Quantitative structure-retention relationships for pyridinium-based ionic liquids used as gas chromatographic stationary phases: convenient software and assessment of reliability of the results

Anastasia Yu. Sholokhova[a], *Dmitriy D. Matyushin[a], Mikhail V. Shashkov[b]

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

[b]Boreskov Institute of Catalysis, 5 Lavrentieva Prospect, Novosibirsk, 630090, Russia

E-mail: *dm.matiushin@mail.ru

Highlights

- The predicted retention index for polyethylene glycol was used as a molecular descriptor
- Reliability and reproducibility of QSRR studies were discussed
- Three pyridinium-based ionic liquids were considered as GC stationary phases
- Retention index data sets for further QSRR studies were created and published
- CHERESHNYA is software for QSRR studies in GC

Abstract

Ionic liquids, i.e., organic salts with a low melting point, can be used as gas chromatographic liquid stationary phases. These stationary phases have some advantages such as peculiar selectivity, high polarity, and thermostability. Many previous works are devoted to such stationary phases. However, there are still no large enough retention data sets of structurally diverse compounds for them. Consequently, there are very few works devoted to quantitative structure-retention relationships (QSRR) for ionic liquid-based stationary phases. This work is aimed to close this gap. Three ionic liquids with substituted pyridinium cations are considered. We provide large enough data sets (123 - 158 compounds) that can be used in further works devoted to QSRR and related methods. We provide a QSRR study using this data set and demonstrate the following. The retention index for a polyethylene glycol stationary phase (denoted as RI_{PEG}), predicted using another model, can be used as a molecular descriptor. The use of this descriptor significantly improves the accuracy of the QSRR model. Both deep learning-based and linear models were considered for RI_{PEG} prediction. The ability to predict the

retention indices for ionic liquid-based stationary phases with high accuracy is demonstrated. Particular attention is paid to the reproducibility and reliability of the QSRR study. It was demonstrated that adding/removing several compounds, small perturbations of the data set can considerably affect the results such as descriptor importance and model accuracy. These facts have to be considered in order to avoid misleading conclusions. For the QSRR research, we developed a software tool with a graphical user interface, which we called CHERESHNYA. It is intended to select molecular descriptors and construct linear equations connecting molecular descriptors with gas chromatographic retention indices for any stationary phase. The software allows the user to generate several hundred molecular descriptors (one-dimensional and two-dimensional). Among them, predicted retention indices for popular stationary phases such as polydimethylsiloxane and polyethylene glycol are used as molecular descriptors. Various methods for selecting (and assessing the importance of) molecular descriptors have been implemented, in particular the Boruta algorithm, partial least squares, genetic algorithms, L1-regularized regression (LASSO) and others. The software is free, open-source and available online.

Keywords Gas chromatography, quantitative structure-retention relationships, molecular descriptors, stationary phases, pyridinium-based ionic liquids.

1. Introduction

Ionic liquids (IL), i.e., organic salts with a melting point below or about room temperature (< 100 °C), have been widely used in analytical chemistry in last decades [1-2]. IL are stable, non-volatile, and liquid in a wide temperature range. Some IL form stable thin films. This makes it possible [2-5] to use them as liquid stationary phases (SP) for gas chromatography (GC). In this case, IL demonstrate high polarity simultaneously with excellent thermal stability [3]. IL are widely used for the separation of various mixtures [5-8]. The selectivity and retention behavior of various IL were reviewed by Yao et al. [4]. Various IL are used as gas chromatographic SP: for instance, derivatives of imidazolium, phosphonium, pyridinium, and guanidinium can be employed [9, 10]. The structures of various IL-based SP are reviewed in Ref. [4, 9]. Several types of IL-based GC columns are commercially distributed by Supelco (owned by Merck Group). These columns are used for various separations [9, 11].

For the use in gas chromatography – mass spectrometry (GC-MS), SP should be particularly thermostable and non-volatile in order to provide low background noise. For less volatile, heavy, and

polar analytes, the SP have to be stable at higher temperatures. In Ref. [12], it was demonstrated that some imidazolium-based IL can be used for GC-MS at temperatures up to 300 °C and have background noise considerably lower than polyethylene glycol-based SP (PEG) and comparable with the non-polar HP-5ms SP.

Methods that predict chromatographic retention using the analyte structure as an input are usually referred to as quantitative structure-retention relationships (QSRR) [13]. One of the application areas of this method is the non-target GC-MS analysis using a mass spectral library search [14-15] for rejection of false candidates. QSRR can be considered as a method that provides an insight into chromatographic separation [16]. When predicting a retention index (RI) based on some molecular descriptors (i.e., numerical values that characterize the structure of a molecule), the contribution of particular molecular descriptors (MD) and a set of selected MD can provide valuable information about the nature of separation, and the model is considered as an interpretable one [16-20]. Almost all work on QSRR for GC is limited to the most typical and well-characterized polymeric SP. In liquid chromatography conditions, more factors influence retention and the use of QSRR to study the separation mechanism is even more common [21-23]. QSSR are also used as a convenient task in order to develop and demonstrate chemometric, statistical, and machine learning methods.

Many hundreds of MD are available by the means of commercial and open-source software [24]. Various types of MD and their use in QSRR in GC-MS are reviewed in Ref. [25]. Diverse machine learning methods (such as support vector machines [20, 25-26], gradient boosting [27], neural networks [20, 25-26]) are used for QSRR. But the most often used are the linear regression methods [25]. Various feature selection approaches can be used in quantitative structure-property research (in particular in QSRR) [24, 28]. Feature selection is especially important when an interpretable model with chemical meaning is required.

Despite the existence of a large number of QSRR studies, most of them use small data sets (less than 1000 compounds) and usually do not answer whether the obtained results will be reproducible if the data set is slightly changed. For example, in Ref. [17], the authors make some qualitative conclusions about retention based on a set of MD chosen using sequential selection. The authors do not study whether the MD selection procedure is reproducible and whether the same MD set will be chosen if the data set is slightly distorted. If a method is unstable to insignificant changes in the data set and random factors, it may lead to misleading conclusions.

Any QSRR study requires a large enough data set of retention values (retention time (RT) or RI) of diverse compounds, and the diversity of data sets affects the results [17]. To the best of our knowledge, such data sets are not available for IL-based SP. For each of SP, the data about the retention are available for a very small number of compounds. Usually these are data about test mixtures for determination of polarity or solvent parameters, or data about several very similar compounds. To the best of our knowledge, there are very few works about the RI prediction and QSRR for IL-based SP, and all of them are focused on one specific class of chemical compounds. In Ref. [29], QSRR for polychlorinated biphenyls and IL-based SP are considered. There are also some works [30-31] that predict the chromatographic properties of IL based on their structure, rather than predict the retention for a given IL based on the structure of the analyte. We focus on the latter task: to predict the retention of diverse compounds on a given IL-based SP.

The majority of previous works devoted to RI prediction consider polydimethylsiloxane, 5%-phenyl-methylpolysiloxane or PEG. For these SP, very large data sets are available. This fact allows for the development of accurate and versatile prediction models [26] and then use the predicted (for these common SP) RI as MD in models developed for other SP. In this work, we investigate whether the predicted RI for PEG is applicable as MD for prediction of RI for IL-based SP.

Since there are still no large and diverse enough data sets and QSRR studies for such SP, this work is aimed to fill this gap by constructing a moderately large structurally diverse retention data set of compounds of various classes for IL-based SP and providing the QSRR study using this data set. Experimental RT and RI were acquired for three promising monocationic and dicationic IL-based SP containing polysubstituted pyridinium cations. This work is also aimed to pay special attention to reliability and reproducibility of the QSRR study. We tested whether small distortions of data sets, such as randomly removing several compounds or adding minor noise to the values, could affect the conclusions of the QSRR study.

2. Materials and Methods

2.1. Chemicals

A collection of 181 organic compounds of diverse chemical nature was used: aromatic and aliphatic alcohols, aldehydes, ketones, heterocycles, and various halogenated compounds. A full list of compounds is provided in Supplementary Material, section S1, and all experimental RT and RI are provided in the online repository https://doi.org/10.6084/m9.figshare.16885009. Most of the

compounds were purchased from Sigma-Aldrich and several from other vendors. The purity of each compound and the correctness of the structure were checked by GC-MS (electron ionization) using matching of observed spectra with spectra from a mass spectral database and matching of RI on standard polar and non-polar SP with reference ones (when available). The NIST 17 database was used for this purpose. A standard mixture of n-alkanes C_7-C_{40} (1000 µg/ml of each component in hexane, Sigma-Aldrich) was used for determination of n-alkanes RI. Acetonitrile (UHPLC-Supergradient PAI-ACS, Panreac) was used to dissolve standard compounds.

2.2. Analysis conditions

1 µl of liquid analytes was dissolved in 0.9 ml of acetonitrile. 1.5 mg of solid analytes was dissolved in 1 ml of acetonitrile. Analyses were carried out using Shimadzu GCMS-TQ8040 (Shimadzu). We mixed up to 10 compounds in one solution (partial concentrations are given above), and in those cases where the peak annotation was not absolutely unambiguous, we remeasured solutions of individual compounds. We measured all compounds using three columns with IL (see below), as well as HP-5 (30 m, 0.32 mm×0.25 µm, Agilent) and SH-Stabilwax (30 m, 0.25 mm×0.1 µm, Shimadzu) columns. The numbers in brackets denote the length, inner diameter of the column, and thickness of the SP layer, respectively. Measurements were made for standard polar and non-polar SP in order to obtain spectra for comparison, as well as to verify that the observed RI match the reference ones. 0.5 µl of the liquid solution was injected to the GC-MS instrument; in order to measure n-alkane RI, a mixture of n-alkanes was added to the sample solution.

GC-MS analyses were carried out under the following conditions. Temperatures of injector and ion source: 250 °C and 200 °C, respectively; carrier gas: He; flow control mode: constant linear velocity; flow rate: 0.6 ml/min; injection split ratio: 1:50. Oven temperatures were programmed as follows: the temperature was raised from 50 °C to 240 °C at 8 °C/min rate and then was kept constant during 15 min. The mass spectrometer was operated in electron ionization (EI) mode at 70 eV, scan rate: 1666 units/s, mass range: 44–500 m/z.

2.3. Capillary columns coated with ionic liquids

Three IL-based GC columns were used: Bis4MPyC6 (30 m, 0.22 mm×0.2 µm), Bis2MPyC9 (25 m, 0.22 mm×0.2 µm), Hex4MPy (18 m, 0.22 mm×0.2 µm). The structures of IL used in these columns are shown in Fig. 1. IL were prepared according to the procedure from Ref. [32]. Cations (in the form

of bromide) were prepared by heating a mixture of corresponding methylpyridine and bromo- or dibromoalkane at 120 °C during 2-6 hours. IL were prepared by the reaction of previously produced bromide with lithium bis(trifluoromethanesulfonyl)imide. The columns were prepared by the static high pressure technique [33] at a constant temperature of 210 °C using tert-butanol as a solvent. The column preparation procedure is described in Ref. [34].

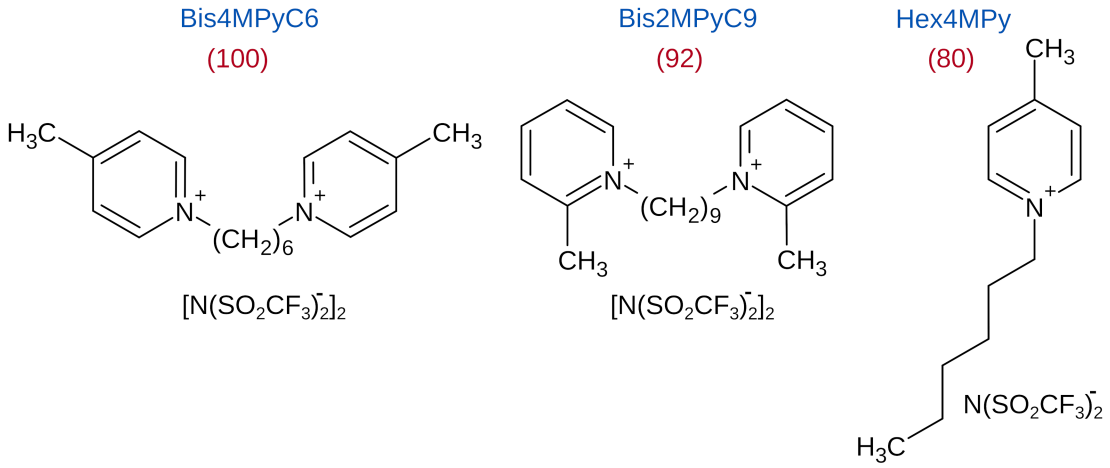

Fig. 1. Structures of the considered IL used as SP. Numbers denote the McReynolds polarity values of SP.

2.4. QSRR modeling and retention index prediction

2.4.1. Prediction of retention indices for PEG

The predicted based on the molecule structure RI for PEG was used as the MD for further prediction of RI for IL-based SP (see Fig. 2A). So, this is a supplementary task for this work. We used two methods for prediction of RI for PEG. The first one is the use of a quite accurate deep learning model, previously described in our previous works [26]. In this case, a multimodal ensemble of two deep neural networks was used. The neural networks were trained using the NIST 17 database. The models use SMILES string representations of models, various MD, and molecular fingerprints used as an input representation of molecules. The models were described in the previous work [26] and used in the unchanged form. The newly developed and described in this work CHERESHNYA software calls our previous software [35, 36] for prediction of RI for the DB-WAX column. This predicted RI value is further referred as the RI_PEG_DL descriptor.

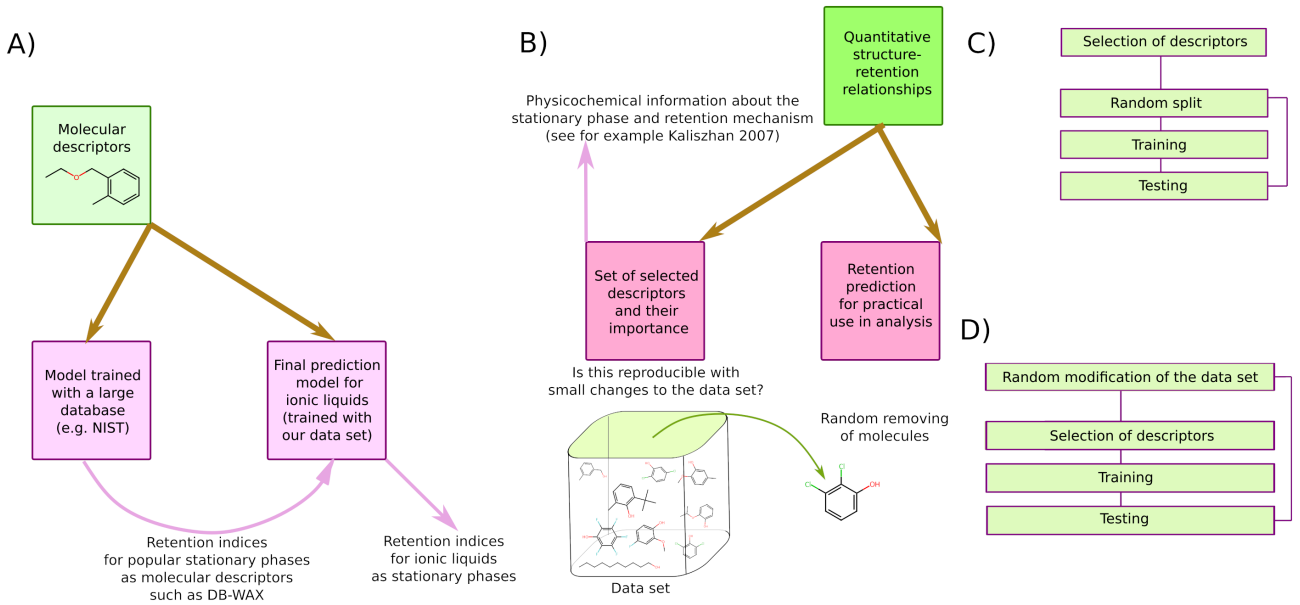

Fig. 2. Graphic illustration of the topics investigated in this work: A – use of predicted for polymeric SP RI as MD for prediction of RI for IL-based SP; B – reproducibility of MD selection and its importance for the use of QSRR for the SP description; C, D – comparison of common cross-validation with the approach used in this work.

The second approach is a linear model for prediction of RI on PEG. The following set of features was used: 243 2D MD and 84 functional groups counters generated using the Chemistry Development Kit, version 2.7.1 (CDK) [37]; 208 2D MD of various types and 42 MQN (so-called Molecular Quantum Numbers [38]) generated using the RDKit library, version 2023.09.4; 4860 Klekota-Roth substructure counters (Klekota-Roth counting fingerprint [39]). The first subset of features (functional groups counters and CDK descriptors) was the same as used in our previous work [26]. All MD were scaled to the range [0; 1]:

$$D_{new} = (D - D_{min}) / (D_{max} - D_{min}) \quad ,$$

where D_{new}, D, D_{min}, D_{max} – scaled, unscaled, minimal and maximal values of a MD.

The NIST 2017 library was used as the training set. Preprocessing of the library is described in our previous work [26], and unsupported compounds were excluded as described there. For each compound, the median value of all values for PEG was used. The compounds that were also measured on IL-based SP were excluded from this data set. Features with zero variation (constant for all molecules), features that are linearly dependent on other features, and features that are not supported for some molecules were excluded. As a result, a data set containing 9408 compounds (1698 features

for each compound) was constructed. This data set was randomly split into training (80%), validation (10%), and test (10%) data sets. The validation set was used for hyperparameters tuning.

The linear model was constructed using support vector regression with a linear kernel. The LibLinear library (version 2.43) was used with the following hyperparameters: C = 0.086; p = 1.44; solver type: L2-regularized L1-loss SVR (dual problem). The final model and additional information are provided in Supplementary material, section S2. The predicted value is further referred as the RI_PEG_LM descriptor. The mean and median square absolute errors for this model were 53.3 and 31.3, respectively, it is comparable with the neural network-based models from Ref. [26].

2.4.2. Molecular descriptor selection for QSRR modeling

A total of 208 MD generated with the RDKit library were used, MQN descriptors were not included. Since the data sets in this case are small and our purpose is a simple and interpretable model, we limited ourselves to the RDKit descriptors, as well as RI_PEG_LM and RI_PEG_DL. We considered 6 methods for selection of MD for QSRR models. The overview and designations of these methods are given in Table 1.

Table 1. Methods of MD selection considered in this work.

Designation	Method of molecular descriptor selection
SEQ_ADD	Sequential addition
LASSO	L1-regularized linear regression with the value of l_1 constant equal to 1.0
BORUTA	Boruta algorithm based on random forest (500 trees), 80 rounds of Boruta algorithm
GA	Genetic algorithm (80 generations)
PLS_VIP	Partial least squares (20 components) with variable importance in projection
SEQ_REM	Sequential removal of molecular descriptors using random forest for assigning the importance scores

In the SEQ_ADD method, at the first stage, the most correlated with the target RI values MD is selected. Then for each of MD that have not yet been selected, the ordinary least squares (OLS) model is built and the MD for which the *f*-factor (goodness of fit) is the largest is selected. In this method, the *f*-factor was calculated using the following equation:

$$F = \frac{(TSS - RSS) * (N_{mol} - N_{desc})}{RSS * (N_{desc} - 1)} \quad ,$$

where TSS – total sum of squares, RSS – residual sum of squares, N_{mol} and N_{desc} are numbers of molecules and MD, respectively.

In the LASSO method, the following term is added to the sum of the squares of deviations:

$$L = l_2 * \sum_i |a_i| \quad ,$$

where a_i – coefficients of the linear model. Such loss function forces some of MD to be almost zero, and MD with coefficients more than 0.1 are selected (all MD were scaled to the range [0; 1] when this threshold value is applied). We use the implementation of L1-regularized regression from the Smile package [40] (version 2.6.0).

The Boruta algorithm is based on other algorithms that can provide importance scores of features. In addition to real features, the same number of "fake" features is added. Fake features are made from real ones by random shuffle of rows. A feature is considered "important" if its importance score is better than the best "fake" feature. The final importance is the number of repeats in which this feature was considered as important. Feature importance scores provided by random forest are based on the decrease of the impurity measure when the corresponding variable is used. For the Boruta algorithm [41], we use our own implementation of the algorithm. We use implementations of random forest from the Smile package [40] (version 2.6.0) with default hyperparameters for initial importance score of features. 80 rounds of the Boruta algorithm were used.

For the GA method, we use the implementation of the genetic algorithm from the Scikit-learn python package (sklearn-genetic, version 0.6.0). The GeneticSelectionCV function with `cv` = 5 is used, OLS linear regression is used as a regression estimator, and the coefficient of determination (R^2) is used as an error measure. For the PLS_VIP method [42], we use the implementation of PLS from the Scikit-learn python package [43] (version 1.4.0). In sequential removing (SEQ_REM), at each stage of the algorithm, the random forest model is built and 10 MD with the least values of importance are removed until the required number of MD are remained. In the last step, less than 10 MD are removed if the total number of MD to be removed cannot be divided by 10.

For SEQ_REM, we also consider MD preselection when MD with a Pearson correlation coefficient $r > 0.8$ are removed. Before training the models, all MD were scaled to the range [0; 1] and the RI values were divided by 1000. This is necessary to avoid too large coefficients and incorrect operation of the framework. The coefficient values given in this work are given without taking into account scaling (for actual values), unless otherwise indicated. The above-mentioned Smile package is used for OLS and the building of final linear equations.

In order to characterize the accuracy, the mean absolute error (MAE), median absolute error (MdAE), root mean square error (RMSE) were computed, 10-fold cross-validation is used, unless otherwise specified. For the comparison purpose, the "black box" models created using previously

published software [26] are considered. A detailed description of this software and machine learning models is given in our previous work [26]. To evaluate the accuracy of the models, we also used 10-fold cross-validation (the “CV” command line option of the above-mentioned software).

2.5. Evaluation of the molecular descriptor selection reproducibility

In previous works, QSRR were often used not for practical prediction of RI, but for characterization of SP in order to draw some conclusions about the nature of retention mechanisms [16] (see Fig. 2B). However, it is not usually tested whether the selection of MD is reproducible with small changes to the data set. If the conclusions characterize the SP in general, then these conclusions should not be altered if the data set (that plays the role of a probe) is slightly changed: for example, several molecules are removed. Thus, the procedure of MD selection was repeated many times and each time 1-25 molecules were removed from the data set. Each time, for most MD selection methods, the set of selected MD was different and conclusions were made taking this change into account. In most cases (unless otherwise specified), 200 repeats are performed and resulting average values of accuracy and importance scores are given.

Also, the accuracy of the model (and of the method in general) depends on the set of selected features. Typically, in QSRR works, a set of MD is selected only once [17-22] and then the accuracy of the model is carefully investigated (see Fig. 2C) using cross-validation or one-leave-out approaches. But since MD selection is a stochastic procedure, such a careful statistical evaluation of the accuracy is not very meaningful because it is built on the basis of a stochastic procedure that was made only once. In this work, we apply the modified procedure and after each alteration of the training set we repeat the MD selection (see Fig. 2D). This approach allows for the evaluation of accuracy of the approach in general, rather than the accuracy of one randomly selected MD set. The evaluation of the reproducibility is made for all three IL, but the detailed results are shown only for Bis4MPyC6, unless otherwise specified. The corresponding data set contains 123 compounds of various classes. All conclusions regarding the comparison of the MD selection methods in terms of reproducibility and accuracy are the same for any of the considered IL.

After removing a given number of molecules and before the MD selection procedure, a preliminary reduction of the MD set is done. The MD that were constant for all molecules or that coincided with other MD up to a linear dependence were removed. The resulting MD set (before the selection procedure) contained 110 - 120 MD (the exact number depends on the exact data set and changes with random removing of molecules).

3. Results and discussion

3.1. Data set for QSRR

Retention data (n-alkane RI and RT) were acquired for 178 compounds (at least for one of the columns with IL) for three columns with IL-based SP. These compounds include 108 aromatic and 70 aliphatic compounds. Among these molecules, 37 are ethers, 49 are phenols, 13 are aldehydes or ketones, and 129 molecules have a hydroxyl group attached to an aliphatic atom. All considered molecules contain carbon, hydrogen, and oxygen. Some of these compounds contain other elements: 26 contain fluorine, 35 contain chlorine, 13 contain bromine, 3 contain iodine, 5 contain nitrogen, and only one contains sulfur. In the final data set for each column, we included only compounds with an n-alkane RI of less than 3500.

During the acquisition of data for the final data sets, all data for each column were measured within 5 days using an autosampler. RT were extracted from chromatograms using GCMSsolution GCMS Postrun Analysis (version 4.50) software. The RT was recorded at the top of the peak. The fraction of compounds (randomly selected) was remeasured after ~15 days after the end of the first acquisition in order to estimate the reproducibility and measurement error. The mean deviations (the average of deviations for multiple compounds) between the results of the first and later measurements are 0.066, 0.026, and 0.030 min. for the Bis4MPyC6, Bis2MPyC9, and Hex4MPy columns, respectively. The mean percentage deviations are 0.53%, 0.44%, and 0.67% for these three columns, respectively. The Bis4MPyC6 column is the longest and the most polar, so the absolute values of the RT are the largest for this column. Due to this, the absolute mean deviation is the largest, while the relative deviation is not.

We also studied whether there was a significant dependence of the RT on the injected volume. No significant dependence was observed: for 5 compounds, the RT was the same for each of the compounds for an injected volume within the range 0.1 – 1 μl. The deviation for successive measurements was not more than 0.01 min. In addition, there is almost no difference whether there was one or multiple different compounds in the solution. Errors of the RI measurement are less than 10 units for almost all compounds. Compounds with RI more than 3500 were not included due to a possible high error in the RI estimation: in this area, peaks of n-alkanes tend to be broad and located closely to each other. The use of RI systems other than those based on n-alkanes can be the scope of further research. The data set containing RT and RI can be downloaded from the Figshare repository https://doi.org/10.6084/m9.figshare.16885009.

3.2. Reproducibility of a QSRR study when data set changes

3.2.1. Stepwise selection with ordinary least squares

The first method of the feature selection employed in this work was “greedy” stepwise selection (SEQ_ADD). MD were added one by one. At each step, the MD that allows achieving the most significant linear regression is selected. The MD selection procedure was repeated 200 times, 10 MD were selected every time. Every time, 25 randomly selected molecules were excluded from the initial data set. Almost every time the set of selected MD was different. If we compare a random pair of MD sets obtained in different runs, then on average ~5.3 out of 10 MD will be the same. In Fig. 3A, the probability to be selected is shown for different MD. Only one MD (fr_benzene, number of benzene rings) is selected in all 200 repeats. This value – the probability of being selected – allows comparing the importance of the MD in a reliable way, while the selection of the MD only once does not allow making any conclusions. The confidence intervals are shown in Fig. 3A for $p = 0.95$, $N = 200$.

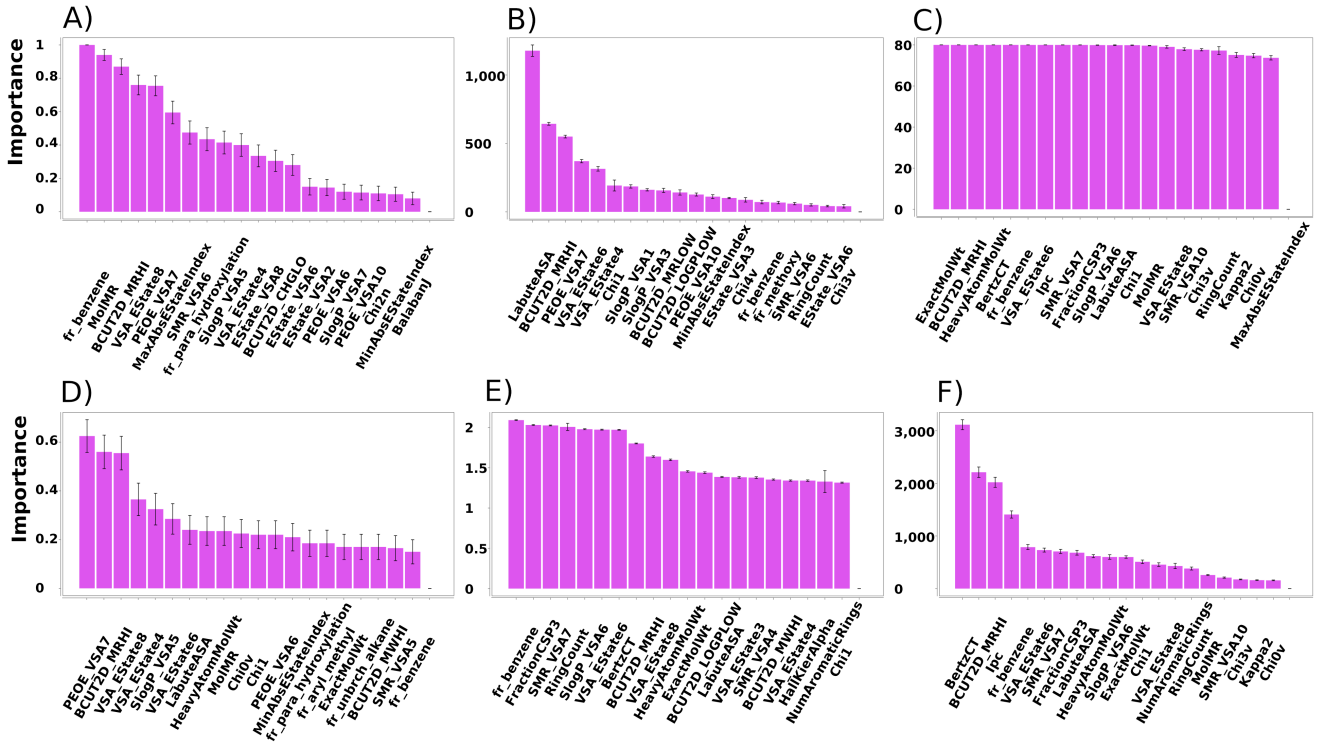

Fig. 3. Importance of various MD for prediction of RI for Bis4MPyC6 SP (all available compounds) estimated with various methods: A – SEQ_ADD; B – LASSO; C – BORUTA; D – GA; E – PLS_VIP; F – SEQ_REM. The ordinate axis denotes the average importance score for 200 repeats, and the error bars show the confidence interval ($p = 0.95$, $N = 200$); the exact meaning of importance score is different for various methods and is described in Section 3.2.1.

Table 2. Cross-validation accuracy of RI prediction for Bis4MPyC6 (all available compounds). Confidence intervals (p = 0.95, N = 200) are shown, the MD selection procedure was performed 200 times with exclusion of 25 random compounds from the data set.

Method	RMSE	MAE	MdAE
SEQ_ADD	140.2 ± 1.5	102.5 ± 1.3	72.9 ± 1.5
LASSO*	142.3 ± 2.0	102.9 ± 1.6	69.4 ± 1.7
BORUTA	230.4 ± 2.6	179.4 ± 2.1	147.8 ± 2.8
GA	147.5 ± 1.9	107.6 ± 1.5	76.5 ± 1.7
PLS_VIP	280.8 ± 4.1	224.7 ± 3.0	190.4 ± 2.9
SEQ_REM	211.1 ± 2.8	161.9 ± 2.7	126.8 ± 3.6

*For all methods except LASSO, the accuracy of OLS regression that uses selected MD is shown; for LASSO, the accuracy of LASSO regression itself is shown instead.

The prediction accuracy of this approach is demonstrated in Table 2. Confidence intervals of the accuracy measures are also shown. Standard deviations for various error measures are in the range of 9 – 11 units. Such relatively large values of standard deviation show that the comparison of accuracy of prediction methods must be done very carefully, the accuracy varies with random modification of the data set. However, in many works it was done based only on one cross-validation experiment [19, 26].

The more molecules we remove from the data set each time we train, the less reproducible the set of selected MD is. Such a dependence is shown in Fig. 4A. The dependence of the average number of the MD selected in both experiments for all pairs of experiments is shown. Even if we remove only one molecule, the MD selection will not be reproducible. In addition, a typical dependence of accuracy on the number of MD is shown in Fig. 4B. The average of 200 repeats is shown, confidence intervals are too narrow to be shown. It can be seen that the prediction error (as expected) decreases with increasing number of MD, but with a further increase in the number it decreases very slowly. Based on this, we decided to select 10 MD.

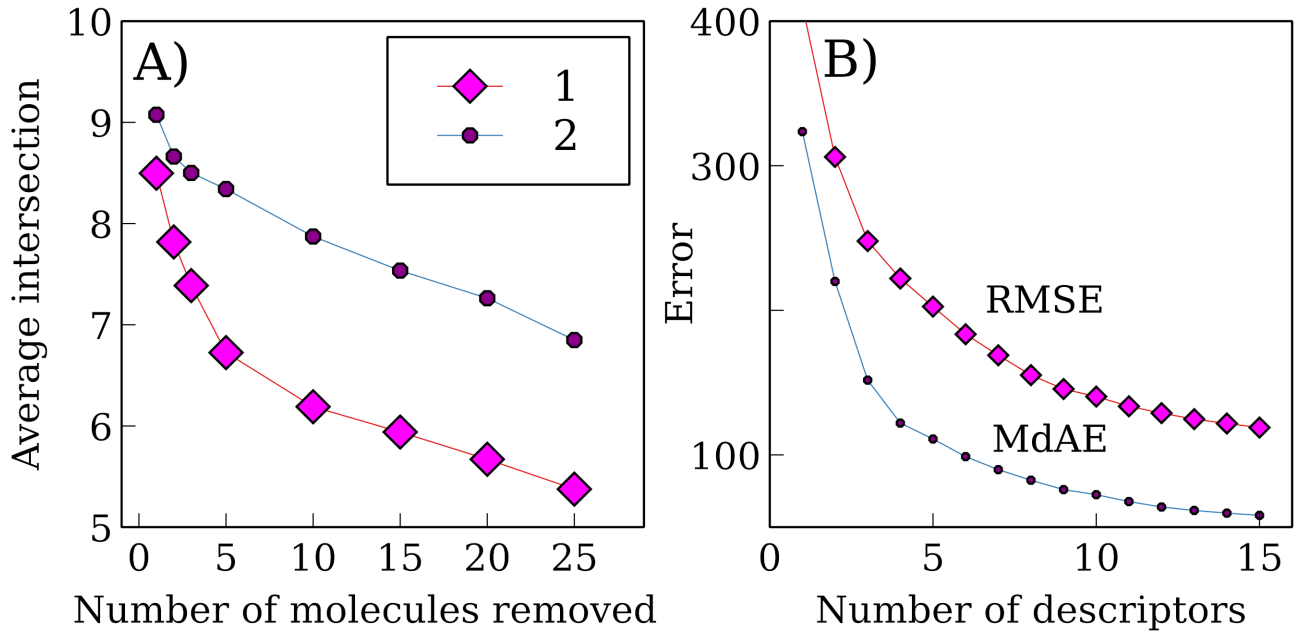

Fig. 4. A – dependence of the average size of the intersection between sets of selected MD for all pairs of repeats with an altered data set on the number of molecules randomly excluded from the data set during each repeat (1 – without preliminary removal of highly correlated MD; 2 – with preliminary removal of MD with the Pearson correlation coefficient $r > 0.8$ with any other MD); B – dependence of cross-validation accuracy on the number of MD.

One reason why MD selection may not be reproducible is that many MD are highly correlated. In Fig. 5, a heatmap given that shows the Pearson correlation coefficient r between some MD for this data set. It can be seen that MD, which at first glance are almost unrelated to each other, are often correlated. For example, the number of methoxy groups (fr_methoxy) and a topological MD characterizing the contribution of polar atoms to the total surface of the molecule (VSA_EState9) are strongly correlated. For each pair of correlated MD, we can arbitrarily remove one of the two. But if we do this, then our qualitative conclusions based on the set of MD may also change depending on which of them we remove. However, we considered such a reduction of the MD set. We removed from the original set all MD having $r > 0.8$ with any of the others. There were 49.9 ± 2.0 MD left (confidence interval, $p = 0.95$, $N = 200$). With this approach to preselection of the MD, we conducted the same experiments in order to evaluate the reproducibility.

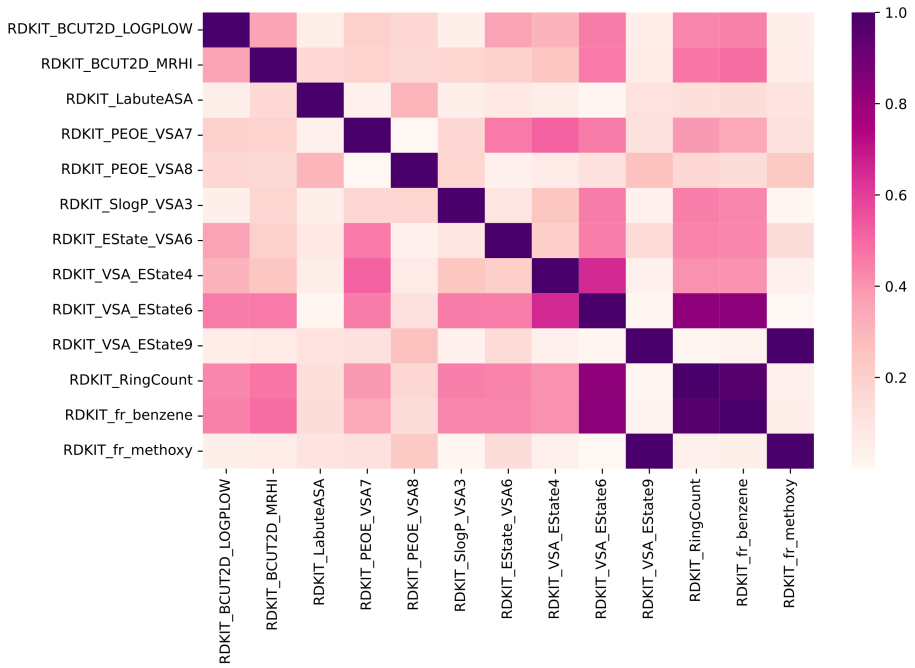

Fig. 5. Heatmap showing the Pearson correlation coefficient for pairs of MD for Bis4MPyC6 SP (all available compounds).

In Fig. 4A, the results of such an experiment are also shown. All MD having a Pearson correlation coefficient $r > 0.8$ with any of the remaining ones were removed from the set. A total of 200 repetitions of MD selection were made using a stepwise method. In this case, for each pair of repetitions, on average ~6.8 MD coincide instead of ~5.3 (25 randomly selected molecules are excluded each time). Reproducibility was slightly improved, as expected, but this approach includes a virtually random removal of half of the MD (from a pair of correlated ones, we randomly choose which one to remove). We do this in a reproducible way (for the same pair of correlated MD, the same MD is removed each time). However, the reproducibility is still not very good and it is clear that as the number of molecules removed from the data set increases, the reproducibility also decreases. Thus, the problem of the selected MD set not being reproducible across the data set changes cannot be explained solely by the presence of highly correlated MD.

Thus, we can draw the following conclusion. The stepwise algorithm for selecting MD for linear regression is not reproducible when small changes in the data set are made, and no “physicochemical” conclusions can be drawn from the set of once selected MD. Unfortunately, a number of previous works [17, 20] made such conclusions.

3.2.2. L1-regularized regression (LASSO)

The LASSO regression (L1-regularized linear regression) is an accurate linear regression and (simultaneously) a MD selection method. When a weighted sum of absolute values of coefficients is added to the loss function, the minimization of the loss function leads to zeroing of most of coefficients. We consider a MD to be selected if its coefficient (for scaled to the [0; 1] range value of the MD) is positive and above the threshold value of 0.1. In Fig. 6, the dependence of the accuracy and number of selected MD on l_1 constant is shown. Smaller values of l_1 result into better accuracy and larger number of important MD. At values $l_1 < 1.0$, the accuracy decreases very slowly with decrease of l_1, and $l_1 = 1.0$ was used for further investigation.

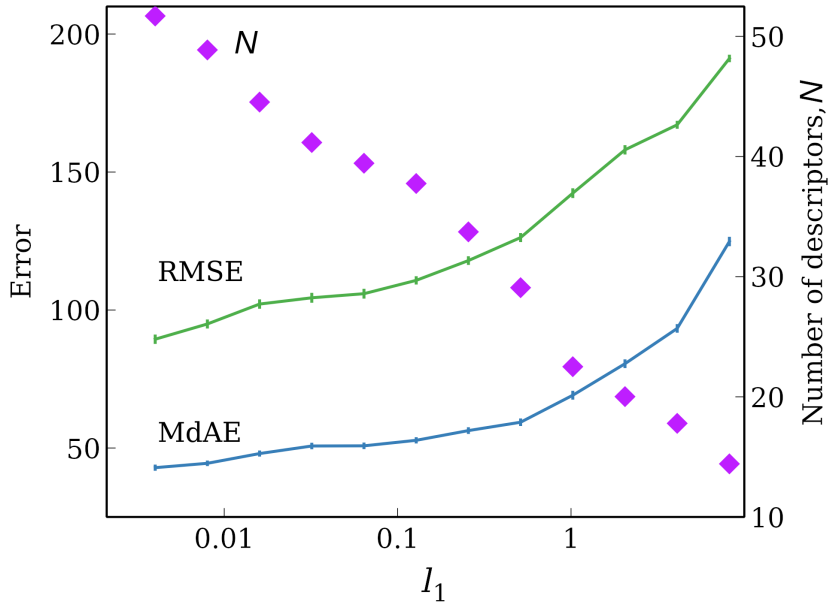

Fig. 6. Dependence of cross-validation accuracy and number of MD (N) with non-zero (>0.1) coefficient on l_1 value in L1-regularized regression (LASSO).

Unlike other MD selection algorithms, the values of accuracy given in Table 2 are given not for the OLS method with 10 MD, but for the LASSO method with $l_1 = 1.0$ itself. The use of MD selected by LASSO in OLS results in very poor accuracy as expected. It can be seen that the accuracy achieved with LASSO is about the same compared with the stepwise algorithm, but in this case much more MD are used (22.5 ± 1.3 on average).

In Fig. 3B, the average values of coefficients (for scaled to the [0; 1] range value of the MD) in LASSO regressions for various MD are shown. 200 repeats were performed, excluding 25 molecules

from the data set each time. It should be noted that the coefficient values are given for MD scaled to the range [0; 1] rather than for initial values. This allows performing a fair comparison of the importance.

It can be clearly seen that the following MD: LabuteASA, BCUT2D_MRHI are the most important (for Bis4MPyC6) and play the largest role. The 5 most important MD and their order can be established reliable. The average number of MD selected in both runs for all pairs of repeats is ~16.0. In general, the LASSO method provides linear regressions of similar accuracy as for the stepwise algorithm, and the selection and importance scores are somewhat more reproducible.

3.2.3. Other molecular descriptor selection algorithms

The next considered algorithm of the MD selection is the Boruta algorithm [42]. The reproducibility of the MD selection is quite high in this case. We performed 200 repeats with random exclusions of 25 molecules from the data set and each time we performed 80 rounds of the Boruta algorithm. For Bis4MPyC6, there are 11 MD (SMR_VSA7, LabuteASA, ExactMolWt, BCUT2D_MRHI, FractionCSP3, VSA_EState6, HeavyAtomMolWt, BertzCT, SlogP_VSA6, Ipc, fr_benzene) that are considered as important in all repeats and in all rounds of the Boruta algorithm. Other MD are sometimes considered as important and sometimes not. The results are shown in Fig. 3C. The average (over 200 repeats) number of rounds of the Boruta algorithm, in each of which the MD is considered as important, is shown. The confidence intervals (N = 200, p = 0.95) are shown. However, the accuracy of the OLS linear regression built using MD selected using the Boruta algorithm is very low (see Table 2), much worse than with step-by-step selection. Consequently, this algorithm is not suitable for constructing linear QSRR, although it allows evaluating the importance of the MD in reproducible way.

The genetic algorithm, as well as the stepwise algorithm, selects MD based on the accuracy of the OLS regression built on this set of MD, while the PLS-VIP and Boruta algorithms select according to criteria that have nothing to do with the accuracy of the OLS regression. Therefore, just as in the case of a step-by-step algorithm, one can expect that the accuracy of the OLS regression built on these MD will be quite high. Indeed, Table 2 shows that the genetic algorithm allows obtaining relatively accurate linear equations. If we compare random pairs of MD sets obtained in different runs, then on average only ~2.2 out of 10 MD are the same for GA. The accuracy of final linear equations for GA is close to that for SEQ_ADD, while the reproducibility of MD selection is significantly worse compared

with SEQ_ADD, LASSO and BORUTA methods at least with the used number of generations. Same as for the SEQ_ADD algorithm, the probability to be selected in each repeat is shown in Fig. 3D.

The last two considered algorithms were PLS_VIP and SEQ_REM. The average importance scores estimated using these methods are shown in Fig. 3E and Fig. 3F, respectively. In case of SEQ_REM, the importance score is estimated using a random forest method based on the decrease of the impurity measure when the corresponding variable is used. As well as the Boruta method, both these methods are reproducible in order to estimate importance of MD but cannot be used for the MD selection for OLS linear regression. The accuracy of PLS regression itself was not investigated in this work.

3.2.4. Comparison of sets of molecular descriptors selected by different algorithms

Different algorithms select different MD sets. We compared MD sets selected by BORUTA, SEQ_ADD and LASSO methods. In case of SEQ_ADD and LASSO we selected 10 most important MD, in case of BORUTA we selected 11 MD, since each of them is considered as important at all iterations of the Boruta algorithm. The Venn diagram for the resulting MD sets is shown in Fig. 7A. Only one MD (BCUT2D_MRHI) is considered as important in all cases for Bis4MPyC6. A total of 4 MD (BCUT2D_MRHI, fr_benzene, LabuteASA, VSA_EState6) are considered as important by at least 2 algorithms simultaneously. The fact that different methods select different sets of MD, and the results of each algorithm are not completely reproducible with minor changes in the data set, shows that in order to draw any “physicochemical” conclusions from such a study, it is necessary to carefully consider issues of reproducibility. Otherwise, one can draw conclusions based on a random (statistically insignificant) result.

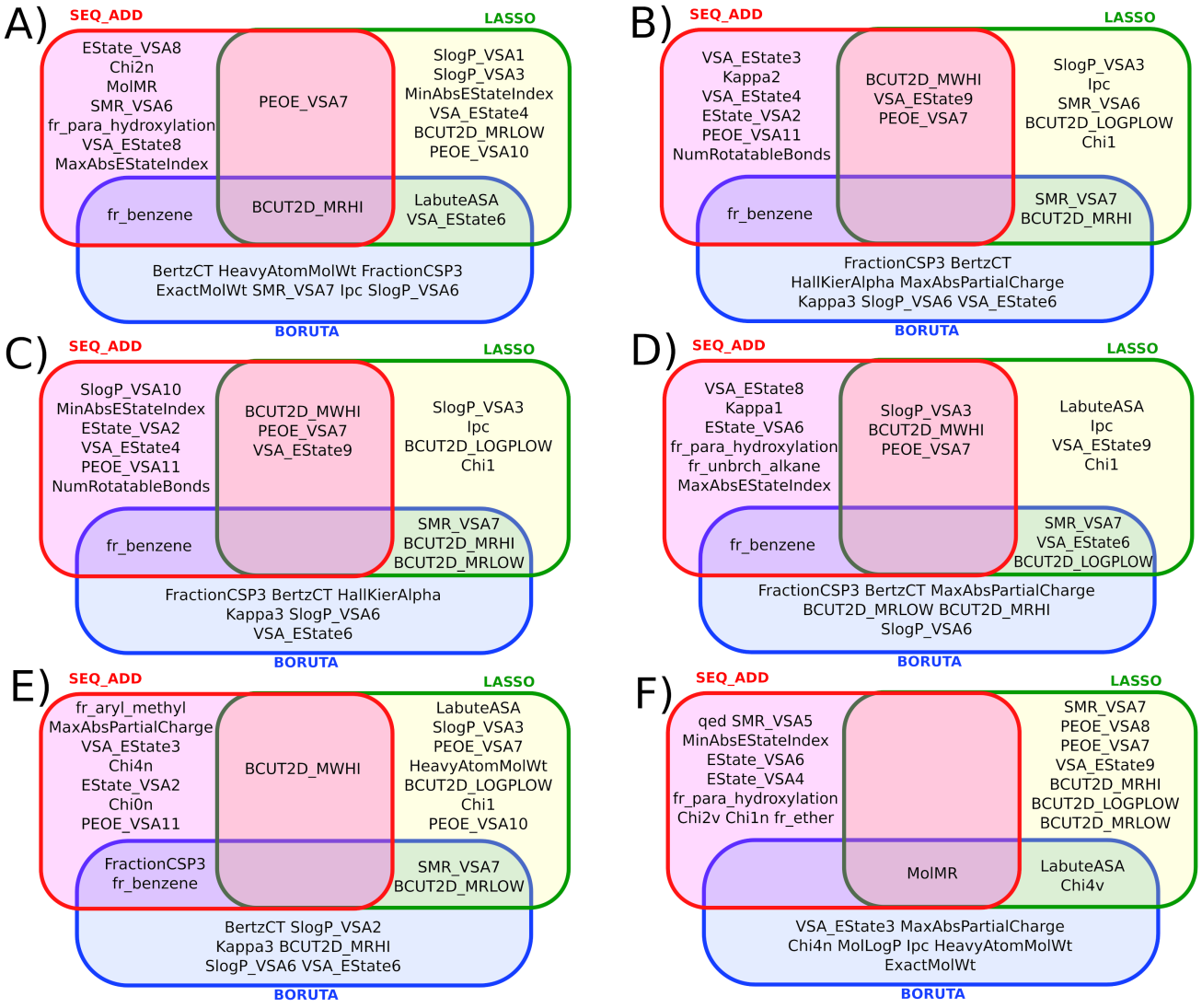

Fig. 7. Venn diagram of MD selected by various MD selection methods for various data sets and SP: A – Bis4MPyC6, including compounds for which RI are not available for other SP; B – Bis4MPyC6; C – Bis2MPyC9; D – Hex4MPy; E – polyethylene glycol; F – 5%-phenylpolydimethylsiloxane. All diagrams B-F are acquired for the identical set of compounds.

3.3. QSRR study for different stationary phases

The data sets acquired for three SP are slightly different, because some compounds are not eluted at reasonable temperatures on some SP, compounds with RI > 3500 are not included in the considered data set (low accuracy of the RI determination in these cases) and due to other reasons. The information about the overlap of these data sets is provided in Supplementary material, Fig. S3. In the previous sections, we demonstrated that even a small difference in a data set severely affects the set of selected MD and their importance. Therefore, the comparison of the sets of selected MD was

performed using the “intersection” data set. For each compound from this data set, RI is available for all three IL, as well as for 5%-phenyl-methylpolysiloxane (also denoted as DB-5 for conciseness, after the common column with such SP) and for polyethylene glycol (43 compounds). It should be noted that such incorrect comparison was made in previous works. For example, in Ref. [17] (Journal of Chromatography A), the authors built QSRR for 4 SP having significantly different data sets (different in dozens of compounds). The authors selected MD using a sequential algorithm and commented on the chemical nature of the separation based on the selected MD set.

In Fig. 8, importance values of MD determined using SEQ_ADD and LASSO methods for five SP are shown. In order to create these plots, 200 repeats were performed with one randomly excluded molecule. The number of excluded molecules was decreased because a much smaller data set is used. Three IL (Bis4MPyC6, Bis2MPyC9, Hex4MPy) and two polymeric SP: polyethylene glycol and 5%-phenyl-methylpolysiloxane were considered. MD selected for Bis4MPyC6 and Bis2MPyC9 are very similar to each other. This is consistent with the fact that these IL are very close in their chemical nature. However, Bis4MPyC6 is more polar, consequently the RI values are higher (the data set consisted of polar molecules) and the absolute values of the coefficients in the LASSO regression before MD are higher. Thus, the PEOE_VSA7 descriptor characterizes the accessible surface of atoms which Gasteiger charge is in the range [-0.05; 0]. Such charges typically have aromatic carbon and other atoms in moderately polar groups, while aliphatic carbons are hidden by positive-charged hydrogens. The BCUT2D_LOGPLOW is the lowest eigenvalue of a matrix which diagonal elements contain contributions of atoms to LogP (factor of lipophilicity) and non-diagonal elements contain information about the connectivity between the corresponding atoms. Both MD are topological and related to the polarity of the molecule, and the average coefficients before them increase with the polarity of the molecule. It should be noted that the most influential according to different MD selection methods fr_benzene (the number of benzene rings) and PEOE_VSA7 are not strongly correlated: the Pearson correlation coefficient is ~0.5 for the considered data set. The topological chi1 descriptor [44] is higher for linear molecules and lower for branched ones and characterizes the shape of the molecule.

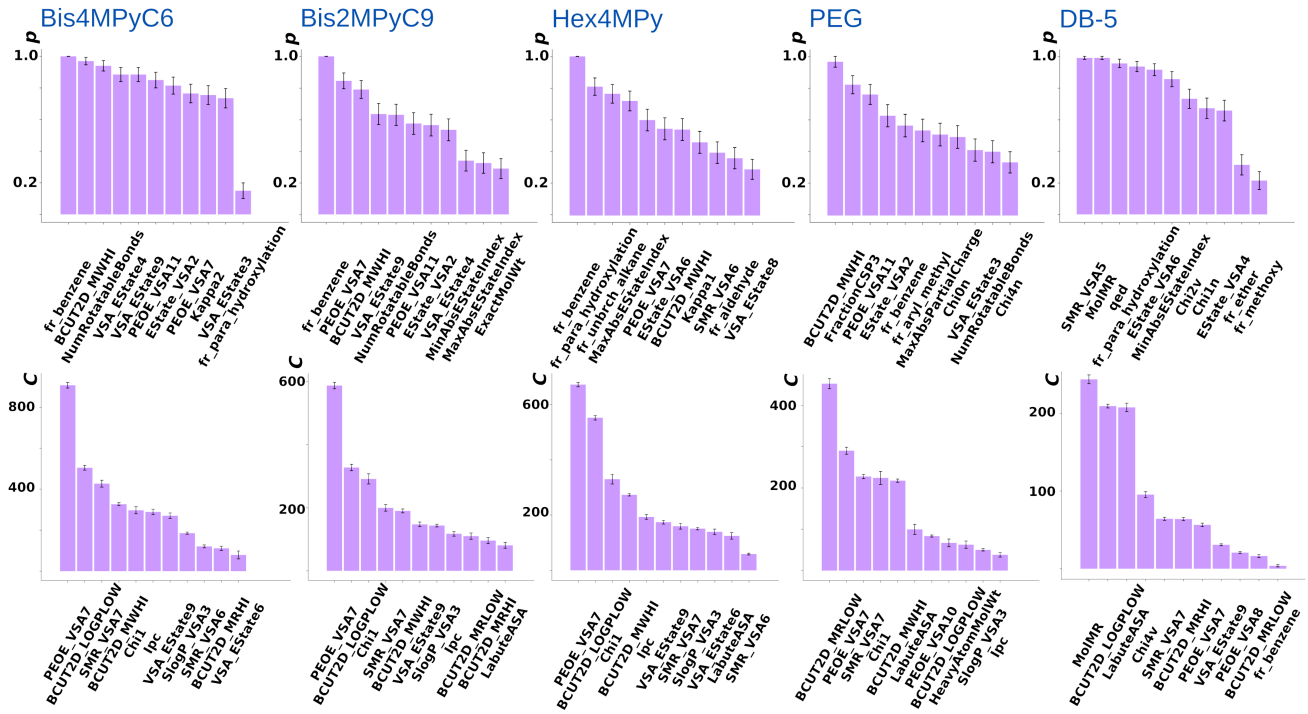

Fig. 8. MD selected by SEQ_ADD (top row) and LASSO (bottom row) methods for five SP. All diagrams are made for the identical set of compounds. The probability *p* to be selected in the SEQ_ADD procedure and the average coefficient *C* in L1-regularized linear regression are shown. Error bars show the confidence interval (p = 0.95, N = 200). 5%-phenylpolydimethylsiloxane SP is denoted as DB-5.

The third IL (Hex4MPy) considerably differs from the first two (Bis4MPyC6 and Bis2MPyC9) in structure, and the set of selected MD also significantly differs. Thus, despite all the above notes that the MD selection is not reproducible when the data set is changed, it is possible to compare SP using QSRR. Polymeric SP are even more different compared with IL-based SP. For Bis4MPyC6 and Bis2MPyC9, the fr_benzene descriptor was selected with a very high probability by the SEQ_ADD method, it is the MD that is the most correlated with RI. For Hex4MPy, it is much less significant according to the same method. The tendency continues with less polar PEG. As for siloxane, it is absent in the top 10. As expected, the polar and aromatic Bis4MPyC6 and Bis2MPyC9 are the most sensitive to aromatic systems.

The difference between the results obtained with different MD selection and MD importance estimation methods is much greater than the difference between SP. Fig. 7B-7F show the Venn diagrams for sets of MD selected using the SEQ_ADD, BORUTA, and LASSO methods. In the case of SEQ_ADD and LASSO we selected 10 the most important MD, in the case of BORUTA we selected

more than 10 MD, because each of them is considered as an important at all iterations of the Boruta algorithm.

Finally, we made the same comparison using the full versions of the data sets. The results are shown in Supplementary material, Fig. S4. It can be clearly seen that the difference between different versions of the data sets is much greater than between different SP using the same data sets. Thus, it can be concluded that QSRR-based comparisons of SP should be made using exactly the same data sets and should be made very carefully. Generally speaking, our results do not confirm the claims that the QSRR with a diverse set of MD (including topological ones) is a truly informative method that allows characterizing SP. In many cases (for example, in the work [17]) the reproducibility is not checked, the data set is not equal for different SP and it can easily result in misleading conclusions.

Table 3 contains information about the accuracy of prediction for 5 considered SP ("intersection" data set, one molecule was excluded in each repeat). It can be seen that SEQ_ADD gives better accuracy compared with LASSO, despite the smaller number of MD (l_1 = 1.0 is used). BORUTA does not select MD useful in OLS regression: the achieved accuracy is not high. The accuracy for IL is worse compared with the accuracy for polymeric SP.

Table 3. Cross-validation accuracy of RI prediction for different SP (equal set of compounds) for SEQ_ADD, LASSO, and BORUTA descriptor selection methods. Confidence intervals (p = 0.95, N = 200) are shown, the MD selection procedure was performed 200 times with exclusion of one random compound from the data set.

	SEQ_ADD		LASSO		BORUTA	
Stationary phase	RMSE	MdAE	RMSE	MdAE	RMSE	MdAE
Bis4MPyC6	107.6 ± 1.3	67.1 ± 1.5	123.2 ± 1.8	82.3 ± 1.8	256.2 ± 2.1	157.9 ± 3.2
Bis2MPyC9	73.4 ± 1.2	40.0 ± 0.9	93.8 ± 1.1	69.0 ± 1.2	172.5 ± 1.3	110.7 ± 2.1
Hex4MPy	100.0 ± 2.4	62.2 ± 1.7	110.9 ± 0.9	80.1 ± 1.7	205.7 ± 1.8	131.8 ± 1.4
PEG	68.4 ± 0.9	35.8 ± 1.0	84.0 ± 1.1	67.3 ± 1.4	134.9 ± 0.9	89.1 ± 1.5
DB-5	26.3 ± 0.3	16.6 ± 0.3	47.8 ± 0.4	25.3 ± 0.5	52.5 ± 0.5	27.3 ± 0.8

3.4. Retention indices for polyethylene glycol as new molecular descriptors

By definition, a MD is a value that can be easily computed from the structure of a molecule and that characterizes ("describes") the structure of a molecule. If we have a model that predicts the RI for a common SP (e.g., polyethylene glycol) and was trained on a large data set (e.g., the NIST 17 RI database) unrelated to the considered data, we can use the predicted RI as the new MD [26].

Table 4 shows the accuracy of predictions for three IL-based SP and various prediction models. Table 5 shows examples of linear QSRR equations for RI prediction. It can be clearly seen that the use of RI_PEG_LM and RI_PEG_DL descriptors improves the accuracy. The improvement of the accuracy is highest for Hex4MPy and Bis2MPyC9 and lowest for Bis4MPyC6. This is consistent with the fact that Bis4MPyC6 is the most polar and the most different from PEG. These MD are the most significant or are among the most significant for all three IL-based SP and for all MD selection methods. Examples of corresponding plots that show the MD importance values when these new MD are used are shown in Fig. 9AB. We did not use these MD together due to the same meaning and the strong correlation.

It should be noted that neither RI_PEG_LM nor RI_PEG_DL is enough to predict RI on IL-based SP alone without the use of other MD. It means that the selectivity and the retention mechanism for IL-based SP is considerably different from such on polyethylene glycol. Fig. 9C shows the dependence of prediction accuracy on the number of MD when RI_PEG_LM or RI_PEG_DL is used (SEQ_ADD MD selection method, 200 repeats). For both MD in all repeats, these MD are always selected in the first iteration. It is clearly seen that the use of these MD alone does not allow achieving reasonable accuracy and it works well together with other MD.

Table 4 and Fig. 9D demonstrate that the accuracy of prediction when using the RI_PEG_DL descriptor is better than when using the RI_PEG_LM descriptor. However, RI_PEG_DL is calculated by a very complex "black box" deep learning model, and this model is not an interpretable model at all. In contrast, RI_PEG_LM is calculated by an easily interpretable linear model based on understandable MD. Thus, when RI_PEG_LM is used as MD, the overall model for IL is a linear model based on MD. Supplementary material, section S2 shows the linear model that was used in order to calculate the RI_PEG_LM descriptor in explicit form. It should also be noted that when this model was trained, the training set did not contain the molecules that are contained in the data sets for IL-based SP. This way we ensured that there was no "data leak" and the molecules used for testing were not seen by the model at any stage of training.

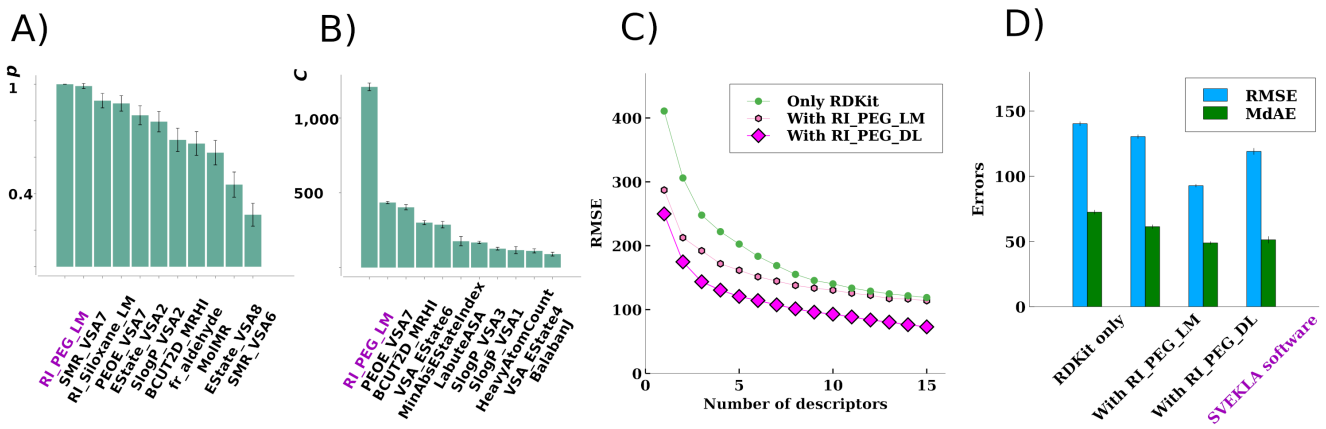

Fig. 9. A – probability p to be selected in the SEQ_ADD procedure for various MD including RI_PEG_LM for Bis4MPyC6 SP (all available compounds); B – average coefficient C in L1-regularized linear regression for various MD including RI_PEG_LM for Bis4MPyC6 SP (all available compounds); C – dependence of the accuracy (RMSE) of RI prediction for Bis4MPyC6 SP (all available compounds) on the number of MD for various sets of MD; D – accuracy (RMSE) of RI prediction for various sets of MD (ordinary least squares) and accuracy of RI prediction using a model developed using previously developed software [26]. Error bars show the confidence interval (p = 0.95, N = 200, except for the bars related to the SVEKLA software, in this case N = 20).

Table 4. Cross-validation accuracy of RI prediction for different MD sets and SP. Confidence intervals (p = 0.95, N = 200 for all cases except for the SVEKLA software) are shown. Each time, 25 molecules were excluded except for the SVEKLA software. For SVEKLA, N = 20 and no random exclusion was used.

	Bis4MPyC6		Bis2MPyC9		Hex4MPy	
Descriptor set	RMSE	MdAE	RMSE	MdAE	RMSE	MdAE
Only RDKit	140.3 ± 1.5	72.5 ± 1.5	159.1 ± 1.8	89.7 ± 1.7	172.2 ± 1.7	90.1 ± 1.5
With RI_PEG_LM	130.4 ± 1.4	61.4 ± 1.4	138.7 ± 1.5	66.6 ± 1.2	151.0 ± 1.4	70.0 ± 1.0
With RI_PEG_DL	92.8 ± 1.0	48.9 ± 1.1	108.6 ± 1.8	53.8 ± 1.0	100.1 ± 1.1	49.0 ± 0.9
SVEKLA software [26]	119.1 ± 2.4	51.3 ± 2.5	136.4 ± 2.8	51.0 ± 1.5	110.9 ± 1.0	55.4 ± 1.7

Table 5. Examples of linear equations for RI prediction. Full data sets are used for each SP, 25 molecules are randomly excluded from each data set.

Descriptor set	Stationary phase	Equation
Only RDKit	Bis4MPyC6	1272.6 - 129.2 * fr_para_hydroxylation + 774.4 * fr_benzene + 11.8 * MolMR - 98.7 * VSA_EState8 - 11.2 * EState_VSA8 + 19.6 * SMR_VSA6 + 19.1 * PEOE_VSA7 + 177.5 * Chi2n + 130.8 * BCUT2D_MRHI - 97.5 * MaxAbsEStateIndex
With RI_PEG_LM	Bis4MPyC6	- 472.7 + 0.9299 * **RI_PEG_LM** - 240.1 * fr_aldehyde + 16.2 * VSA_EState7 - 8.7 * EState_VSA8 - 14.2 * EState_VSA2 + 21.8 * SlogP_VSA2 + 20.4 * SMR_VSA7 + 11.6 * PEOE_VSA7 + 101.7 * BCUT2D_MRHI + 91.6 * MinAbsEStateIndex
With RI_PEG_DL	Bis4MPyC6	- 154.3 + 1.1221 * **RI_PEG_DL** + 31.6 * fr_unbrch_alkane - 85.0 * fr_para_hydroxylation - 68.2 * MolLogP - 30.3 * VSA_EState5 - 7.4 * EState_VSA2 + 16.5 * SlogP_VSA2 + 17.6 * SMR_VSA7 + 10.2 * PEOE_VSA7 + 3.4 * BCUT2D_MWHI
With RI_PEG_DL	Bis2MPyC9	- 129.4 + 1.0446 * **RI_PEG_DL** - 132.3 * fr_para_hydroxylation + 212.0 * fr_nitro_arom_nonortho - 17.2 * VSA_EState8 + 24.0 * VSA_EState6 + 19.9 * EState_VSA10 + 10.3 * TPSA - 16.1 * SlogP_VSA7 - 6.3 * SlogP_VSA3 - 591.6 * MaxPartialCharge
With RI_PEG_DL	Hex4MPy	2043.6 + 1.3862 * **RI_PEG_DL** - 160.9 * fr_aryl_methyl - 182.6 * fr_C_O_noCOO + 60.7 * VSA_EState4 + 31.4 * SMR_VSA1 + 5.7 * PEOE_VSA7 - 645.4 * BCUT2D_LOGPHI - 101.3 * BCUT2D_MWLOW - 122.2 * FpDensityMorgan1 - 1.4 * HeavyAtomMolWt

3.5. Comparison with previously published software

A complex two-stage method was recently developed [26] that allows using all benefits of deep learning for RI prediction using training sets with ~100-200 compounds. The idea of this software is similar to the considered above: deep learning models predict, using a molecule structure, RI for multiple common SP (siloxanes, polyethylene glycol), and then these predicted RI are used as input features for a new model for the given SP and data set. Together with these features (RI for a set of SP), other MD are also used. The difference with the approach considered in this work is that this set of features is fed to a linear support vector regression model (with a non-linear kernel) with predefined parameters without any MD preselection. This software (we call it SVEKLA) [26, 35] allows creating a

machine learning model for any SP easily, but these models are not interpretable and use an excessive set of features.

Table 4 demonstrates the accuracy achieved by SVEKLA software and the accuracy achieved by linear regression. The accuracy achieved by SVEKLA software is approximately the same or even worse compared with the use of linear equations with the RI_PEG_DL descriptor. But this model uses much less features (and only one deep learning-based MD), is linear and is much more interpretable. RI_PEG_LM is calculated using a linear model. The use of this MD is the most simple and interpretable way to accurately predict RI for IL. A graphical comparison of the accuracy of several different approaches is shown in Fig. 9D.

3.6. CHERESHNYA – interactive software for QSRR studies in gas chromatography

We have developed the interactive software for QSRR studies in GC and called it CHERESHNYA, the example of a screenshot is shown in Fig. 10. This software allows the interactive MD generation (2D MD supported by RDKit and CDK packages), MD selection, building of linear (OLS) models for QSRR in GC. The newly developed RI_PEG_LM and RI_PEG_DL descriptors are also supported, as well as similar MD for polydimethylsiloxane, 5%-phenyl-methylpolysiloxane, 94%-dimethyl-6%-cyanopropyl-phenyl-polylsiloxane. All MD selection methods listed in Table 1 and described in section 2.4.2 are implemented in this software. The software is written in the Java programming language, Smile framework [40] is used. PLS-VIP and GA methods are implemented using Scikit-learn package. The molecular editor JSME [45] is integrated into the software for interactive MD computation and RI prediction. The figures (heatmaps, bar plots) shown in this article are generated using this software. The reproducibility study can be automatically provided using it.

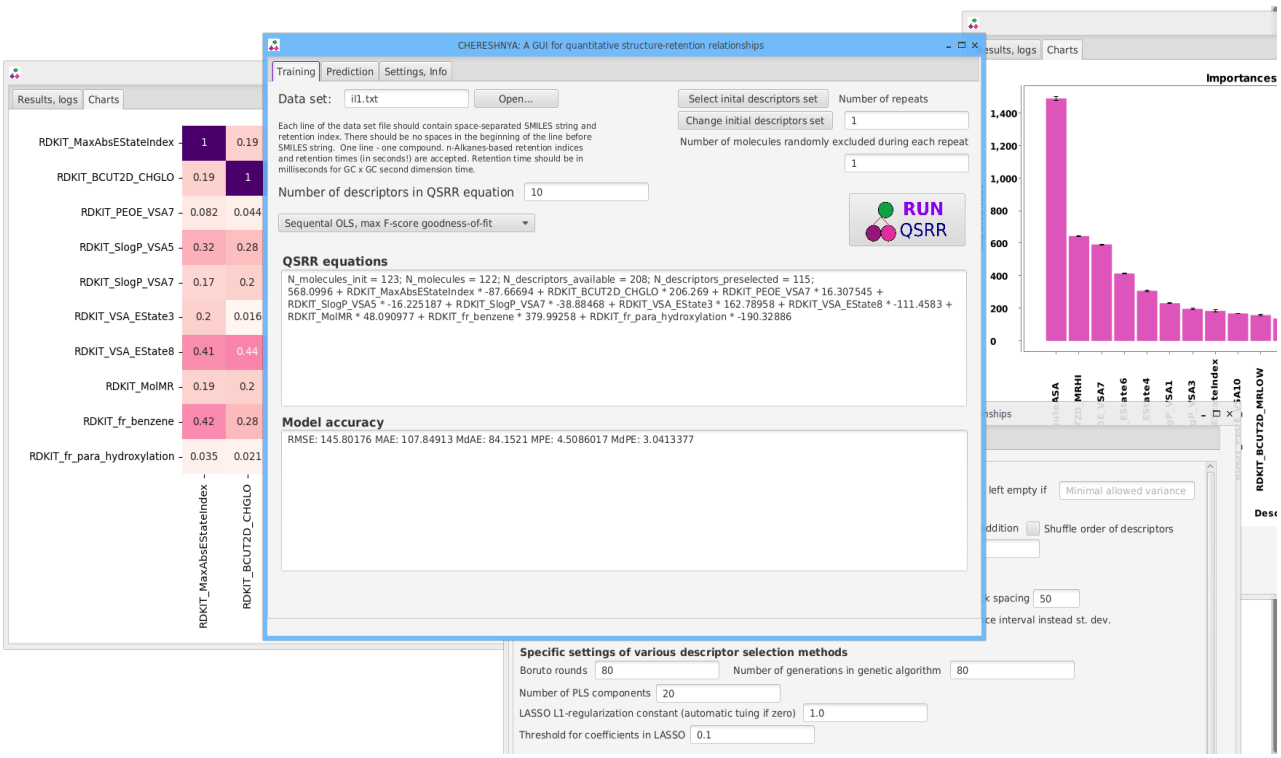

Fig. 10. Screenshot of the CHERESHNYA software (two copies of software run).

The software is free, open-source and available under the GNU General Public License (version 3.0), all components and dependencies are also free software. The prebuilt binaries are available for Linux and Windows operating systems. The software can be downloaded from the repository: https://github.com/mtshn/chereshnya

4. Conclusions

In this work, a data set of retention indices on three ionic liquid-based stationary phases was acquired for a diverse set of molecules of various classes. This is the first such data set to be published. This data set can be used in further QSRR studies and as a benchmark in works about machine learning. Using this data set, a study devoted to reproducibility of the descriptor selection and descriptor importance estimation was carried out.

Methods for selecting descriptors for constructing linear quantitative structure-retention relationships are not reproducible with respect to changes in the data set. Different selection methods give different results. Conclusions about the retention mechanism and comparison of stationary phases based on such quantitative relationships must be made with extreme caution. Some previous works did not carry out any checks on the reproducibility of the selection of descriptors, but qualitative

conclusions were drawn from the fact which descriptors were selected. Such conclusions are unreliable and should be avoided.

The selectivity of the considered stationary phases significantly differs from the selectivity of polyethylene glycol. The retention on ionic liquids cannot be directly computed using only the retention index on polyethylene glycol. However, the retention index on polyethylene glycol predicted using a machine learning model (trained on other, non-overlapping data) is a very good descriptor for predicting retention indices on ionic liquids. Sufficiently accurate linear models for retention index prediction were developed for these stationary phases.

The interactive software with a graphical user interface for QSRR studies in gas chromatography that includes calculation of various descriptors, descriptor selection and other tasks was developed. This software is free, open-source and can be downloaded from the above-mentioned Github repository.

Funding

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

Data availability

The data set for further QSRR studies containing retention times and indices can be downloaded from the Figshare repository https://doi.org/10.6084/m9.figshare.16885009.

References

1. Ho, T. D.; Zhang, C.; Hantao, L. W.; Anderson, J. L. Ionic Liquids in Analytical Chemistry: Fundamentals, Advances, and Perspectives. *Analytical Chemistry* **2014**, *86*(1), 262–285. doi:10.1021/ac4035554.

2. *Ionic Liquids in Analytical Chemistry*; Elsevier, 2022. doi:10.1016/C2019-0-04941-2.

3. Poole, C. F.; Poole, S. K. Ionic liquid stationary phases for gas chromatography. *Journal of Separation Science* **2011**, *34*(8), 888–900. doi:10.1002/jssc.201000724.

4. Yao, C.; Anderson, J. L. Retention characteristics of organic compounds on molten salt and ionic liquid-based gas chromatography stationary phases. *Journal of Chromatography A* **2009**, *1216*(10), 1658–1712. doi:10.1016/j.chroma.2008.12.001.

5. Cagliero, C.; Bicchi, C. Ionic liquids as gas chromatographic stationary phases: how can they change food and natural product analyses? *Analytical and Bioanalytical Chemistry* **2020**, *412*(1), 17–25. doi:10.1007/s00216-019-02288-x.

6. Aslani, S.; Armstrong, D. W. Ionic liquids as gas chromatography stationary phases. In *Ionic Liquids in Analytical Chemistry*; Elsevier, 2022; pp 171–202. doi:10.1016/B978-0-12-823334-4.00011-4.

7. De Boer, J.; Blok, D.; Ballesteros-Gómez, A. Assessment of ionic liquid stationary phases for the determination of polychlorinated biphenyls, organochlorine pesticides and polybrominated diphenyl ethers. *Journal of Chromatography A* **2014**, *1348*, 158–163. doi:10.1016/j.chroma.2014.05.001.

8. Cagliero, C.; Mazzucotelli, M.; Rubiolo, P.; Marengo, A.; Galli, S.; Anderson, J. L.; et al. Can the selectivity of phosphonium based ionic liquids be exploited as stationary phase for routine gas chromatography? A case study: The use of trihexyl(tetradecyl) phosphonium chloride in the flavor, fragrance and natural product fields. *Journal of Chromatography A* **2020**, *1619*, 460969. doi:10.1016/j.chroma.2020.460969.

9. Poole, C. F.; Lenca, N. Gas chromatography on wall-coated open-tubular columns with ionic liquid stationary phases. *Journal of Chromatography A* **2014**, *1357*, 87–109. doi:10.1016/j.chroma.2014.03.029.

10. Shashkov, M. V.; Sidelnikov, V. N. Properties of columns with several pyridinium and imidazolium ionic liquid stationary phases. *Journal of Chromatography A* **2013**, *1309*, 56–63. doi:10.1016/j.chroma.2013.08.030.

11. Ros, M.; Escobar-Arnanz, J.; Sanz, M. L.; Ramos, L. Evaluation of ionic liquid gas chromatography stationary phases for the separation of polychlorinated biphenyls. *Journal of Chromatography A* **2018**, *1559*, 156–163. doi:10.1016/j.chroma.2017.12.029.

12. Shashkov, M. V.; Sidelnikov, V. N. Mass spectral evaluation of column bleeding for imidazolium-based ionic liquids as GC liquid phases. *Analytical and Bioanalytical Chemistry* **2012**, *403*(9), 2673–2682. doi:10.1007/s00216-012-6020-9.

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

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

15. Su, Q.-Z.; Vera, P.; Salafranca, J.; Nerín, C. Decontamination efficiencies of post-consumer high-density polyethylene milk bottles and prioritization of high concern volatile migrants. *Resources, Conservation and Recycling* **2021**, *171*, 105640. doi:10.1016/j.resconrec.2021.105640.
16. Kaliszan, R. QSRR: Quantitative Structure-(Chromatographic) Retention Relationships. *Chemical Reviews* **2007**, *107*(7), 3212–3246. doi:10.1021/cr068412z.
17. Yan, J.; Cao, D.-S.; Guo, F.-Q.; Zhang, L.-X.; He, M.; Huang, J.-H.; 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* **2012**, *1223*, 118–125. doi:10.1016/j.chroma.2011.12.020.
18. Ahmadi, S.; Lotfi, S.; Hamzehali, H.; Kumar, P. A simple and reliable QSPR model for prediction of chromatography retention indices of volatile organic compounds in peppers. *RSC Advances* **2024**, *14*(5), 3186–3201. doi:10.1039/D3RA07960K.
19. Chen, X.; Li, H.-D.; Guo, F.-Q.; Yan, J.; Cao, D.-S.; Liang, Y.-Z. QSRR Study on Flavor Compounds of Diverse Structures on Different Columns with the Help of New Chemometric Methods. *Chromatographia* **2013**, *76*(5–6), 241–253. doi:10.1007/s10337-012-2349-7.
20. Sepehri, B.; Ghavami, R.; Farahbakhsh, S.; Ahmadi, R. Machine learning-based quantitative structure–retention relationship models for predicting the retention indices of volatile organic pollutants. *International Journal of Environmental Science and Technology* **2022**, *19*(3), 1457–1466. doi:10.1007/s13762-021-03271-9.
21. Fan, Y.; Deng, Y.; Yang, Y.; Deng, X.; Li, Q.; Xu, B.; et al. Modelling and predicting liquid chromatography retention time for PFAS with no-code machine learning. *Environmental Science: Advances* **2024**, *3*(2), 198–207. doi:10.1039/D3VA00242J.
22. Walczak-Skierska, J.; Szultka-Młyńska, M.; Pauter, K.; Buszewski, B. Study of chromatographic behavior of antibiotic drugs and their metabolites based on quantitative structure-retention relationships with the use of HPLC-DAD. *Journal of Pharmaceutical and Biomedical Analysis* **2020**, *184*, 113187. doi:10.1016/j.jpba.2020.113187.
23. Svrkota, B.; Krmar, J.; Protić, A.; Otašević, B. The secret of reversed-phase/weak cation exchange retention mechanisms in mixed-mode liquid chromatography applied for small drug molecule analysis. *Journal of Chromatography A* **2023**, *1690*, 463776. doi:10.1016/j.chroma.2023.463776.
24. Danishuddin; Khan, A. U. Descriptors and their selection methods in QSAR analysis: paradigm for drug design. *Drug Discovery Today* **2016**, *21*(8), 1291–1302. doi:10.1016/j.drudis.2016.06.013.

738 25. Zhokhov, A. K.; Loskutov, A. Yu.; Rybal'chenko, I. V. Methodological Approaches to the
739 Calculation and Prediction of Retention Indices in Capillary Gas Chromatography. *Journal of*
740 *Analytical Chemistry* **2018**, *73*(3), 207–220. doi:10.1134/S1061934818030127.
741 26. Matyushin, D. D.; Sholokhova, A. Yu.; Buryak, A. K. Deep Learning Based Prediction of Gas
742 Chromatographic Retention Indices for a Wide Variety of Polar and Mid-Polar Liquid Stationary
743 Phases. *International Journal of Molecular Sciences* **2021**, *22*(17), 9194. doi:10.3390/ijms22179194.
744 27. Matyushin, D. D.; Sholokhova, A. Yu.; Buryak, A. K. Gradient boosting for the prediction of gas
745 chromatographic retention indices. *Сорбционные и хроматографические процессы* **2019**, *19*(6),
746 630–635. doi:10.17308/sorpchrom.2019.19/2223.
747 28. Goodarzi, M.; Jensen, R.; Vander Heyden, Y. QSRR modeling for diverse drugs using different
748 feature selection methods coupled with linear and nonlinear regressions. *Journal of Chromatography B*
749 **2012**, *910*, 84–94. doi:10.1016/j.jchromb.2012.01.012.
750 29. Escobar-Arnanz, J.; Sanz, M. L.; Ros, M.; Sanz, J.; Ramos, L. Potential of topological descriptors
751 to model the retention of polychlorinated biphenyls in different gas chromatography stationary phases,
752 including ionic liquid-based columns. *Journal of Chromatography A* **2020**, *1616*, 460844.
753 doi:10.1016/j.chroma.2019.460844.
754 30. Rabhi, F.; Hussard, C.; Sifaoui, H.; Mutelet, F. Characterization of bis(fluorosulfonyl)imide based
755 ionic liquids by gas chromatography. *Journal of Molecular Liquids* **2019**, *289*, 111169.
756 doi:10.1016/j.molliq.2019.111169.
757 31. Kulsing, C.; Nolvachai, Y.; Zeng, A. X.; Chin, S.; Mitrevski, B.; Marriott, P. J. From Molecular
758 Structures of Ionic Liquids to Predicted Retention of Fatty Acid Methyl Esters in Comprehensive Two-
759 Dimensional Gas Chromatography. *ChemPlusChem* **2014**, *79*(6), 790–797.
760 doi:10.1002/cplu.201300410.
761 32. Shashkov, M. V.; Sidelnikov, V. N.; Zaikin, P. A. Selectivity of stationary phases based on
762 pyridinium ionic liquids for capillary gas chromatography. *Russian Journal of Physical Chemistry A*
763 **2014**, *88*(4), 717–721. doi:10.1134/S0036024414040268.
764 33. Ilkova, E. L.; Mistryukov, E. A. A Simple Versatile Method for Coating of Glass Capillary
765 Columns. *Journal of Chromatographic Science* **1971**, *9*(9), 569–570. doi:10.1093/chromsci/9.9.569.
766 34. Shashkov, M. V.; Sidel'nikov, V. N. Single cation ionic liquids as high polarity thermostable
767 stationary liquid phases for capillary chromatography. *Russian Journal of Physical Chemistry A* **2012**,
768 *86*(1), 138–141. doi:10.1134/S0036024412010268.

769 35. Sholokhova, A. Yu.; Matyushin, D. D.; Grinevich, O. I.; Borovikova, S. A.; Buryak, A. K.
770 Intelligent Workflow and Software for Non-Target Analysis of Complex Samples Using a Mixture of
771 Toxic Transformation Products of Unsymmetrical Dimethylhydrazine as an Example. *Molecules* **2023**,
772 *28*(8), 3409. doi:10.3390/molecules28083409.
773 36. Software for predicting gas chromatographic retention indices and mass spectra.
774 <https://github.com/mtshn/svekla>.
775 37. Willighagen, E. L.; Mayfield, J. W.; Alvarsson, J.; Berg, A.; Carlsson, L.; Jeliazkova, N.; et al. The
776 Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure
777 searching. *Journal of Cheminformatics* **2017**, *9*(1), 33. doi:10.1186/s13321-017-0220-4.
778 38. Nguyen, K. T.; Blum, L. C.; van Deursen, R.; Reymond, J. Classification of Organic Molecules by
779 Molecular Quantum Numbers. *ChemMedChem* **2009**, *4*(11), 1803–1805.
780 doi:10.1002/cmdc.200900317.
781 39. Klekota, J.; Roth, F. P. Chemical substructures that enrich for biological activity. *Bioinformatics*
782 **2008**, *24*(21), 2518–2525. doi:10.1093/bioinformatics/btn479.
783 40. Smile - Statistical Machine Intelligence and Learning Engine. <haifengl.github.io>.
784 41. Kursa, M. B.; Jankowski, A.; Rudnicki, W. R. Boruta – A System for Feature Selection.
785 *Fundamenta Informaticae* **2010**, *101*(4), 271–285. doi:10.3233/FI-2010-288.
786 42. Mehmood, T.; Liland, K. H.; Snipen, L.; Sæbø, S. A review of variable selection methods in Partial
787 Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems* **2012**, *118*,
788 62–69. doi:10.1016/j.chemolab.2012.07.010.
789 43. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; et al. Scikit-learn:
790 Machine Learning in Python. arXiv June 5, 2018. <http://arxiv.org/abs/1201.0490> Accessed 24.03.19.
791 44. Hall, L. H.; Kier, L. B. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in
792 Structure-Property Modeling. In *Reviews in Computational Chemistry*. Lipkowitz, K. B., Boyd, D. B.,
793 Eds.; Wiley, 1991; Vol. 2, pp 367–422. doi:10.1002/9780470125793.ch9.
794 45. Bienfait, B.; Ertl, P. JSME: a free molecule editor in JavaScript. *Journal of Cheminformatics* **2013**,
795 *5*(1), 24. doi:10.1186/1758-2946-5-24.