18] (NIST 17); MAE = 33.2, MdAE = 18.0, MPE = 1.96%, MdPE = 1.03% [16] (NIST 08); MdAE = 46.0, MdPE = 3.2% [17] (NIST 05); RMSE = 90-115 [21] (NIST 05)	16.6
ESSOILS	MAE = 43.5, MdAE = 28.6, MPE = 3.03%, MdPE = 2.08% [16]; MdPE = ~2.5-2.7% [9]	28.0
OUF	RMSE = 78-88; R^2 = 0.93 [22]	-
FLAVORS	MAE = 34.3, MdAE = 18.8, MPE = 2.93%, MdPE = 1.54% [16]; RMSE = 88.2 [29]	26.5
FLNET1	MAE = 50.3, MdAE = 48.4, MPE = 4.71%, MdPE = 4.22% [28]	54.7
FLNET5	MAE = 51.0, MdAE = 45.7, MPE = 4.76%, MdPE = 4.03% [28]	57.3
SET184	MAE = 11.4, MdAE = 8.7, MPE = 1.67%, MdPE = 1.20% [64]; MPE = 2% [67]	12.3
GMD	MAE = 63.6, MdAE = 38.1, MPE = 3.39%, MdPE = 2.15% [16]; MdPE = ~2.5-2.7% [9]	12.2

* – Relative accuracy gain achieved in this work compared with the best of previous works, in terms of MAE ($100\%*(MAE_{previous\ work}-MAE_{this\ work})/MAE_{previous\ work}$). ** – Different versions of NIST were used in different works.

For comparison with the previously trained models that do not take into account the exact type of a stationary phase, we trained two of them [16-17] using the NIST 17 data set. We used a data set with one reference RI value for each compound. Such data sets were used in works [16-17]. But we tested these models using the same test set (multiple RI records for one compound) as for the models reported in this work. The results of the comparison and further details are given in the Supplementary material, section S5. CNN1D, MLP, and the stacking model outperform previously reported models in this comparison. This is expected because, as far as we know, the best previously published model [16] is CNN, which is close to the CNN1D model reported in this work, but it does not take into account information about the stationary phase, has fewer channels in CNN layers (120 instead of 300) and nodes in a dense layer (200 instead of 600), and uses L2-regularization ($l_2 = 10^{-5}$).

We studied how l_2 value affects the accuracy of CNN1D. When we use the NIST 17 data set for training, MAE increases with growth of l_2 value. The best accuracy was observed with $l_2 = 0$. The comparison is given in tabular form in the Supplementary material, section S5.

D. COMPARISON OF PREDICTION ACCURACY WITH ACCURACY OF REFERENCE DATA

Finally, we attempted to estimate the accuracy of experimental data. For pairs of data sets that use the same stationary phases, we selected overlapping subsets, i.e., subsets of the compounds that are contained in both data sets. For each subset, two different prediction accuracies can be calculated: using the first and using the second data set as a source of reference values. Both prediction accuracy values and deviations of reference (experimental) data are shown in Table 8 for a few pairs of data sets. Correlation plots for experimental and predicted data for these subsets are shown in the Supplementary material, section S4.

The accuracy of prediction is (as expected) still worse than the accuracy of experimental data. However, for a pair of FLNET1-FLAVORS data sets, the deviations between reference and predicted values are only slightly worse than the deviations between reference values from different sources. For GMD-FIEHNLIB pair, a quite large deviation between the experimental values is observed. This pair of data sets uses different types of semi-standard non-polar stationary phases and different types of RI: FAME-based RI and Kovats RI. In this comparison, all RI values were converted to Kovats RI, but this conversion introduces some error [5]. For GMD-OUF pair, the deviation between the experimental values is unexpectedly low. Despite the fact that predicted RI have worse accuracy than experimental reference values, the prediction accuracy closely approaches the accuracy of experimental data for some data sources and classes of compounds.

TABLE 8
COMPARISON OF PREDICTION ACCURACY AND DEVIATIONS BETWEEN DATA FROM VARIOUS SOURCES

Data sets	Deviations of experimental data		Prediction accuracy	
	MAE	MdAE	MAE	MdAE
FLNET1-FLAVORS	17.9	6	16.7-22.3	6.9-8.1
FLNET5-ESSOILS	13.3	5	17.6-18.8	11.1-12.6
GMD-FIEHNLIB	41.2	18.1	47.5-56.8	26.3-31.0
GMD-OUF	11.0	2.6	28.6-41.7	19.4-24.3

Further improvement in prediction accuracy is significantly complicated by noise in training data. It seems unlikely to create a versatile (near-universal) RI prediction method that will predict with accuracy of ~5 units in terms of MdAE. Further possible improvement in accuracy can be achieved by using more information about the separation conditions (e.g., temperature), by using molecular graph convolutional networks, by using support vectors regression and other machine learning methods together with gradient boosting. However, all these efforts will be limited by the accuracy of experimental data.

IV. PRACTICAL APPLICATION AND FURTHER RESEARCH DIRECTIONS

The primary application of retention index prediction is augmenting of spectral libraries for GC-MS library search [8] and identification of metabolites based on RI, mass spectra, and list of candidates [9]. In both cases, RI data are used together with mass spectra. The use of RI for GC-MS library search was recently discussed in detail as well as the dependence of search accuracy on RI accuracy [8]. Another important application of this work is the quality control of experimental databases and the detection of wrong experimental data. In this regard, it is important that we develop several different models that are trained independently. If all models give results close to each other, which significantly differ from the “experimental” value in the database, it is probably incorrect reference data. This approach is not useful for detecting minor experimental errors but can help detect errors caused, for example, by wrong structure annotation.

Further research directions can include further improvement in accuracy using more models for stacking and retention index prediction for stationary phases other than standard and semi-standard non-polar. These can be polar stationary phases, semi-polar stationary phases (such as DB-624, DB-1701), ionic liquid stationary phases. Only relatively small data sets are available for these phases. This fact makes the prediction task more difficult (small training sets) and even more important. Transfer learning techniques can be used in this case. These models can also be used for liquid chromatography, for which large data sets recently became available [73]. Another direction of further research is the elaboration of more detailed explanations of how and why these models work: a more detailed analysis of the importance of certain descriptors, a detailed research how hyperparameters affect accuracy, a study of feature maps. A better understanding of how models works, rather than using them as “black boxes”, probably will allow achieving better accuracy. Graph convolutional neural networks can also be used to improve accuracy. At the moment this model and most other models for prediction of retention index consider optical and *cis-trans* isomers as identical compounds. The possibility of stable conformational isomers with different retention time is also ignored. Taking geometric isomerism into account can be a direction of further research.

V. CONCLUSIONS

Four machine learning models were developed for prediction of gas chromatographic retention indices: 1D and 2D convolutional neural networks, deep residual multilayer perceptron, and gradient boosting. Each of these models

perform at the same level as the best previously reported models or better. The linear meta-model can be applied to combine the results of these models and to obtain even more accurate predictions. The stacking model outperforms four base-level models and any of the previously reported machine learning models for retention index prediction. This is true for various compounds: both for volatiles with a low content of atoms other than hydrogen and carbon as well as for trimethylsilyl- derivatives of highly polar compounds. For some external test data sets, the accuracy of the model approaches the accuracy of the experimental data that were estimated by comparing data from different sources. It was also shown that the use of information about the type of the stationary phase allows improving the prediction compared with considering all standard and semi-standard non-polar phases as equal. Further model improvement is complicated by random errors in the experimental data.

Compared with previous works, we achieved significantly better accuracy for various test data sets and proposed two new accurate RI prediction models: 2D convolutional neural network and multilayer perceptron with two inputs. In previous works, different models were trained for different stationary phases, or the difference between similar stationary phases was not made. Unlike previous works, information not only about compounds but also about the stationary phase was used by our models. It is the use of different "modalities": various representations of the molecule and information about the stationary phase that allows achieving the best accuracy. This work is the first application of multimodal machine learning to RI prediction. We share source code and parameters of the pre-trained models. The models are ready for use by metabolomics scientists and analytical chemists. Unlike some of previous works, we do not use non-free proprietary software for computation of molecular descriptors.

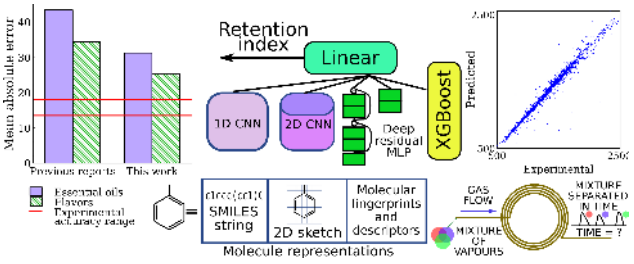

FIGURE 9. Graphical illustration of the conception and major findings of this work. Gas chromatographic separation and the machine learning model used in this work are schematically depicted. The bar plot shows the prediction accuracy (MAE) for FLAVORS and ESSOILS data sets obtained in this and previous works and rough estimation of the accuracy of the experimental data. The scatter plot shows the correlation between predicted and reference values for these two data sets.

Graphical illustration of the conception and major findings of this work is shown in Fig. 9. Developed models can be used for GC-MS library search, for GC-MS identification of compounds using in silico methods, for experiment design development, and for detection of possible wrong data in databases. Some source code and pre-trained models parameters are provided online: https://www.doi.org/10.6084/m9.figshare.12651680

REFERENCES

[1] G. Tarján *et al.*, "Thirtieth anniversary of the retention index according to Kováts in gas-liquid chromatography," *Journal of Chromatography A*, vol. 472, pp. 1–92, Jan. 1989, doi: 10.1016/S0021-9673(00)94099-8.

[2] B. d'Acampora Zellner, C. Bicchi, P. Dugo, P. Rubiolo, G. Dugo, and L. Mondello, "Linear retention indices in gas chromatographic analysis: a review," *Flavour Fragr. J.*, vol. 23, no. 5, pp. 297–314, Sep. 2008, doi: 10.1002/ffj.1887.

[3] E. Kováts, "Gas-chromatographische Charakterisierung organischer Verbindungen. Teil 1: Retentionsindices aliphatischer Halogenide, Alkohole, Aldehyde und Ketone," *HCA*, vol. 41, no. 7, pp. 1915–1932, 1958, doi: 10.1002/hlca.19580410703.

[4] D. L. Vassilaros, R. C. Kong, D. W. Later, and M. L. Lee, "Linear retention index system for polycyclic aromatic compounds," *Journal of Chromatography A*, vol. 252, pp. 1–20, Jan. 1982, doi: 10.1016/S0021-9673(01)88394-1.

[5] T. Kind *et al.*, "FiehnLib: Mass Spectral and Retention Index Libraries for Metabolomics Based on Quadrupole and Time-of-Flight Gas Chromatography/Mass Spectrometry," *Anal. Chem.*, vol. 81, no. 24, pp. 10038–10048, Dec. 2009, doi: 10.1021/ac9019522.

[6] M. Vinaixa, E. L. Schymanski, S. Neumann, M. Navarro, R. M. Salek, and O. Yanes, "Mass spectral databases for LC/MS- and GC/MS-based metabolomics: State of the field and future prospects," *TrAC Trends in Analytical Chemistry*, vol. 78, pp. 23–35, Apr. 2016, doi: 10.1016/j.trac.2015.09.005.

[7] X. Domingo-Almenara, J. Brezmes, G. Venturini, G. Vivó-Truyols, A. Perera, and M. Vinaixa, "Baitmet, a computational approach for GC–MS library-driven metabolite profiling," *Metabolomics*, vol. 13, no. 8, p. 93, Aug. 2017, doi: 10.1007/s11306-017-1223-x.

[8] D. D. Matyushin, A. Yu. Sholokhova, A. E. Karnaeva, and A. K. Buryak, "Various aspects of retention index usage for GC-MS library search: A statistical investigation using a diverse data set," *Chemometrics and Intelligent Laboratory Systems*, vol. 202, p. 104042, Jul. 2020, doi: 10.1016/j.chemolab.2020.104042.

[9] F. Qiu, Z. Lei, and L. W. Sumner, "MetExpert: An expert system to enhance gas chromatography–mass spectrometry-based metabolite identifications," *Analytica Chimica Acta*, vol. 1037, pp. 316–326, Dec. 2018, doi: 10.1016/j.aca.2018.03.052.

[10] J. N. Wei, D. Belanger, R. P. Adams, and D. Sculley, "Rapid Prediction of Electron–Ionization Mass Spectrometry Using Neural Networks," *ACS Cent. Sci.*, vol. 5, no. 4, pp. 700–708, Apr. 2019, doi: 10.1021/acscentsci.9b00085.

[11] H. Ji, H. Deng, H. Lu, and Z. Zhang, "Predicting a Molecular Fingerprint from an Electron Ionization Mass Spectrum with Deep Neural Networks," *Anal. Chem.*, vol. 92, no. 13, pp. 8649–8653, Jul. 2020, doi: 10.1021/acs.analchem.0c01450.

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

[13] J. Kopka *et al.*, "GMD@CSB.DB: the Golm Metabolome Database," *Bioinformatics*, vol. 21, no. 8, pp. 1635–1638, Apr. 2005, doi: 10.1093/bioinformatics/bti236.

[14] R. P. Adams, *Identification of essential oil components by gas chromatography/mass spectorscopy*, 4th ed. Carol Stream, Ill: Allured Pub. Corp, 2007.

[15] H. Arn and T. E. Acree, "Flavornet: A database of aroma compounds based on odor potency in natural products," in *Developments in Food Science*, vol. 40, Elsevier, 1998, p. 27.

[16] D. D. Matyushin, A. Yu. Sholokhova, and A. K. Buryak, "A deep convolutional neural network for the estimation of gas chromatographic retention indices," *Journal of Chromatography A*, vol. 1607, p. 460395, Dec. 2019, doi: 10.1016/j.chroma.2019.460395.

[17] S. E. Stein, V. I. Babushok, R. L. Brown, and P. J. Linstrom, "Estimation of Kováts Retention Indices Using Group

Contributions," *J. Chem. Inf. Model.*, vol. 47, no. 3, pp. 975–980, May 2007, doi: 10.1021/ci600548y.

[18] D. D. Matyushin, A. Yu. Sholokhova, and A. K. Buryak, "Gradient boosting for the prediction of gas chromatographic retention indices," *sorpchrom*, vol. 19, no. 6, pp. 630–635, Dec. 2019, doi: 10.17308/sorpchrom.2019.19/2223.

[19] Y. Marrero-Ponce, S. J. Barigye, M. E. Jorge-Rodríguez, and T. Tran-Thi-Thu, "QSRR prediction of gas chromatography retention indices of essential oil components," *Chem. Pap.*, vol. 72, no. 1, pp. 57–69, Jan. 2018, doi: 10.1007/s11696-017-0257-x.

[20] E. Dossin *et al.*, "Prediction Models of Retention Indices for Increased Confidence in Structural Elucidation during Complex Matrix Analysis: Application to Gas Chromatography Coupled with High-Resolution Mass Spectrometry," *Anal. Chem.*, vol. 88, no. 15, pp. 7539–7547, Aug. 2016, doi: 10.1021/acs.analchem.6b00868.

[21] V. V. Mihaleva, H. A. Verhoeven, R. C. H. de Vos, R. D. Hall, and R. C. H. J. van Ham, "Automated procedure for candidate compound selection in GC-MS metabolomics based on prediction of Kovats retention index," *Bioinformatics*, vol. 25, no. 6, pp. 787–794, Mar. 2009, doi: 10.1093/bioinformatics/btp056.

[22] T. Matsuo, H. Tsugawa, H. Miyagawa, and E. Fukusaki, "Integrated Strategy for Unknown EI–MS Identification Using Quality Control Calibration Curve, Multivariate Analysis, EI–MS Spectral Database, and Retention Index Prediction," *Anal. Chem.*, vol. 89, no. 12, pp. 6766–6773, Jun. 2017, doi: 10.1021/acs.analchem.7b01010.

[23] K. Héberger, "Quantitative structure–(chromatographic) retention relationships," *Journal of Chromatography A*, vol. 1158, no. 1–2, pp. 273–305, Jul. 2007, doi: 10.1016/j.chroma.2007.03.108.

[24] A. K. Zhokhov, A. Yu. Loskutov, and I. V. Rybal'chenko, "Methodological Approaches to the Calculation and Prediction of Retention Indices in Capillary Gas Chromatography," *J Anal Chem*, vol. 73, no. 3, pp. 207–220, Mar. 2018, doi: 10.1134/S1061934818030127.

[25] Z. Garkani-Nejad *et al.*, "Prediction of gas chromatographic retention indices of a diverse set of toxicologically relevant compounds," *Journal of Chromatography A*, vol. 1028, no. 2, pp. 287–295, Mar. 2004, doi: 10.1016/j.chroma.2003.12.003.

[26] B. Hemmateenejad, K. Javadnia, and M. Elyasi, "Quantitative structure–retention relationship for the Kovats retention indices of a large set of terpenes: A combined data splitting-feature selection strategy," *Analytica Chimica Acta*, vol. 592, no. 1, pp. 72–81, May 2007, doi: 10.1016/j.aca.2007.04.009.

[27] Z. Garkani-Nejad, "Use of Self-Training Artificial Neural Networks in a QSRR Study of a Diverse Set of Organic Compounds," *Chroma*, vol. 70, no. 5–6, pp. 869–874, Sep. 2009, doi: 10.1365/s10337-009-1241-6.

[28] J. Yan *et al.*, "Comparison of quantitative structure–retention relationship models on four stationary phases with different polarity for a diverse set of flavor compounds," *Journal of Chromatography A*, vol. 1223, pp. 118–125, Feb. 2012, doi: 10.1016/j.chroma.2011.12.020.

[29] C. Rojas, P. R. Duchowicz, P. Tripaldi, and R. P. Diez, "QSPR analysis for the retention index of flavors and fragrances on a OV-101 column," *Chemometrics and Intelligent Laboratory Systems*, vol. 140, pp. 126–132, Jan. 2015, doi: 10.1016/j.chemolab.2014.09.020.

[30] C. Rojas, P. R. Duchowicz, P. Tripaldi, and R. Pis Diez, "Quantitative structure–property relationship analysis for the retention index of fragrance-like compounds on a polar stationary phase," *Journal of Chromatography A*, vol. 1422, pp. 277–288, Nov. 2015, doi: 10.1016/j.chroma.2015.10.028.

[31] A. Mauri, V. Consonni, M. Pavan, and R. Todeschini, "Dragon software: An easy approach to molecular descriptor calculations," *Match*, vol. 56, no. 2, pp. 237–248, 2006.

[32] V. H. Masand and V. Rastija, "PyDescriptor

Anticancer Compound Sensitivity Prediction via Multimodal Attention-Based Convolutional Encoders," Mol. Pharmaceutics, vol. 16, no. 12, pp. 4797–4806, Dec. 2019, doi: 10.1021/acs.molpharmaceut.9b00520.

[53] G. B. Goh, K. Sakloth, C. Siegel, A. Vishnu, and J. Pfaendtner, "Multimodal Deep Neural Networks using Both Engineered and Learned Representations for Biodegradability Prediction," arXiv:1808.04456 [cs, stat], Sep. 2018, Accessed: Sep. 29, 2020. [Online]. Available: http://arxiv.org/abs/1808.04456.

[54] X. Fei, Q. Zhang, and Q. Ling, "Vehicle Exhaust Concentration Estimation Based on an Improved Stacking Model," *IEEE Access*, vol. 7, pp. 179454–179463, 2019, doi: 10.1109/ACCESS.2019.2958703.

[55] S. Guo, H. He, and X. Huang, "A Multi-Stage Self-Adaptive Classifier Ensemble Model With Application in Credit Scoring," *IEEE Access*, vol. 7, pp. 78549–78559, 2019, doi: 10.1109/ACCESS.2019.2922676.

[56] N. T. Cockroft, X. Cheng, and J. R. Fuchs, "STarFish: A Stacked Ensemble Target Fishing Approach and its Application to Natural Products," J Chem Inf Model, vol. 59, no. 11, pp. 4906–4920, 25 2019, doi: 10.1021/acs.jcim.9b00489.

[57] Md. K. Hasan, Md. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers," IEEE Access, vol. 8, pp. 76516–76531, 2020, doi: 10.1109/ACCESS.2020.2989857.

[58] Y. Xiong, Q. Wang, J. Yang, X. Zhu, and D.-Q. Wei, "PredT4SE-Stack: Prediction of Bacterial Type IV Secreted Effectors From Protein Sequences Using a Stacked Ensemble Method," Frontiers in Microbiology, vol. 9, p. 2571, 2018, doi: 10.3389/fmicb.2018.02571.

[59] R. Bouwmeester, L. Martens, and S. Degroeve, "Comprehensive and Empirical Evaluation of Machine Learning Algorithms for Small Molecule LC Retention Time Prediction," Anal. Chem., vol. 91, no. 5, pp. 3694–3703, Mar. 2019, doi: 10.1021/acs.analchem.8b05820.

[60] R. Bouwmeester, L. Martens, and S. Degroeve, "Generalized Calibration Across Liquid Chromatography Setups for Generic Prediction of Small-Molecule Retention Times," Anal. Chem., vol. 92, no. 9, pp. 6571–6578, May 2020, doi: 10.1021/acs.analchem.0c00233.

[61] H.-F. Chen, "Quantitative predictions of gas chromatography retention indexes with support vector machines, radial basis neural networks and multiple linear regression," *Analytica Chimica Acta*, vol. 609, no. 1, pp. 24–36, Feb. 2008, doi: 10.1016/j.aca.2008.01.003.

[62] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. Willighagen, "The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics," *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 2, pp. 493–500, Mar. 2003, doi: 10.1021/ci025584y.

[63] W. Jennings and T. Shibamoto, *Qualitative analysis of flavor and fragrance volatiles by glass capillary gas chromatography*. San Francisco: Academic Press, 1980.

[64] A. M. Veselinović, D. Velimorović, B. Kaličanin, A. Toropova, A. Toropov, and J. Veselinović, "Prediction of gas chromatographic retention indices based on Monte Carlo method," *Talanta*, vol. 168, pp. 257–262, Jun. 2017, doi: 10.1016/j.talanta.2017.03.024.

[65] J. A. Rijks and C. A. Cramers, "High precision capillary gas chromatography of hydrocarbons," *Chromatographia*, vol. 7, no. 3, pp. 99–106, Mar. 1974, doi: 10.1007/BF02269819.

[66] Sadtler Research Laboratories, Ed., *The Sadtler standard gas chromatography retention index library*. Philadelphia, Pa: Sadtler Research Laboratories, 1985.

[67] A. Yan, R. Zhang, M. Liu, Z. Hu, M. A. Hooper, and Z. Zhao, "Large artificial neural networks applied to the prediction of retention indices of acyclic and cyclic alkanes, alkenes, alcohols, esters, ketones and ethers," *Computers & Chemistry*, vol. 22, no. 5, pp. 405–412, Sep. 1998, doi: 10.1016/S0097-8485(98)00001-1.

[68] S. E. Stein and R. L. Brown, "Estimation of normal boiling points from group contributions," *J. Chem. Inf. Model.*, vol. 34, no. 3, pp. 581–587, May 1994, doi: 10.1021/ci00019a016.

[69] *Eclipse Deeplearning4j Development Team. Deeplearning4j: Open-source distributed deep learning for the JVM, Apache Software Foundation License 2.0. http://deeplearning4j.org*

[70] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[71] G. B. Goh, N. O. Hodas, C. Siegel, and A. Vishnu, "SMILES2Vec: An Interpretable General-Purpose Deep Neural Network for Predicting Chemical Properties," arXiv:1712.02034 [cs, stat], Mar. 2018, Accessed: Sep. 29, 2020. [Online]. Available: http://arxiv.org/abs/1712.02034.

[72] K. Preuer, G. Klambauer, F. Rippmann, S. Hochreiter, and T. Unterthiner, "Interpretable Deep Learning in Drug Discovery," in Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, vol. 11700, W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, Eds. Cham: Springer International Publishing, 2019, pp. 331–345.

[73] X. Domingo-Almenara et al., "The METLIN small molecule dataset for machine learning-based retention time prediction," Nat Commun, vol. 10, no. 1, p. 5811, Dec. 2019, doi: 10.1038/s41467-019-13680-7.

DMITRIY D. MATYUSHIN received a graduate degree and has been working as a staff scientist at the laboratory of physicochemical principles of chromatography and chromatography – mass spectrometry, Institute of Physical Chemistry and Electrochemistry, Moscow, since 2016. His scientific interests include chemoinformatics and computational chemistry and their applications to analytical chemistry, in particular to environmental and metabolomics analysis.

Now he is working on problems of untargeted analysis and identification of small molecules in complex mixtures. His most recent works are devoted to application of deep learning methods to search in databases containing gas chromatographic – mass spectrometric data.

ALEKSEY K. BURYAK has been working at the Institute of Physical Chemistry and Electrochemistry since 1986 after graduating from the Department of Chemistry of Moscow State University. PhD since 1986, Doctor of Sciences since 2000; he became the director of the Institute in 2016.

His interests are broad and include many fields of analytical chemistry and physical chemistry: mass spectrometry and physical basis of ionization and fragmentation in mass spectrometry, surface chemistry and physics of chromatographic separation, information processing in analytical chemistry. His early works are devoted to identification of isomers using predicted retention values.

Prof. Buryak has been a Corresponding Member of the Russian Academy of Sciences since 2019, a member of the editorial board of several scientific journals.