separations

Article

Comparative Prediction of Gas Chromatographic Retention Indices for GC/MS Identification of Chemicals Related to Chemical Weapons Convention by Incremental and Machine Learning Methods

Albert Kireev [1,*], Sergey Osipenko [2], Gary Mallard [3], Evgeny Nikolaev [4] and Yury Kostyukevich [1,*]

[1] Center for Molecular and Cellular Biology, Skolkovo Institute of Science and Technology, Bolshoy Boulevard 30, Bld. 1, 121205 Moscow, Russia
[2] Center for Computational and Data-Intensive Science and Engineering, Skolkovo Institute of Science and Technology, Bolshoy Boulevard 30, Bld. 1, 121205 Moscow, Russia
[3] Teal Consulting, Chevy Chase, MD 20815, USA
[4] Laboratory of Mass-Spectrometry, Skolkovo Institute of Science and Technology, Bolshoy Boulevard 30, Bld. 1, 121205 Moscow, Russia
* Correspondence: a.kireev@skoltech.ru (A.K.); y.kostyukevich@skoltech.ru (Y.K.)

Abstract: During on-site verification activities conducted by the Technical Secretariat of Organization for the Prohibition of Chemical Weapons, identification by gas chromatography retention indices (RI) data, in addition to mass spectrometry data, increase the reliability of factual findings. However, reference RIs do not cover all the possible chemical structures. That is why it is important to have models to predict RIs. Applicable only for narrow data sets of chemicals with a fixed scaffold (G- and V-series gases as example), the non-learning incremental method demonstrated predictive median absolute and percentage errors of 2–4 units and 0.1–0.2%; these are comparable with the experimental bias in RI measurements in the same laboratory with the same GC conditions. It outperforms the accuracy of two reported machine learning methods–median absolute and percentage errors of 11–52 units and 0.5–2.8%. However, for the whole Chemical Weapons Convention (CWC) data set of chemicals, when a fixed scaffold is absent, the incremental method is not applicable; essential machine learning methods achieved accuracy: median absolute and percentage errors of 29–33 units and 0.5–2.2%, depending on the machine learning method. In addition, we have developed a homology tree approach as a convenient method for the visualization of the CWC chemical space. We conclude that non-learning incremental methods may be more accurate than the state-of-the-art machine learning techniques in particular cases, such as predicting the RIs of homologues and isomers of chemicals related to CWC.

Keywords: gas chromatography; retention indexes; database; predictions; machine learning; deep learning

Citation: Kireev, A.; Osipenko, S.; Mallard, G.; Nikolaev, E.; Kostyukevich, Y. Comparative Prediction of Gas Chromatographic Retention Indices for GC/MS Identification of Chemicals Related to Chemical Weapons Convention by Incremental and Machine Learning Methods. *Separations* **2022**, *9*, 265. https://doi.org/10.3390/separations9100265

Academic Editor: Josef Cvačka

Received: 15 July 2022
Accepted: 8 September 2022
Published: 22 September 2022

1. Introduction

Gas chromatography (GC) is one of the core analytical methods recommended for the analysis of chemicals related to the Chemical Weapons Convention (CWC) [1]. GC hyphenated with electron ionization mass spectrometry (GC/MS) is a predominant experimental technique established by the Technical Secretariat of Organization for the Prohibition of Chemical Weapons (OPCW) for factual findings during on-site verification activities with sampling and analysis. The chemicals that have been identified for the application of verification measures are listed in the schedules contained in the Annex on Chemicals. In the sense of the Chemical Weapons Convention, Schedule 1 substances are chemicals that can either be used as chemical weapons themselves or used in the manufacture of chemical weapons; and which have no, or very limited, uses outside of chemical warfare.

Examples are mustard and nerve agents; and substances that are solely used as precursor chemicals in their manufacture. Schedule 2 substances are chemicals that can either be used as chemical weapons themselves or used in the manufacture of chemical weapons; however, they have small-scale applications outside of chemical warfare and thus, can be legitimately manufactured in small quantities. An example is thiodiglycol, which can be used in the manufacture of mustard agents; however, it is also used as a solvent in inks. Schedule 3 substances are chemicals that can either be used as toxic chemical weapons themselves or used in the manufacture of chemical weapons; however, they also have legitimate large-scale industrial uses. Examples of these substances are phosgene, which has been used as a chemical weapon, but which is also a precursor in the manufacture of many legitimate organic compounds; and triethanolamine, used in the manufacture of nitrogen mustard, but also commonly used in toiletries and detergents. All the schedules are sub-divided into Part A substances, which are chemicals that can be used directly as weapons; and Part B, which are precursors useful in the manufacture of chemical weapons. The scope of chemicals related to CWC is not limited to chemical warfare agents; their precursors; and degradation products or derivatives. It depends on the context of the activity; for instance, under which article of CWC, the mission is conducted, or in which framework of proficiency tests, analysis is carried out. Depending on the conditions, riot control agents, non-scheduled toxic industrial chemicals, or novel agents may also be relevant [1]. The origin of the samples can include industrial, environment, food, medicine, and biological matrices.

The absence of retention indexes (RI) [2] can lead to a false positive identification during an inspection or an investigation of alleged use; and could have adverse diplomatic consequences [1]. The involvement of retention parameters together with mass spectral data improves the accuracy of mass spectrometry identification. The results of a library search by mass spectra can provide relevance to the schedule of CWC and its scaffold. Usually, a mass spectrum formed and collected with electron ionization does not contain molecular ions; in this case, RI would be helpful to provide a clue about the length and isomerization of alkyl chains inside of the Schedule. The GC RI used for identification are part of the official OPCW Central Analytical Database (OCAD) [3]. OCAD for CWC-related chemicals is carefully curated and well-maintained by a validation group. The common problem inherent in such analytical banks is that RI data are limited for the included chemicals; and are absent for chemicals that are not in the database. As an example, the OCAD version 21, issued in 2019, contains 5292 RIs for 4482 chemicals. For comparison, when all the possible permutations and combinations are considered, Schedules 1.A.1–1.A.12 of CWC alone contain more than 1,300,000 possible chemicals; this is the case without including protonated salts and unspecified alkyl quaternary salts. The vast majority of possible structures belong to phosphorus-containing nerve agents, their precursors, and degradation products. After the addition of four more entries, 1.A.13–1.A.16, to the Schedules of CWC in 2020 (among them, "Novichok agents"), the problem became even more challenging. This problem may be overcome using estimated RI values.

There are several approaches to generate in silico RI values from the structure. Existing methods for RI predictions can be divided into quantitative structure–retention relationship (QSRR) methods based on molecular descriptors (MD); deep learning approaches that operate with graph-based or text-based representations; and non-learning methods based on functional group increments [4,5]. Recent advances in the theory and applications of the deep neural networks justified the appearance of several models to predict RI using 1D or 2D convolutional networks (1D-CNN [6], 2D-CNN [7]); graph neural networks [8]; or ensemble models that use both representations and human-designed features as input [9]. Despite the different architectures, all the mentioned models were trained on various releases of the NIST RI database. A model with graph neural networks allows predicting the Kovats retention index with a mean unsigned error of 28 index units as compared to 44; the putative best result using a convolutional neural network [8]. Ensemble models [9] were tested using various diverse data sets: flavor compounds, essential oils, and metabolomics-

related compounds. The achieved accuracy was as follows: the median absolute and percentage errors were 6–40 units and 0.8–2.2%. Accuracy depends on a test data set. Although all the models demonstrate excellent accuracy, they are still limited by the chemical nature of the compounds in the training set that restricts the applicability domain; in addition, it is probable that more specific methods may enhance the predictive accuracy for narrow domain tasks, such as modeling the RIs of CWC compounds.

The main purpose of this work was to evaluate broad domain models trained on an NIST RI database and available for RI predictions with respect to the CWC chemicals; and to compare them with narrow-domain models trained on OCAD and with non-learning approaches.

2. Materials and Methods

2.1. Datasets

Retention data from the OPCW chemical analysis database (OCAD, v.21) was kindly provided by the OCPW. OCAD initially contains 4397 unique compounds with RI values in the range 488–3309, along with OPCW names. The description of compounds includes their chemical name, OPCW code, CAS, and schedule number [3].

The typical instrument information includes:

GC column–HP-MS or DB-5MS, 30 m × 0.25 mm × 0.25 μm;

GC temperature program 40 °C (1 min), 10 °C/min, 280 °C (10 min);

Carrier Gas Helium, 1 mL/min (constant flow);

Injection temperature 200 °C;

Reference standard GC/MS Sample Preparation Kit (C-I/DEC.71), standard chemicals, OPCW GC/MS test mixture chemicals.

During preparation, validation and standardization procedures, SMILES strings [10] were assigned for each compound; 31 molecules containing deuterium were removed. The list of heavy atoms in the OCAD molecules was limited to C, N, O, F, Si, P, S, Cl, Se, and As.

The fine-tuning of the Transformer-CNN [11] model was performed on the NIST 17 RI database. Only Kováts and linear indexes for non-polar and semi-standard non-polar columns were used. Multiple values were averaged.

2.2. Machine Learning Models

The 1D-CNN-based predictor was introduced recently and was downloaded from the source [6]. The predictor is provided as a Java application that contains the pre-trained 1D-CNN model. DeepReI software [7] is an R-package that includes optimized weights for a 2D-CNN model. Both models take SMILES strings as input and incorporate the one-hot encoding algorithm.

Transformer-CNN is a deep learning approach that involves data-intensive pre-training in a self-supervised mode using a large non-labeled library. The resulting pre-trained model can be further fine-tuned on a number of regression or classification tasks. The details are available in the original publication [11].

A narrow applicability domain model specified for CWC-related chemicals was trained with XGBoost library on descriptors from the Mordred library [12]. The OCAD was used as training set in 5-fold cross-validation mode (CV). The hyperparameters were optimized via grid search 5-fold CV protocol. The initial grid is presented in the supplementary materials (Table S1). The optimal tree depth, number of estimators, and learning rate were 4, 1000, and 0.05, respectively.

2.3. The Increment Predictions

The proposed incremental method was based on the observation that the RI difference in two pairs of molecules that differ by the side alkyl chain inside the pair and by scaffold between the pairs should be preserved (Figure 1). Such an effect is a result of the "relative" nature of the Kováts or linear indexes that are derived as retention values, normalized on the n-alkanes retention times; while the last have RIs of 100 multiplied by the length of the

chain by definition. As the structure of the CWC includes a number of homologues series and contains a significant number of molecules, such observation can be applied to predict the RI values for chemical compounds in most CWC schedules (which are organized by scaffolds). Thus, the prediction requires at least three molecules with known RI values and appropriate substituents. Moreover, a set of combinations may be used to obtain reliable and statistically significant predictions.

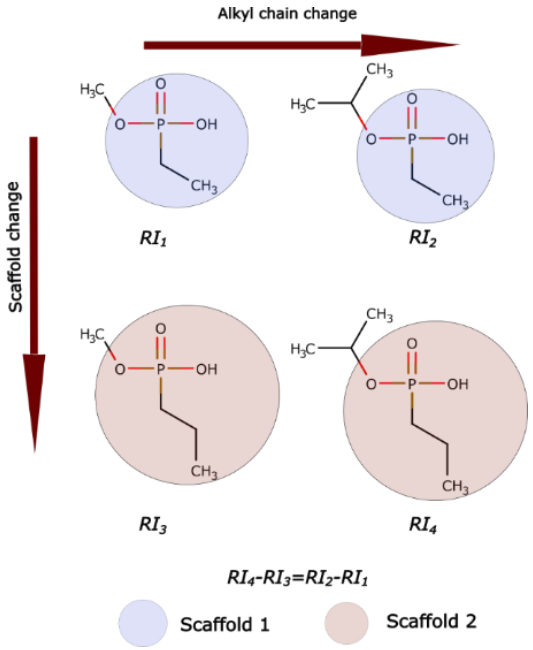

Figure 1. The explanation of the proposed increment-based approach.

The isolating of necessary molecules may be done manually using, for example, IUPAC names. However, for convenience, we constructed a Python script that extracts scaffolds by cutting off the terminal alkyl chains; and compares the substituents to find molecules that are required to obtain predictions.

A predictor that is based on a narrow domain XGBoost model and the implementation of the increment method is available as a prototype of a web application at https://ri-cwc.anvil.app (accessed on 1 September 2022).

3. Results and Discussion

3.1. Evaluating Available RI Prediction Models

The state-of-the-art approach in predicting RIs is to use NIST-based deep learning models with various architectures. In fact, these models demonstrate excellent predictive accuracy with a mean absolute error of about 25–30 units in the cross-validation mode [6]. Nevertheless, the authors often report that the accuracy may dramatically decrease when the model is evaluated on external datasets. The possible explanation is that such datasets contain molecules that are out of the chemical space covered by the training set. As OCAD content is structurally different from molecules in the NIST 17 database (Figure 2A) and these databases have a very limited intersection (Figure 2B), it was essential to evaluate the state-of-the-art models on CWC chemicals. We have chosen two models that were recently proposed for RI predictions. These models reflect the most popular 1D-CNN [6] and 2D-CNN [7] architectures. For a GNN-based predictor [8] we fine-tuned Transformer-CNN [11], which was reported to be very powerful for various QSPR tasks [11] on the NIST 17 database. The models' behavior on OCAD is presented in Table 1. 1D-CNN demonstrated the lowest value of MAE; which is, however, higher than that reported for the cross-validation mode. It is worth noticing that these open-source predictors may fail with predictions for particular molecules. For example, some models [6] were trained on

molecules with restricted elemental composition; and cannot make predictions for arsenic or selenic compounds, which are included into CWC. For that reason, such molecules were excluded from the evaluation where required. It is necessary to note that the 1D-CNN and 2D-CNN models were trained on the subsets of the NIST database taken for various stationary phases. While 2D-CNN was trained to predict RI values for the semi-standard non-polar phases, 1D-CNN was trained on both non-polar and semi-standard non-polar retention data. However, since the replicated training data were averaged for both phases, the difference in their selectivity was thus neglected; moreover, it is not obvious how it might affect the overall accuracy.

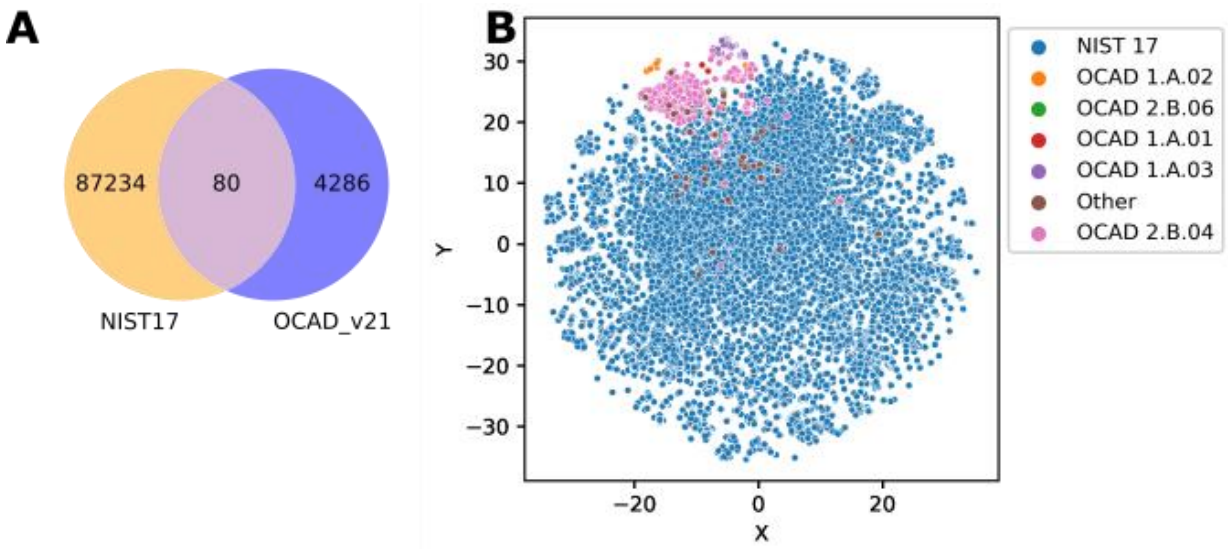

Figure 2. (**A**) Venn diagram of the NIST 17 RI database and OCAD v. 21. (**B**) t-SNE plot of the chemical space covered by the NIST 17 database and OCAD v21. Morgan fingerprints were used for the calculations.

Table 1. Evaluating various publicly available models on OCAD.

Model	Mean Absolute Error	Median Absolute Error	Mean Relative Error, %	Median Relative Error, %
1D-CNN	39.95	28.77	2.68	1.88
2D-CNN	51.46	38.00	3.32	2.53
Transformer-CNN	48.11	33.45	3.23	2.17

Figure 3 demonstrates the distribution of the errors for the evaluated methods. 1D-CNN provides a narrower peak around zero, which means it provides more accurate predictions for the majority of chemicals from OCAD. Considering this together with the MAE values, we can conclude that the predictor based on 1D-CNN is the most appropriate for filling missed RI values in OCAD.

3.2. Domain Specific Modeling

Although large and diverse training sets are required to create a model with good generalization ability, a model trained on a restricted chemical domain may be a better fit for specific tasks. For that reason, we attempted to create such a model trained on OCAD that can predict the RI of CWC compounds. As OCAD contains less than 5000 compounds, deep neural networks would definitely overfit. Therefore, we turned to traditional ML models and evaluated a gradient boosting algorithm that is available via the XGBoost [13] Python library. This algorithm was previously successfully applied for liquid chromatography (LC) retention predictions [14–16]. Molecular descriptors from the Mordred [12] library were used as input. The mean absolute error in 5-fold CV was 16.2 $\pm$ 0.86 units; it was reduced by approximately 20 units, compared with the broad domain deep learning models.

However, we need to outline that this model is appropriate only for compounds that are structurally similar for compounds from CWC; in addition, the MAE value was obtained in cross-validation mode.

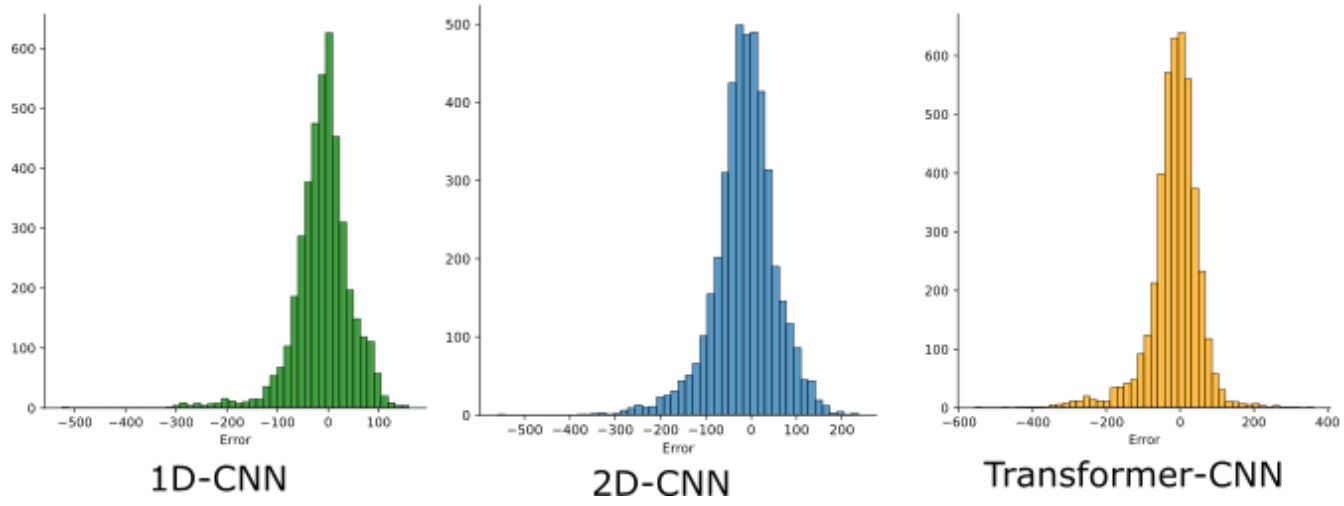

Figure 3. Error distribution for the evaluated deep learning methods [6,7,11] on OCAD data. The X-axis represents the difference between the experimental and predicted values.

3.3. Increment-Based Method

Machine learning based predictions still have a lack of accuracy for several reasons. First, they rely on the data from the NIST database that is collected in non-standardized conditions; and may vary even for the same type of chromatographic column. Second, the diversity of molecules in the NIST base cannot cover the whole chemical space; and the predictions may be distorted for particular classes of chemicals. In such cases, non-learning methods may provide a solution for accurate predictions. Here, we propose using a special case of the increment method that considers RI as a linear combination of the number of functional groups multiplied by their increment values. The increment values were estimated on a large set of molecules to fit the experimental values. This method was widely applied for a decade because of relatively high accuracy [17] and is now being replaced by deep learning models. Our implementation of the increment-based approach is based on the observation that the RI difference between molecular pairs that vary on the alkyl side chain should be preserved when the scaffold is changed and the side chains remain the same (Figure 1). Since the CWC Schedules are organized by scaffolds (i.e., "soman Schedule", "VX Schedule") and contain homologous compounds that differ by their alkyl side chains, this approach has the potential to improve RI predictions within a Schedule (Table 2). To evaluate this approach, we created a Python script that searches required homologous molecules to make predictions. Another option is to use OPCW names to identify scaffolds and side chains; this can be done with high throughput using electronic spreadsheets. In the cases where several combinations are available for prediction, the average values may be used as output to add confidence in the prediction results and estimate the prediction error. A set of RI differences for various alkyl radical pairs, along with standard deviations, is available in the supplementary materials (Table S2).

Table 2. An example of initial RI data, taken from the OPCW central analytical database (OCAD).

Chemical Name	CAS	Formula	RIs in OCAD	Schedule
1,2-Dimethylbutyl methylphosphonofluoridate	1005239-89-3	C7H16FO2P	1087	1.A.01
1,2-Dimethylbutyl ethylphosphonofluoridate	1005258-27-4	C8H18FO2P	1180	1.A.01
1,2-Dimethylbutyl isopropylphosphonofluoridate	1005249-16-0	C9H20FO2P	1231	1.A.01
1,2-Dimethylbutyl propylphosphonofluoridate	N0014	C9H20FO2P	1264	1.A.01
1,2-Dimethylpropyl ethylphosphonofluoridate	N0032	C7H16FO2P	1087	1.A.01

An example of the application of the proposed approach for filling missing values in OCAD is given in Figure 4.

Example data from OCAD

RO linked to phosphorus	Methyl phosphonofluoridate	Ethyl phosphonofluoridate	Isopropyl phosphonofluoridate	Propyl phosphonofluoridate
1-Propylcyclohexyl	1420	1517	**Missing value 1**	**Missing value 2**
1,2-Dimethylbutyl	1087	1180	1231	1264
Difference	**333**	**337**	**335±3**	**335±3**

Missing value 1 = 1231+335 = 1566

Missing value 2 = 1264+335 = 1599

Figure 4. An example of using RI differences of homologous compounds to predict missing RI values. RO is an alkyl-O substitute.

This method was evaluated on the Schedule 1.A.01 (alkylphosphonofluoridates) of the CWC using OCAD data and compared with the ML predictions. An issue was to choose a reasonable approach to evaluate the XGboost model trained on OCAD, because the molecules that are evaluated should be unseen by the model. To provide a comprehensive evaluation, we moved molecules with particular substituents at phosphorous atoms (methyl, ethyl, propyl) from the training set to the test set; re-trained the models; and calculated the MAE in these separate groups with the ML and non-learning method. Finally, the whole Schedule 1.A.01 was removed and used as a test set. In other words, the model was pre-trained after data division to train and test the set. The results, summarized in Table 3, demonstrate that the proposed non-learning way may significantly enhance the RI predictions and approach to the experimental error.

Table 3. The performance of various approaches on the 1.A.01 CWC Schedule.

Test Set	1D-CNN	XGboost	Increment-Based
Methyl–P	35	39	4.0
Ethyl–P	11	21	1.8
Propyl–P	52	20	3.4
1.A.01	30	27	3.0

We applied the proposed method to predict the missing RI values for some chemicals from CWC for practical use (Table S3).

3.4. Homology Trees

In order to improve the clarity of the incremental method and to visualize the chemical space of the OCAD database, taking in account the presence of many homologous molecules, the homology trees concept can be used. The homologue series formed by C_CH_{2C} for molecules with the molecular formula $C_CH_{2C+Z}N_VO_OS_S$ are referred to as the $ZN_NO_OS_S$ [18] class. The double bond equivalent is determined as DBE = C + 1-H/2 + N/2. Here, H is the number of hydrogens, H = 2C + Z; one can see the relationship between Z and DBE. The idea of the homology trees approach is to organize a chemical space as a graph in which nodes are molecules and nodes are connected with an edge only if corresponding molecules differ by the aliphatic –CH2- unit. The algorithm that was used for the building of the homology tree is presented in Figure 5A. First, we compute all the homology classes (see above). The histogram of such classes for the OCAD database is given in Figure 5B. Then, for the selected class, we can build the homology tree (Figure 5C). In such a tree, the nodes are marked with the value of the retention index for the corresponding molecule; and the edges are marked with the retention time difference. Using the homology tree approach is convenient to see how the value of the retention index varies with the growth of the aliphatic chain (even for a branched chain).

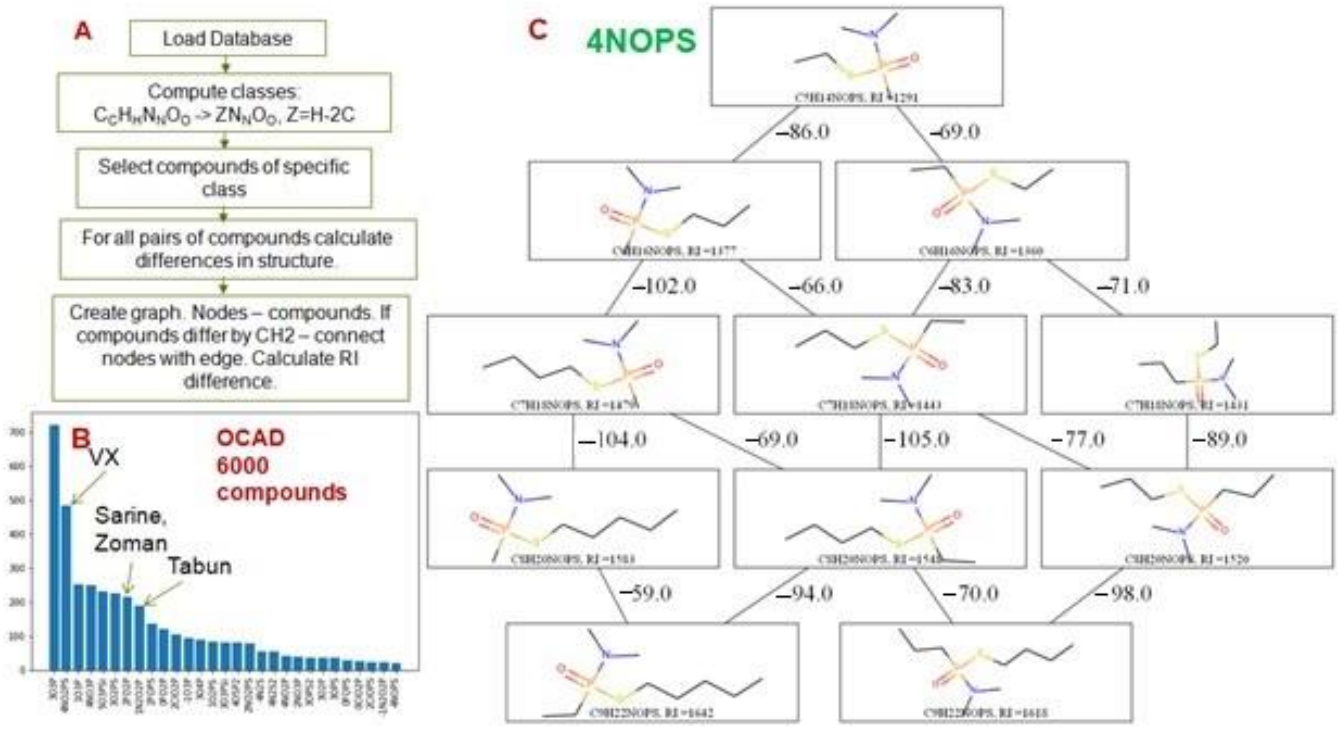

Figure 5. (**A**) an algorithm for computing the homology tree; (**B**) a histogram of the classes of the compounds present in the OCAD database; (**C**) a sample homology tree for the 4NOPS class.

It is worth noting that the homology trees approach can be used for the visualization of any arbitrary chemical space. In order to facilitate it, we have developed a free WEB application available at https://skolmix-homology-trees.anvil.app/ (accessed on 1 September 2022) which allows building a fragmentation tree for any arbitrary dataset. The interface of the application is shown in Figure 6. The format of the dataset is SMILES and some scalar value separated by space. The button "determine classes" calculates the homology classes and corresponding number of compounds. The homology tree will be computed for the selected class. The resulting tree is not always connected. In this case, it is possible to obtain connected subgraphs and plot the homology trees for them.

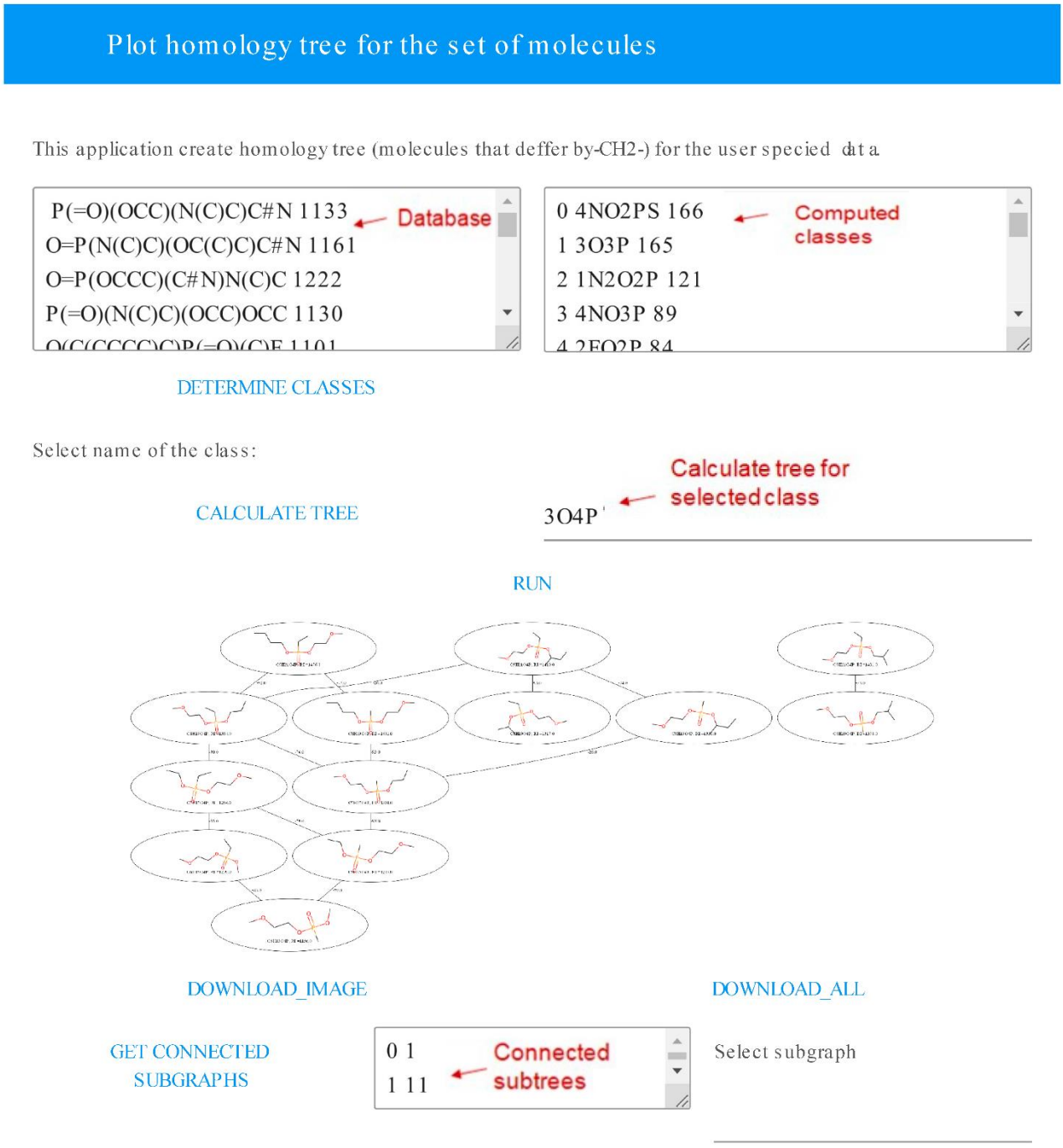

Figure 6. The WEB interface of the developed application for the building of homology trees for an arbitrary dataset.

4. Conclusions

We can conclude that although deep learning methods are extremely powerful and became state-of-the-art, non-learning methods may provide an additional option for narrow tasks; for example, in modeling the molecular properties of structurally similar compounds. It was shown that for special cases, non-learning approaches might demonstrate a predictive error that is comparable with the experimental bias in RI measurements; while even narrow-domain machine learning models have a lack of accuracy. This is extremely important for the analysis guided by CWC; this is because any mistakes and false-positive identifications may have serious political consequences.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/separations9100265/s1, Table S1: Set of hyperparameters for grid search cross-validation: Table S2: Examples of calculated difference of retention index (RI) using molecular pairs matched by RO radical. 1A1, 1A2, 1A3 and selected 2B4 with 1A1 data sets are considered: Table S3: Retention Indices predicted by R-P differences for 1A1 toxic chemicals with available combination of RO and RP, but missing RI data in OCAD library.

Author Contributions: The manuscript was written through the contributions of all authors. All authors have read and agreed to the published version of the manuscript.

Funding: The research was supported by the Russian Scientific Foundation, grant No. 18-79-10127.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Vanninen, P. *Recommended Operating Procedures for Analysis in the Verification of Chemical Disarmament*, 2017th ed.; University of Helsinki: Helsinki, Finland, 2017.
2. Kováts, E. Gas-chromatographische Charakterisierung organischer Verbindungen. Teil 1: Retentionsindices aliphatischer Halogenide, Alkohole, Aldehyde und Ketone. *Helv. Chim. Acta* **1958**, *41*, 1915–1932. [CrossRef]
3. Mesilaakso, D.M. The OPCW Central Analytical Database. In *Chemical Weapons Convention Chemicals Analysis*; Mesilaakso, D.M., Ed.; Finnish Institute for Verification of the Chemical Weapons Convention (VERIFIN), University of Helsinki: Helsinki, Finland, 2005; pp. 133–149. [CrossRef]
4. Matyushin, D.D.; Sholokhova, A.Y.; Buryak, A.K. Deep Learning Driven GC-MS Library Search and Its Application for Metabolomics. *Anal. Chem.* **2020**, *92*, 11818–11825. [CrossRef] [PubMed]
5. Zhokhov, A.K.; Loskutov, A.Y.; Rybal'Chenko, I.V. Methodological Approaches to the Calculation and Prediction of Retention Indices in Capillary Gas Chromatography. *J. Anal. Chem.* **2018**, *73*, 207–220. [CrossRef]
6. Matyushin, D.D.; Sholokhova, A.; Buryak, A.K. A deep convolutional neural network for the estimation of gas chromatographic retention indices. *J. Chromatogr. A* **2019**, *1607*, 460395. [CrossRef] [PubMed]
7. Vrzal, T.; Malečková, M.; Olšovská, J. DeepReI: Deep learning-based gas chromatographic retention index predictor. *Anal. Chim. Acta* **2020**, *1147*, 64–71. [CrossRef] [PubMed]
8. Qu, C.; Schneider, B.I.; Kearsley, A.J.; Keyrouz, W.; Allison, T.C. Predicting Kováts Retention Indices Using Graph Neural Networks. *J. Chromatogr. A* **2021**, *1646*, 462100. [CrossRef] [PubMed]
9. Matyushin, D.D.; Buryak, A.K. Gas Chromatographic Retention Index Prediction Using Multimodal Machine Learning. *IEEE Access* **2020**, *8*, 223140–223155. [CrossRef]
10. Weininger, D.; Weininger, A.; Weininger, J.L. Smiles-Documentation. Available online: https://docs.chemaxon.com/display/docs/SMILES.html (accessed on 29 August 2022).
11. Karpov, P.; Godin, G.; Tetko, I.V. Transformer-CNN: Swiss knife for QSAR modeling and interpretation. *J. Chemin* **2020**, *12*, 17. [CrossRef] [PubMed]
12. Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: A molecular descriptor calculator. *J. Chemin* **2018**, *10*, 4. [CrossRef] [PubMed]
13. Chen, T.Q.; Guestrin, C. Assoc Comp, XGBoost: A Scalable Tree Boosting System, Kdd'16. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [CrossRef]
14. Bonini, P.; Kind, T.; Tsugawa, H.; Barupal, D.K.; Fiehn, O. Retip: Retention Time Prediction for Compound Annotation in Untargeted Metabolomics. *Anal. Chem.* **2020**, *92*, 7515–7522. [CrossRef] [PubMed]
15. Osipenko, S.; Bashkirova, I.; Sosnin, S.; Kovaleva, O.; Fedorov, M.; Nikolaev, E.; Kostyukevich, Y. Machine learning to predict retention time of small molecules in nano-HPLC. *Anal. Bioanal. Chem.* **2020**, *412*, 7767–7776. [CrossRef] [PubMed]
16. Osipenko, S.; Botashev, K.; Nikolaev, E.; Kostyukevich, Y. Transfer learning for small molecule retention predictions. *J. Chromatogr. A* **2021**, *1644*, 462119. [CrossRef] [PubMed]
17. Stein, S.E.; Babushok, V.I.; Brown, A.R.L.; Linstrom, P.J. Estimation of Kováts Retention Indices Using Group Contributions. *J. Chem. Inf. Model.* **2007**, *47*, 975–980. [CrossRef]
18. Marshall, A.G.; Rodgers, R.P. Petroleomics: The Next Grand Challenge for Chemical Analysis. *Acc. Chem. Res.* **2003**, *37*, 53–59. [CrossRef] [PubMed]