Exploring the Chemical Subspace of RPLC: a Data Driven Approach

Denice van Herwerden,[*,†] Alexandros Nikolopoulos,[†] Leon P. Barron,[‡,†] Jake W. O'Brien,[¶,†] Bob W. J. Pirok,[†] Kevin V. Thomas,[¶] and Saer Samanipour[*,†,§,¶]

[†] *Van 't Hoff Institute for Molecular Sciences (HIMS), University of Amsterdam, Amsterdam*

[‡] *MRC Centre for Environment and Health, Environmental Research Group, School of Public Health, Faculty of Medicine, Imperial College London, W12 0BZ, United Kingdom*

[¶] *Queensland Alliance for Environmental Health Sciences (QAEHS), The University of Queensland, Australia*

[§] *UvA Data Science Center, University of Amsterdam, Amsterdam*

E-mail: d.vanherwerden@uva.nl; s.samanipour@uva.nl

Abstract

The chemical space is comprised of a vast number of possible structures, of which an unknown portion comprises the human and environmental exposome. Such samples are frequently analyzed using non-targeted analysis via liquid chromatography (LC) coupled to high-resolution mass spectrometry often employing a reversed phase (RP) column. However, prior to analysis, the contents of these samples are unknown and could be comprised of thousands of known and unknown chemical constituents. Moreover, it is unknown which part of the chemical space is sufficiently retained and eluted using RPLC. Therefore, we present a generic framework that uses a data driven approach to predict whether molecules fall 'inside', 'maybe' inside, or 'outside' of the

RPLC subspace. Firstly, three retention index random forest (RF) regression models were constructed that showed that molecular fingerprints are able to predict RPLC retention behavior. Secondly, these models were used to setup the dataset for building a RPLC RF classification model. The RPLC classification model was able to correctly predict whether a chemical belonged to the RPLC subspace with an accuracy of 92% for the testing set. Finally, applying this model to the 91737 small molecules (i.e., $\leq$1000 Da) in NORMAN SusDat showed that 19.1% fall 'outside' of the RPLC subspace. Knowing which chemicals are outside of the RPLC subspace can assist in reducing potential candidates for library searching and avoid screening for chemicals that will not be present in RPLC data.

Introduction

The chemical space refers to a collection of all possible organic structures - for example, the GBD-17 database includes 116 billion possible organic molecules with a maximum of 17 atoms, which is only a fraction of the chemical space.[1–8] Increasing the number of atoms only drastically increases these numbers and shows how vast the chemical space actually is. Even though these are possible structures, not all of them are likely to be present in the human and environmental exposome.[8] When evaluating the exposome, the main difficulty is that the contents of the samples taken are unknown prior to analysis and may comprise of thousands of both known and unknown constituents, particularly for small molecules (i.e., molecular weight $\leq$ 1000 Da).[9–16] A frequently used approach for analyzing such samples is non-targeted analysis (NTA) via liquid chromatography (LC) coupled to high-resolution mass spectrometry (HRMS), for which a reversed phase (RP) LC selectivity is often used.[8] However, it is not yet known what part of the chemical space is covered by RPLC. The knowledge of the covered subspace also contains crucial information on chemicals that might not be visible in the final data even though they were present in the sample.[3]

Knowing what is separable with RPLC can have an improved outcome for both NTA and suspect screening. For NTA, the aim is to identify as much as possible of the potentially thousands of chemicals present in samples coming from, for example, biological or environmental backgrounds. Eliminating the potential candidates that fall outside of the chemical subspace of the selectivity (e.g., RPLC), reduces the number of false positive identifications. On the other hand, suspect screening is also a frequently used approach, where samples are screened for lists or even databases of compounds. Defining the subspace of a selectivity can reduce the number of potential candidates in these compound lists, reducing the computational time required and the false positive matches with chemicals that cannot possibly be measured with this technique.

Separation data is usually limited to the mere assessment of whether the analyte retention time could fit in the range of the candidate's chemical class.[17–20] To take better advantage of the LC data, retention times are required to be initially converted to retention indices (r_i), since the former are significantly influenced by the chromatography conditions, such as temperature, mobile phase composition, and gradients.[20,21] On the other hand, r_i values provide a robust and highly reproducible way to express retention in liquid chromatography.[20] High reproducibility makes inter-laboratory results comparable, enabling both m/z and r_i comparison with a reference and resulting in more confident suspect shortlisting.

As for any r_i system, different chromatography conditions should have negligible influence on the r_i value of the analytes, suggesting that there is a correlation between the r_i values and structural properties, expressed as molecular descriptors. This is the main principle used by the quantitative structure-retention relationship (QSRR) based models,[22] enabling the construction of QSRR models that either use all or a selection of descriptors to predict r_i values.[23–26] However, difficulties arise when calculating descriptors due to convergence issues related to calculation time-out or local minima.[25–27] Moreover, descriptors can often be difficult to interpret, since they contain mathematical representations of the molecular structure. Alternatively, molecular fingerprints directly encode the molecular structure, making them more descriptive/understandable to interpret in relation to the chemical and do not require structural optimization (i.e., only uses 2D structural information), making them a potential alternative to descriptors.

In this paper, we present a data driven approach for a generic framework that enables quick screening of the RPLC chemical space, assuming that the molecules are in solution and can be injected into a system. A set of regression and classification models were built to assess whether a structure can theoretically be analyzed via RPLC. To build the RPLC classification model, firstly, we show the potential of using fingerprints for the prediction of r_i values for three retention index series, confirming that molecular fingerprints contain information on RPLC retention behavior. Three commonly used scales, namely: the n-alkylamide system, containing the n-alkylamide homologous series from n-propanamide to n-tetradecanamide (C3-C14)[28], the r_i system developed by Aalizadeh et al. from the University of Athens referred to as UoA, comprising of 18 reference compounds that were computationally selected in order to achieve a broad and reliable r_i reference system[29], and the cocamide diethanolamine homologous series that is comprised of C(n = 0-23)-DEA chemicals[30] were employed for our model building. Secondly, we show the performance of the RPLC classification model and apply the model on a set of 91737 small molecules (i.e., molecular weight $\leq$ 1000 Da) from the NORMAN substance database (SusDat).

Experimental Section

Overall Workflow

The overall workflow for this work can be found in figure 1 and the details are explained in the following sections. In brief, a total of four random forest (RF) models were built, of which three were r_i RF regression models (Figure 1**A**) and the fourth a RPLC RF classification model (Figure 1**B**). For building these models, a type of molecular fingerprint needed to be selected and the dataset obtained before model optimization and performance testing (Figure 1**C**). These models were used for evaluating the potential of using molecular fingerprints for prediction of retention behavior in RPLC and for setting up two of the classes for the fourth RF classification model. The latter refers to the 'inside' and 'maybe' inside class. Here, the 'maybe' class represents the chemicals that are poorly retained (i.e., close to t_0) or require relatively high amounts of organic modifier to elute. All chemicals in between the 'maybe' regions are classified as 'inside'. For the RPLC classification model, a dataset with chemicals that were 'inside', 'maybe' inside, and 'outside' of the RPLC subspace was constructed (Figure 1**B**). Finally, the application of the RPLC classification model was showcased by applying it on the NORMAN SusDat database, which is a collection of expert curated environmentally relevant chemicals that have been actively used for screening of complex samples. All datasets for constructing the models and the NORMAN SusDat database can be found on Figshare.[31]

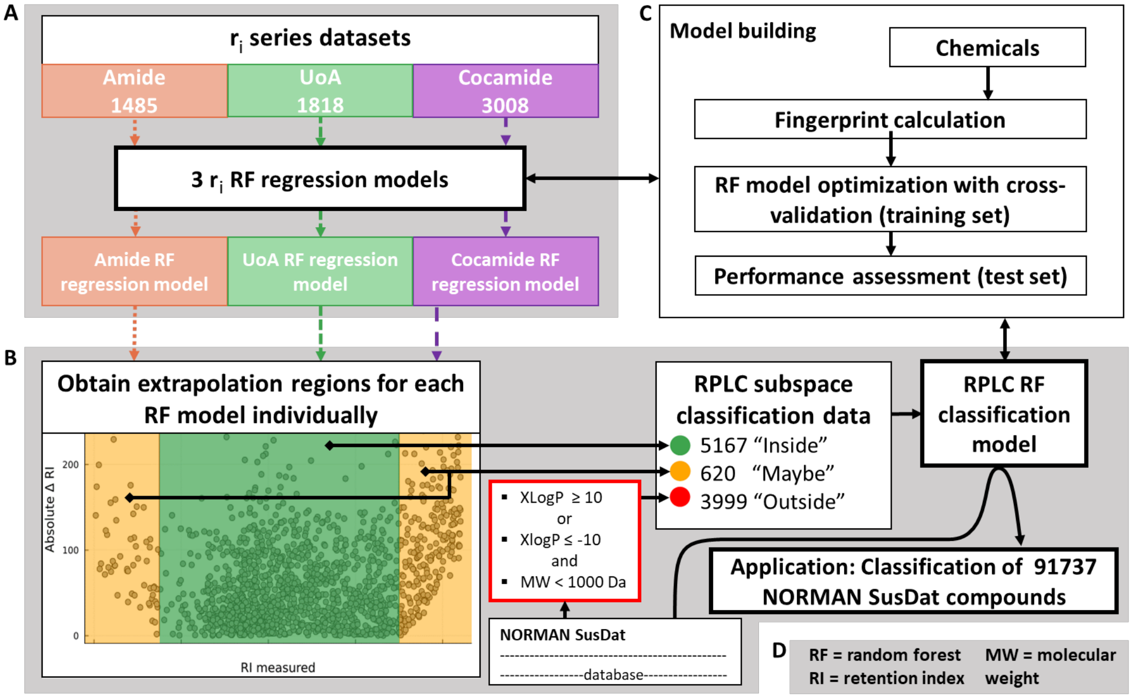


Figure 1: Workflow for construction of the RPLC classification model, comprising of the construction of three r_i RF regression models (**A**) and the construction the RPLC dataset for the RPLC RF classification model, which was applied to NORMAN SusDat(**B**). Finally **C** shows the model setup and **D** contains an overview of the abbreviations.

Fingerprint Calculations

The RF models were built using a combination of two different fingerprint series as inputs, which included the AtomPairs2DFingerprintCount (2DAPC) and PubChem fingerprints,[32] obtained through PaDEL.[33] The 2DAPC fingerprints counted the number of times two atoms were present with a certain distance between themselves. For example, the molecule with the SMILES 'NC(CC)CN' contains two times a distance of 3 between a C and N atom (i.e., C-x-x-N in the 2D molecular structure). The distances included ranges from 1 to 10 and the elements considered were C, N, O, Cl, I, Br, F, P, S, Si, B, and X, where X represents all halogens, yielding a total of 780 2DAPC fingerprints. As for the PubChem fingerprints, only the portion of fingerprints containing ring information was used (i.e., PubChem fingerprint 115 - 262). These fingerprints were converted and reduced to a total of 10 additional variables, which were the number of rings with a size of 3, 4, 5, 6, 7, 8, 9, 10, the number of aromatic rings, and the number of hetero-aromatic rings. Since the PubChem fingerprints are binary, there were multiple columns describing the same information but only differing in the number of a ring of a certain size. For example, for a ring size of 3, there were 2 fingerprints, namely PubChem fingerprint 115 and 122, which were described as more than 1 ring with a size of 3 or more than 2 rings with a size of 3, respectively. In case a molecule contained 2 rings with a size of 3, the PubChem fingerprints 115 would be 0 and 122 would be 1, which was converted to a single variable for our model containing the number of rings with a size of 3, meaning that this variable would be equal to 2 for this example case. An overview of which PubChem fingerprints were used for each of the 10 reduced PubChem variables can be found in table S2.

Retention Index Random Forest Regression Models

To show that fingerprints can be used to describe retention behavior in RPLC and for setting up the dataset for the RPLC classification model, random forest (RF) regression models were built using three different retention index series (Figure 1**A**). The three series used for this, were the amide[28], University of Athens (UoA)[29], and cocamide series.[30] For each of the series, the measured r_i were obtained from their respective articles, yielding 1485, 1818, and 3008 unique chemicals with measured r_i values for the amide, UoA, and cocamide series, respectively. For all chemicals, the 2DAPC and PubChem fingerprints were calculated according to Section 'Fingerprint Calculations'. For each r_i series, data was split into a training and test set, at random, with a ratio of 0.85:0.15, ensuring similar coverage of the r_i range in both sets. The test set was only used for testing and thus never used for training. For optimization of the RF regression models, the training set was used with a 0.8:0.2 split for training and cross-validation, respectively. This ratio of split has been shown to be effective in such data sets.[25,26,34,35] The RF regression models used a third of the features (i.e., 264) for training each tree. The parameters that were optimized were the minimum number of samples per leaf and the number of trees. The minimum number of samples per leaf tested were 4, 6, 8, 10, 15, and 20. The tested number of trees were 50, 100, 150, 200, 250, 300, 350, 400, 500, 600, 700, 800, 900, and 1000. In addition, the random state for splitting the cross-validation set and selection of the features in the RF models for each tree was also varied with values of 1, 2, and 3. The accuracy of the cross-validation set for each possible combination of the minimum number of samples per leaf, number of trees, and random state was used for the optimization of the RF models. After obtaining the optimized models for the amide, UoA, and cocamide series, the applicability domains were assessed according to Section 'Applicability Domain Calculations'. Finally, for each r_i series, the optimized model and applicability domain assessment were applied on the test set to evaluate the performance of the model on unseen data.

RPLC Random Forest Classifier

The dataset for building the RPLC classifier model was comprised of three classes: 'inside', 'maybe', and 'outside' the RPLC subspace (Figure 1**B**). The 'outside' chemicals were obtained from the NORMAN SusDat database based on their extreme XLogP values. Here, the XLogP was chosen rather than the logD due to the fact that it is easier to predict, more stable, and more accurate.[36] For the 'outside' case, a total of 3999 compounds with a XLogP value above 10 or below -10 and with a molecular weight below 1000 Da were obtained. As for the 'inside' and 'maybe' chemicals, these were obtained from the experimentally defined r_i values by the three r_i series. For each of the series, the absolute difference between the predicted and measured r_i (i.e., the residuals) versus the measured r_i values were plotted and the regions of extrapolation were identified. These regions were obtained based on the increasing residuals that were caused by the inherent over estimation and under estimation of a RF regression model, which are associated with either extremely low or extremely high

r_i values, respectively. These regions correspond to chemicals that elute close to t_0 or are very difficult to elute from the column (i.e., require a relatively high percentage of organic modifier). The chemicals with a measured r_i in these extrapolation regions were labeled as ‘maybe’ and the remaining chemicals were labeled as ‘inside’ the RPLC subspace. This yielded a total of 620 ‘maybe’ and 5167 ‘inside’ compounds. Whenever a chemical was found in multiple classes (i.e., it was present in mulitple datasets of the r_i models), it was removed from the lower ranking RPLC classes and kept in the highest ranking RPLC class (i.e., ‘inside’ > ‘maybe’ > ‘outside’ RPLC class rank). For example, if a chemical was found in the ‘maybe’ region for UoA and in the ‘inside’ for Cocamide, it would be classified as ‘inside’. More details on the division between the ‘inside’ and ‘maybe’ classification can be found in Section ‘RPLC Classification Model’ as these are based on the results of the three RF regression models.

The dataset described above was used for building the RPLC classifier model with a training set/test set split of 0.85:0.15, ensuring equal distribution of each class in both sets. The optimized RF classifier model was obtained using the same approach as for the RF regression models (see Section ‘Retention Index Random Forest Regression Models’). For this model, the applicability domain was also obtained as described below. Finally, the optimized RPLC classification model and applicability domain assessment was applied to the test set and the performance was evaluated.

RPLC Space Prediction for NORMAN SusDat

To showcase the model’s potential, it was applied to the NORMAN SusDat database.[5] For this, the 2DAPC and reduced PubChem fingerprints for a total of 91737 chemicals with a molecular weight below 1000 Da from SusDat were calculated. These fingerprints were then used to calculate the leverage of each chemical with the RPLC classifier training set, as explained in the next section ‘Applicability Domain Calculations’, and to apply the RPLC classifier model to each of the SusDat chemicals. To visualize the coverage of each class (i.e., ‘inside’, ‘maybe’, and ‘outside’ the RPLC subspace), the molecular weight was plotted against the XLogP, which were obtained from the descriptor calculations of PaDEL.

Applicability Domain Calculations

Applicability domain calculations were used to assess whether the training data, used in the random forest models, sufficiently covered the variable space for new chemicals on which the models need to be applied.[25,37] This was done through leverage calculations of a chemical with the entire training set, yielding a distance of that chemical to the training set. Equation 1 shows how the leverage is calculated, where X is the training data matrix and x_i is the sample vector, both containing the 2DAPC and reduced PubChem fingerprints for our models. To set a threshold for this, the leverage was calculated for all training samples with the entire training set of a model, yielding values between 0 and 1. Then, a leverage threshold was obtained that covered 95% of the training data. If a chemical, compared to the training set of the model in question, had a value lower than the leverage threshold, the compound was within the applicability domain, and, if the value was above the leverage threshold, the results should be taken with care as the training data might not be sufficiently describing the variable space for the new compound.

$$l_{ii} = x_i(X^T X)^{-1} x_i \tag{1}$$

Calculations and Code Availability

The calculations and development of the models were executed on a personal computer with 12 CPUs and 32 GB of RAM, using Windows 10. The r_i regression and RPLC classification models were developed and evaluated with the Julia programming language (v1.6). The code for using the r_i regression models and RPLC space prediction model is available at: `https://bitbucket.org/Denice_van_Herwerden/riprediction/src/main/`. This Julia package contains functions for obtaining the required 2DAPC and reduced PubChem fingerprints and for using the r_i regression models and RPLC sub space classification model.

Results and discussion

Retention Index Random Forest Regression Models

All three r_i regression models obtained an accuracy of 81% for the training set and, for the test set. The amide, UoA, and cocamide models had an accuracy of 68%, 70%, and 67%, respectively. The r_i regression models were built and optimized for the amide, UoA and cocamide series. Grid optimization of each of these models showed that the number of trees did not influence the performance of the model (Figures S1, S2, and S3). Therefore, to keep the model light, 200 trees were selected. As for the minimum number of samples per leaf, 8 was found to be the optimum, based on the training and cross-validation accuracy. When evaluating the predicted versus the measured r_i values for these models a trend of over prediction for lower r_i values and under prediction of higher r_i values was found(Figures S4, S6, and S8), corresponding to the regions where the RF regression models were extrapolating. These regions were used for establishing the 'maybe' areas for the RPLC classification dataset.

Most compounds (i.e., 88.5%) in our test set appeared to be within the applicability domain of each model. To obtain the applicability domains of these models, a 95% leverage threshold of 0.189 for amide, 0.652 for UoA, and 0.424 for cocamide was found for the training sets. For the training set the leverage values range between 0 and 1, meaning that the lower threshold for the amide model showed how similar most of the amide compounds were to each other, while for the UoA and cocamide models, the higher thresholds corresponded with the larger variety of chemical structures found in the dataset. When the leverage calculations were applied on the test sets for these models, a total of 22, 34, and 54 compounds were found to be outside of the applicability domain for the amide, UoA, and cocamide r_i models, respectively. This does not necessarily mean that the predicted outcome for these cases was wrong, as can be seen in figures S4, S6, and S8. Here, most chemicals outside the applicability domain still follow the trend of the other data points. However, the outcome should be taken with care as the model might insufficiently cover the chemical space for a new compound in question, especially for leverage values > 1. It should be noted that the largest training set leverage value obtained from our applicability domain calculations was 1.

The cocamide RF regression model used the most fingerprints for the prediction of the r_i indices (i.e., 215 fingerprints), while the UoA and amide r_i models used 165 and 61, respectively. The low number of fingerprints used for amide was not surprising due to the fact that the compounds in this r_i series are only comprised of C, H, N, and O. Hence, the amide r_i model only used the 2DAPC fingerprint counts with a certain distance between C, N, and O atoms. At first sight, this was also noticeable when comparing the top 20 most important fingerprints for the three r_i models (S3). The most contributing fingerprints for the amide r_i model were the distances 1 till 7 between two C atoms with importance ranging between 27% and 4%. As for the UoA r_i model, C-Cl and C-X distance begin to contribute more to the model and the most important fingerprint (i.e., distance 7 between C-C) only contributes 9.6%, having an overall more divided importance between a larger group of contributing features than the amide model. Finally, a similar trend was also observed for the cocamide model, except that the C-X distances start to play a more important role than the C-Cl distances, which could be explained by the higher number of halogens present in the compounds from the cocamide dataset. This variability in important features used in each

r_i regression model shows that different structures may be better captured by one r_i model vs another, due to the diversity of training set in terms of chemical structures. This, also, further indicates the need for a more generic model incorporating the information from all three r_i models.

Overall, these models show that a combination of the 2DAPC fingerprints and the reduced PubChem fingerprints can be used to predict r_i values. All three models performed almost equally well with negligible deviations for the training set accuracy. However, depending on the chemicals for which r_i would be predicted, it is advised to evaluate which model would be most suitable based on the leverage applicability domain calculations.

RPLC Classification Model

To build the RPLC classification model, it was assumed that the chemicals are in solution and that the chemicals can be injected into a system. Additionally, the model focuses on whether an analyte could be analyzed with RPLC regardless of experimental parameters or sample pretreatment. The dataset for this was comprised of 5167 'inside', 620 'maybe' inside, and 3999 'outside' chemicals for the RPLC subspaces. The 'outside' cases were obtained from NORMAN SusDat with extreme XLogP values, while the 'inside' and 'maybe' cases came from the three r_i regression models. In figures S10, S11, and S12 the extrapolation limits for each of the models are defined. For r_i range for the 'inside' RPLC subspace for the amide, UoA, and cocamide series were 350-900, 100-900, and 250-1300, respectively. All compounds that had a higher or lower r_i value for the corresponding range of the model it was coming from, were classified as 'maybe' inside the RPLC subspace, due to the fact that these chemicals either elute close to t_0 or require high percentages of organic eluent to be eluted.

The final optimized classification model resulted in an accuracy of 94% and 92% for the training and test set, respectively (Figures 2, and S15). In this case 200 trees and 8 minimum samples per leaf was found to be the optimum for the model (Figure S13). For the training and test set, 90.8% and 87.7% of the 'inside' and 'maybe' cases were correctly classified, 7.4% and 9.3% of the 'inside' and 'maybe' cases were wrongly classified as a 'maybe' or 'inside' case, respectively, and 1.7% and 3.0% of the 'inside' and 'maybe' cases were wrongly classified as 'outside'. For the 'outside' cases, 0.7% and 1.5% of the cases were wrongly classified as an 'inside' or 'maybe' case and 99.3% and 98.5% of the cases was correctly classified as an 'outside' case for the training and test set, respectively. Overall, considering that the wrongly classified 'inside' and 'maybe' cases as 'maybe' and 'inside', respectively, still are considered part of the RPLC subspace, the performance of the model was very good with only 2.4% of all cases being wrongly classified as 'inside' or 'maybe' while being an 'outside' or vise versa for the test set.

As for the applicability domain of the RPLC classification model, the 95% leverage threshold of the training set was 0.209 (Figure S14). In total, 102 compounds from the test set (i.e., 6.9%) had a leverage with the training set that was higher than 0.209, of which 31 cases had leverage values above 1. Out of these 102 cases only 10 were wrongly classified and had leverage values ranging between 0.209 to the most extreme (i.e., 809.255), showing that in this case higher leverage values did not necessarily mean that the model would have a higher error. However, it should be noted that cases with a very large leverage should be considered with extra care, as they may have a higher level of uncertainty.

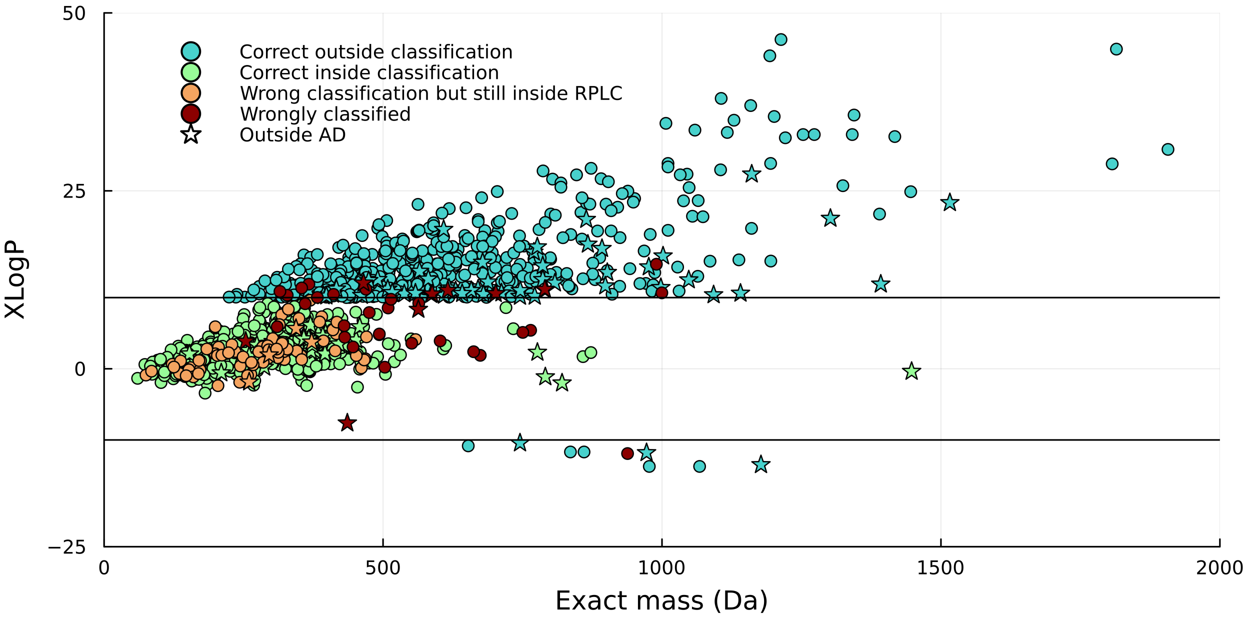


Figure 2: XLogP values versus the molecular weight for the RPLC classification test set. In blue are the correctly classified 'outside' cases, in green are the correctly classified 'inside' and 'maybe' cases, in orange are the wrongly classified 'inside' cases as 'maybe' and vice versa, in red the wrongly classified 'inside' and 'maybe' cases as 'outside' and the wrongly classified 'outside' cases as 'inside'. The star markers show the compounds that were outside the 95% applicability domain of the RPLC classification training set

A total of 280 features were contributing to the RPLC classification model. This is more than for each of the three r_i regression models, which was expected due to the higher variety in chemical structures used in the RPLC classification model. The 20 most contributing features are mainly described by ring related features and distances between combinations C, N, and O atoms. A previous version of the model that was tested, using only the 2DAPC fingerprints, frequently wrongly classified 'inside' as 'outside' due to the high degree of cyclicity in the chemical structures (e.g., InChIKey: IUKLSMSEHKDIIP-BZMYINFQSA-N). Hence, the addition of the reduced PubChem fingerprints better captures these chemical properties. As a result, the number or rings with a size of 6, the minimum number of aromatic rings, and the number of rings with a size of 5 were also part of the top 20 most contributing features.

In total, considering the extreme misclassifications, 9 out of 599 'outside' chemicals were wrongly classified as 'inside' or 'maybe' inside the RPLC subspace and 14 out of the 767 'inside' and 12 out of the 102 'maybe' cases were classified as 'outside' the RPLC subspace. Two of the nine wrongly classified 'outside' cases were organic complexes that, in the mobile phase, would be analyzed as multiple smaller molecules (e.g., Gadopentetic acid dimeglumine salt). Also, another case was a surfactant containing a positive and negative charge (i.e., 4-Dodecyl-2-[(2-nitrophenyl)azo]phenol). This case was a chemical that falls 'outside' of the RPLC space due to its predicted XLogP value of 10.452. However, the charges on this molecule would make it difficult to calculate this value accurately. Lexidronam was one of the 'maybe' cases that was classified as 'outside', due to a large leverage value of 26.0 and the fact that it elutes at t_0 (i.e., amide scale r_i of 206 versus urea $r_i = 200$), indicating the need for special gradients to be able to retain such a chemical. As for the 'inside' cases that were wrongly classified as 'outside', generally larger, branched (e.g., SCHEMBL312614), or hydrolyzing (e.g., Bis[2-(perfluorohexyl)ethyl] Phosphate, respectively) chemicals showed higher likelihood of such misclassifications. Again these are structures that may require very specific adjustment of experimental condition (e.g., pH of mobile phase) to fit them within the RPLC analyzable chemical subspace.

Overall, our RPLC classification model was highly successful in identifying the chemical structures that are easily analyzable via RPLC (i.e., 'inside' cases) as well as the 'maybe' and 'outside' cases. The classification model used a combination of similar molecular fingerprints as those used by the three r_i models, taking advantage of all the structural information.

NORMAN SusDat Chemical Space Prediction

Finally, the RPLC classification model was applied to a set of small molecules (i.e., molecular weight < 1000) from the NORMAN SusDat database. In total, 80503 chemicals were within the applicability domain with leverage values ≤ 0.209, 6570 compounds had leverage values between 0.209 and 1, and 4664 compounds had even larger leverages. This showed that the

RPLC classification model was suitable for a large variety, 87.8%, of compounds present in SusDat. The model predicted that 79.0% of the compounds would fit 'inside' the RPLC subspace, 2.0% was 'maybe' in this space, and 19.1% was 'outside' of the RPLC subspace. Examples of molecules classified as 'inside', 'maybe', and 'outside' were carbamazepine, sudan I, and coronene, respectively. When comparing the relationship between XlogP and r_i, it is clearly observable that these parameters, even though relatively linear, are insufficient to determine if a chemical fits the RPLC subspace, figure 3. In figures S16,S17, and S18, the XlogP values of the chemicals with the same r_i range vary between -10 to +10 units.

Using the developed classification models implies that for screening RPLC samples against databases such as SusDat, 1/5 of the overall time can be saved, which becomes even more significant when applying it to larger sample sets. Additionally, this will result in higher confidence identifications when performing database matching for an RPLC NTA method with SusDat, by reducing the overall number of potential candidates and thus false positive identifications.

The amide r_i model is the least suited scale based on its applicability domain coverage since only 44500 (i.e., 48.5%) chemicals fell within the applicability domain. For the chemicals that were outside the applicability domain, 18988 had a leverage value between 0.189 and 1 (i.e., similar to the full training set) and 28249 had an even higher leverage value. As for the UoA and cocamide r_i models, 71022 (i.e., 77.4%) and 74252 (i.e., 80.9%) compounds were within the applicability domain. For the UoA model, 3421 and 17294 chemicals had a leverage value below and above 1, respectively, and the cocamide model had 5947 chemicals with a leverage value below 1 and 11538 chemicals with higher leverage values. Figures S16, S17, and S18 show the coverage of the 'inside', 'maybe', and 'outside' RPLC classes in terms of the XLogP values versus the predicted r_i values for the amide, UoA, and cocamide series. As expected the chemicals classified as 'maybe' inside RPLC are mainly clustering around the lower and higher r_i values. While the chemicals classified as 'outside' the RPLC space span the entire r_i range for each of the three r_i series, suggesting that r_i prediction would also be insufficient to define the boundaries of the RPLC subspace.

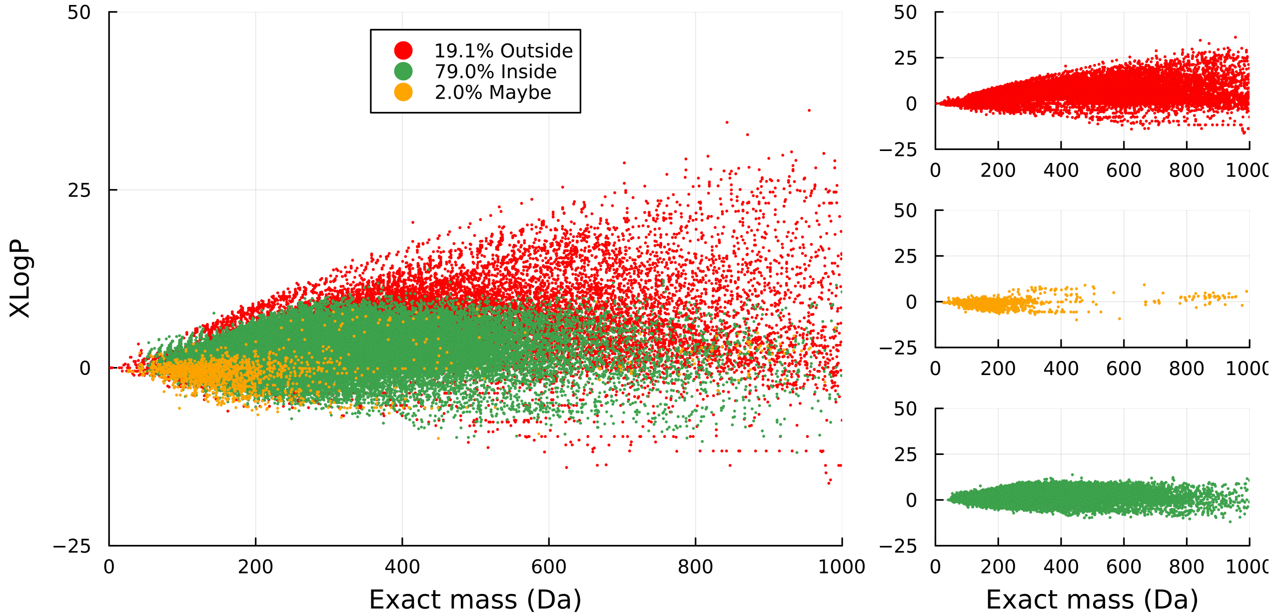


Figure 3: XLogP values versus the molecular weight for the NORMAN SusDat database compounds with a molecular weight below 1000 Da. In red, orange, and green are the compounds that were classified as 'outside', 'maybe', and 'inside' the RPLC chemical space, respectively. The subplots on the left show the coverage of the individual classes.

Potentials and Limitations

Overall, we developed four models for exploration of the RPLC subspace. The r_i regression models showed that fingerprints can be used for describing RPLC retention indices. Consequently, these fingerprints were used for RPLC classification model building. This model was able to predict whether chemicals were 'inside', 'maybe' inside, or 'outside' of RPLC chemical subspace with an accuracy of 92% on the test set. Applying the RPLC classification model on NORMAN SusDat showed that 19.1% of the compounds were classified as 'outside' the RPLC subspace. This means that, when performing identification on NTA RPLC samples, candidates classified as ‘outside’ compounds are unlikely to be the true structure of the chemical and can be removed to reduce the number of false positive identifications. In terms of suspect screening, it can save computational time since the ‘outside’ chemicals fall ‘outside’ of the RPLC subspace and thus should not be screened for. Additionally, 87.8% of NORMAN SusDat was within the applicability domain of the RPLC classifier, showing good coverage of a variety of compounds. The RPLC classification model also showed that the XLogP or r_i values alone are not sufficient to define the RPLC subspace.

The RPLC classification model overall did have more difficulties with regard to more bulky and branched or surfactant-like chemicals. Additionally, the model was not able to properly predict the RPLC subspace class of chemicals that are organic complexes, due to the fact that in solution those are dissociated into multiple individual structures. The latter is not a major limitation for the model itself, since, using expert knowledge, they can be easily identified. Generally, as knowledge on analyzable chemicals with RPLC grows, the model could easily be rebuilt and expanded for the range of analytes. Ideally, when sufficient data becomes available, selectivity classification models could be constructed for other selectivities (e.g., HILIC). This allows for further understanding of what part of the chemical space is actually covered by the selectivities used in NTA and what we are missing.

Moreover, the RPLC classification model uses a data driven approach and is intended for quick screening of the RPLC chemical space. The model assumes that compounds are analyzable with RPLC regardless of the chemicals solubility, experimental parameters, or pretreatment steps taken. This means that it cannot be assumed that chemicals ‘inside‘ the RPLC space will be analyzable with every RPLC method. Here, the method subspace plays a major role when looking at what individual NTA methods can cover, becoming an even more complex issue due to the fact that sample pretreatment, gradient program’s, and RP column selectivities play a large influence on this. Defining the method chemical space would be the next step in understanding what part of the vast chemical space we are covering and, more importantly, excluding with our current NTA methods.

Acknowledgement

The authors thank the Environmental Monitoring and Computational Mass Spectrometry (www.emcms.info) group for their insights and feedback. The Queensland Alliance for Environmental Health Sciences. Finally, the University of Queensland gratefully acknowledges the financial support from the Queensland Department of Health. J.W.O is the recipient of an NHMRC Emerging Leadership Fellowship (EL1 2009209).

Supporting Information Available

Overview of performance for using different types of molecular fingerprints, composition of reduced PubChem fingerprints, optimization, prediction, leverage, and feature importance results for the 3 RF regression models and the RPLC classification model, and the RPLC classification of NORMAN SusDat visualized by plotting the XLogP values versus the predicted r_i values for the three r_i regression models.

Author Information

Corresponding Author:
Saer Samanipour
Van 't hoff institute for molecular sciences (HIMS),
University of Amsterdam,
the Netherlands
Email: s.samanipour@uva.nl

Denice van Herwerden
Van ’t hoff institute for molecular sciences (HIMS),
University of Amsterdam,
the Netherlands
Email: d.vanherwerden@uva.nl

References

(1) Ruddigkeit, L.; Deursen, R. V.; Blum, L. C.; Reymond, J. L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *Journal of Chemical Information and Modeling* **2012**, *52*, 2864–2875.

(2) Reymond, J. L. The Chemical Space Project. *Accounts of Chemical Research* **2015**, *48*, 722–730.

(3) Black, G. et al. Exploring chemical space in non-targeted analysis: a proposed ChemSpace tool. *Analytical and Bioanalytical Chemistry* **2023**, *415*, 35–44.

(4) Williams, A. J.; Grulke, C. M.; Edwards, J.; McEachran, A. D.; Mansouri, K.; Baker, N. C.; Patlewicz, G.; Shah, I.; Wambaugh, J. F.; Judson, R. S.; Richard, A. M. The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. *J. Cheminf.* **2017**, *9*, 61.

(5) NORMAN SusDat. `https://www.norman-network.com/nds/susdat/`.

(6) Lipinski, C. A.; Dominy, B. W.; Feeney, P. J. drug delivery reviews Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. 1997.

(7) Bohacek, R. S.; McMartin, C.; Guida, W. C. The art and practice of structure-based drug design: A molecular modeling perspective. 1996.

(8) Hulleman, T.; Turkina, V.; O, J. W.; Chojnacka, A.; Thomas, K. V.; Samanipour, S. Critical assessment of covered chemical space with LC-HRMS non-targeted analysis.

(9) Schulze, B.; Jeon, Y.; Kaserzon, S.; Heffernan, A. L.; Dewapriya, P.; O'Brien, J.; Gomez Ramos, M. J.; Ghorbani Gorji, S.; Mueller, J. F.; Thomas, K. V.; Samanipour, S. An assessment of quality assurance/quality control efforts in high resolution mass spec-

trometry non-target workflows for analysis of environmental samples. *TrAC Trends in Analytical Chemistry* **2020**, *133*, 116063.

(10) Schymanski, E. L. et al. Non-target screening with high-resolution mass spectrometry: Critical review using a collaborative trial on water analysis. *Anal. Bioanal.Chem.* **2015**, *407*, 6237–6255.

(11) Werner, E.; Heilier, J.-F.; Ducruix, C.; Ezan, E.; Junot, C.; Tabet, J.-C. Mass spectrometry for the identification of the discriminating signals from metabolomics: Current status and future trends. *J. Chromatogr. B* **2008**, *871*, 143 – 163, Hyphenated Techniques for Global Metabolite Profiling.

(12) Samanipour, S.; Reid, M. J.; Bæk, K.; Thomas, K. V. Combining a Deconvolution and a Universal Library Search Algorithm for the Nontarget Analysis of Data-Independent Acquisition Mode Liquid Chromatography-High-Resolution Mass Spectrometry Results. *Environ. Sci. Technol.* **2018**, *52*, 4694–4701.

(13) Samanipour, S.; Kaserzon, S.; Vijayasarathy, S.; Jiang, H.; Choi, P.; Reid, M. J.; Mueller, J. F.; Thomas, K. V. Machine learning combined with non-targeted LC-HRMS analysis for a risk warning system of chemical hazards in drinking water: A proof of concept. *Talanta* **2019**, *195*, 426 – 432.

(14) Brack, W.; Hollender, J.; de Alda, M. L.; Müller, C.; Schulze, T.; Schymanski, E.; Slobodnik, J.; Krauss, M. High-resolution mass spectrometry to complement monitoring and track emerging chemicals and pollution trends in European water resources. *Environ. Sci. Eur.* **2019**, *31*, 62.

(15) Minkus, S.; Bieber, S.; Letzel, T. Spotlight on mass spectrometric non-target screening analysis: Advanced data processing methods recently communicated for extracting, prioritizing and quantifying features. *Analytical Science Advances* **2022**, *3*, 103–112.

(16) van Herwerden, D.; O'Brien, J. W.; Choi, P. M.; Thomas, K. V.; Schoenmakers, P. J.; Samanipour, S. Naive Bayes classification model for isotopologue detection in LC-HRMS data. *Chemometrics and Intelligent Laboratory Systems* **2022**, *223*, 104515.

(17) Gertsman, I.; Barshop, B. A. Promises and pitfalls of untargeted metabolomics. *Journal of Inherited Metabolic Disease* **2018**, *41*, 355–366.

(18) Watson, D. G. A ROUGH GUIDE TO METABOLITE IDENTIFICATION USING HIGH RESOLUTION LIQUID CHROMATOGRAPHY MASS SPECTROMETRY IN METABOLOMIC PROFILING IN METAZOANS. *Computational and Structural Biotechnology Journal* **2013**, *4*, e201301005.

(19) Fedorova, E. S.; Matyushin, D. D.; Plyushchenko, I. V.; Stavrianidi, A. N.; Buryak, A. K. Deep learning for retention time prediction in reversed-phase liquid chromatography. *Journal of Chromatography A* **2022**, *1664*.

(20) Rigano, F.; Arigò, A.; Oteri, M.; Tella, R. L.; Dugo, P.; Mondello, L. The retention index approach in liquid chromatography: An historical review and recent advances. *Journal of Chromatography A* **2021**, *1640*.

(21) Smith, R. M. *Chapter 3 Retention index scales used in high-performance liquid chromatography*; 1995; pp 93–144.

(22) Aalizadeh, R.; Thomaidis, N. S.; Bletsou, A. A.; Gago-Ferrero, P. Quantitative Structure–Retention Relationship Models To Support Nontarget High-Resolution Mass Spectrometric Screening of Emerging Contaminants in Environmental Samples. *Journal of Chemical Information and Modeling* **2016**, *56*, 1384–1398.

(23) Lamparczyk, H.; Radecki, A. The role of electric interactions in the retention index concept; Implications in quantitative structure-retention studies. *Chromatographia* **1984**, *18*, 615–618.

(24) Farkas, O.; Zenkevich, I. G.; Stout, F.; Kalivas, J. H.; Héberger, K. Prediction of retention indices for identification of fatty acid methyl esters. *Journal of Chromatography A* **2008**, *1198-1199*, 188–195.

(25) Samanipour, S.; O'Brien, J. W.; Reid, M. J.; Thomas, K. V.; Praetorius, A. From Molecular Descriptors to Intrinsic Fish Toxicity of Chemicals: An Alternative Approach to Chemical Prioritization. *Environmental Science and Technology* **2022**,

(26) Boelrijk, J.; van Herwerden, D.; Ensing, B.; Forré, P.; Samanipour, S. Predicting RPLC retention indices of structurally unknown chemicals from mass spectrometry data. *Journal of Cheminformatics* **2023**, *15*, 1–12.

(27) Alygizakis, N.; Konstantakos, V.; Bouziotopoulos, G.; Kormentzas, E.; Slobodnik, J.; Thomaidis, N. S. A Multi-Label Classifier for Predicting the Most Appropriate Instrumental Method for the Analysis of Contaminants of Emerging Concern. *Metabolites* **2022**, *12*.

(28) Hall, L. M.; Hill, D. W.; Menikarachchi, L. C.; Chen, M. H.; Hall, L. H.; Grant, D. F. Optimizing artificial neural network models for metabolomics and systems biology: An example using HPLC retention index data. *Bioanalysis* **2015**, *7*, 939–955.

(29) Aalizadeh, R. et al. Development and Application of Liquid Chromatographic Retention Time Indices in HRMS-Based Suspect and Nontarget Screening. *Analytical Chemistry* **2021**, *93*, 11601–11611.

(30) Aalizadeh, R.; Nikolopoulou, V.; Thomaidis, N. S. Development of Liquid Chromatographic Retention Index Based on Cocamide Diethanolamine Homologous Series (C(n)-DEA). *Analytical Chemistry* **2022**, *94*, 15987–15996.

(31) van Herwerden, D.; Samanipour, S. Dataset for: RPLC chemical space prediction. 2023; `https://figshare.com/articles/dataset/Dataset_for_RPLC_chemical_space_prediction/22587262`.

(32) PubChem substructure fingerprint. `https://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.pdf`.

(33) Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry* **2011**, *32*, 1466–1474.

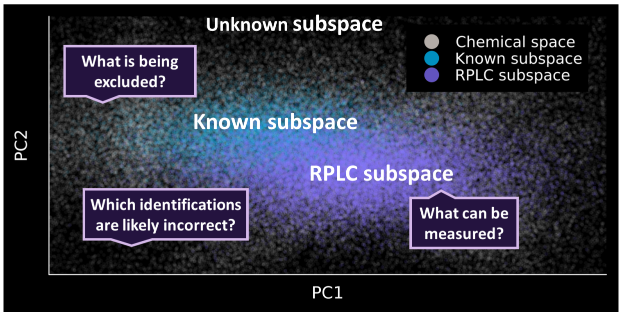
(34) Yang, F.; van Herwerden, D.; Preud'homme, H.; Samanipour, S. Collision Cross Section Prediction with Molecular Fingerprint Using Machine Learning. *Molecules* **2022**, *27*.

(35) Barron, L.; Loftus, N. Gradient retention time predictopm for 653 pesticides on a biphenyl column using machine learning. *Chromatography Today* **2019**,

(36) Klamt, A.; Eckert, F.; Reinisch, J.; Wichmann, K. Prediction of cyclohexane-water distribution coefficients with COSMO-RS on the SAMPL5 data set. *Journal of Computer-Aided Molecular Design* **2016**, *30*.

(37) Aalizadeh, R.; von der Ohe, P. C.; Thomaidis, N. S. Prediction of acute toxicity of emerging contaminants on the water flea Daphnia magna by Ant Colony Optimization-Support Vector Machine QSTR models. *Environmental science. Processes impacts* **2017**, *19*, 438–448.

TOC Graphic


Unknown subspace
● Chemical space
● Known subspace
● RPLC subspace
What is being excluded?
Known subspace
RPLC subspace
Which identifications are likely incorrect?
What can be measured?
PC2
PC1