analytical chemistry

Article

pubs.acs.org/ac

Retention Time Prediction Improves Identification in Nontargeted Lipidomics Approaches

Fabian Aicheler,*[†] Jia Li,[‡] Miriam Hoene,[§] Rainer Lehmann,[§,∥,⊥] Guowang Xu,[‡] and Oliver Kohlbacher[†,∥,⊥]

[†]Applied Bioinformatics, Center for Bioinformatics, Quantitative Biology Center, and Department of Computer Science, University of Tuebingen, Sand 14, 72076 Tuebingen, Baden-Württemberg, Germany

[‡]Key Laboratory of Separation Science for Analytical Chemistry, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian, Liaoning 116023, China

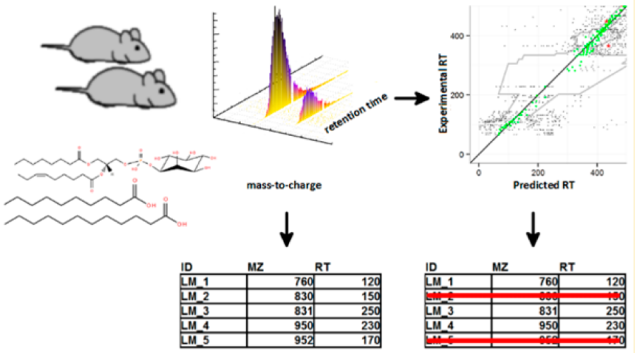
[§]Division of Clinical Chemistry and Pathobiochemistry, Department of Internal Medicine IV, University Hospital Tuebingen, 72076 Tuebingen, Baden-Württemberg, Germany

[∥]Department of Molecular Diabetology, Institute for Diabetes Research and Metabolic Diseases of the Helmholtz Centre Munich at the University of Tuebingen, 72076 Tuebingen, Baden-Württemberg, Germany

[⊥]German Center for Diabetes Research (DZD), 72076 Tuebingen, Baden-Württemberg, Germany

Supporting Information

ABSTRACT: Identification of lipids in nontargeted lipidomics based on liquid-chromatography coupled to mass spectrometry (LC-MS) is still a major issue. While both accurate mass and fragment spectra contain valuable information, retention time (t_R) information can be used to augment this data. We present a retention time model based on machine learning approaches which enables an improved assignment of lipid structures and automated annotation of lipidomics data. In contrast to common approaches we used a complex mixture of 201 lipids originating from fat tissue instead of a standard mixture to train a support vector regression (SVR) model including molecular structural features. The cross-validated model achieves a correlation coefficient between predicted and experimental test sample retention times of $r = 0.989$. Combining our retention time model with identification via accurate mass search (AMS) of lipids against the comprehensive LIPID MAPS database, retention time filtering can significantly reduce the rate of false positives in complex data sets like adipose tissue extracts. In our case, filtering with retention time information removed more than half of the potential identifications, while retaining 95% of the correct identifications. Combination of high-precision retention time prediction and accurate mass can thus significantly narrow down the number of hypotheses to be assessed for lipid identification in complex lipid pattern like tissue profiles.



ID	MZ	RT
LM_1	760	120
LM_2	830	150
LM_3	831	250
LM_4	950	230
LM_5	952	170

ID	MZ	RT
LM_1	760	120
~~LM_2~~	~~830~~	~~150~~
LM_3	831	250
LM_4	950	230
~~LM_5~~	~~952~~	~~170~~

Recent years have seen an increased interest in high-throughput analysis of the lipidome. Modern high-resolution mass spectrometric technologies enable comprehensive structural characterization and quantification of a large range of lipid molecules within cellular lipidomes.[1,2] Lipidomics has proven to be a powerful tool in biomedical research enabling biomarker discovery, mechanistic studies, drug development, and therapeutic evaluation in various disease areas including cancer, diabetes, neurodegenerative diseases, and cardiovascular disorders.[3,4] As is the case with other mass-spectrometry-based techniques, lipid identification without tandem mass spectral information poses a challenge. Isobaric lipids are an obvious problem for identification via accurate mass. Even on high-resolution instruments with mass accuracies better than 1 ppm databases cannot uniquely identify the majority of small molecules.[5] The reasons for such ambiguities differ between application areas: In proteomics, the makeup of peptides from amino acids gives rise to a large number of possible isobar permutations. Metabolomics considers mostly smaller molecules; however, metabolites cover a much more diverse chemical space, which increases the odds of misidentifications. Finally in the case of lipidomics, multiple fatty acid chains and frequent occurrence of unsaturated bonds in those chains lead to a large isobaric chemical space to consider.

The inadequacy of mass information alone for large-scale compound identification implies that additional information is required. This can either come from mass spectrometry (MS)

Received: March 25, 2015
Accepted: July 5, 2015

DOI: 10.1021/acs.analchem.5b01139

– typically tandem mass spectra (MS/MS spectra) – or from the chromatographic separation. MS/MS identification against reference libraries is commonly used for various analyte classes. Metabolite fragment spectra can be searched against for example the METLIN[6] or MassBank[7] database. The LIPID MAPS Web tools[8] support fragment search for various lipid classes. Extending the lipid database covered by LIPID MAPS, LipidBlast[9] supplies a synthetic spectral database covering about 200,000 compounds. One possible approach for database independent MS/MS-based lipid identification is LipidXplorer:[10] It employs user-defined, lipid class-specific fragmentations to examine data for viable combinations of precursor and product ions. However, understanding of lipid fragmentation rules is still incomplete. Further, MS/MS-based identification is complicated by the multitude of different instrumental setups in lipidomics, which influence fragmentation in different ways.

Chromatographic retention time is a compound-specific property that is readily available for LC-MS-based lipidomics experiments. The use of Retention Indices[11,12] (RI) for compound identification is a common approach in gas chromatography (GC). Smaller fatty acids in particular are suited for analysis via GC, and the FiehnLib[13] library for example contains RI for various lipids. GC still sees use in lipidomics, and research improving on used methods is ongoing.[14−20] For separation based on liquid chromatography (LC), retention time prediction is less trivial due to the more complex interactions with the stationary phase. As a consequence, retention time prediction in LC is still an active research area. Publications on this topic range from rather specific to more general predictors for whole compound classes.[21,22] Nevertheless, lipidomics methods with low technical retention variability are practical, for example with reversed-phase LC columns.[23,24] This allows for identification strategies leveraging solid knowledge on elution behavior and chromatographic separation of lipid classes.[24]

Existing prediction methods applicable to lipidomics were mostly designed for metabolites in general. Publications in this area have shown varying degrees of success. Creek et al. constructed a linear regression model using six parameters for hydrophilic-interaction chromatography (HILIC) columns, which focuses strongly on the logarithm of the distribution coefficient logD and derived properties.[25] Similarly, Stanstrup and co-workers adapted the use of logD as sole information for a simple yet effective t_R prediction method and applied it to a metabolite annotation pipeline.[26] However, logD-based methods are currently limited by the still lacking prediction accuracy of the distribution coefficient.[26] Moreover, logD calculation methods are often commercial, with licensing designed to protect their experimental data, methodology, or results. This complicates publication or use of their predictions in academics: Here, algorithmic methods should ideally be freely available with transparent inner workings. Another current approach for metabolite t_R prediction based on artificial neural networks (ANNs) is used in antidoping screening.[27]

Specialized methods exist for peptide retention time, for which a multitude of successful predictors have been published,[28−31] as well as for DNA retention time prediction.[32,33] Support vector regression is a popular choice for predictor models in these fields. In the most trivial case, peptide t_R predictors assume that peptide retention is defined by the sum of the contributions of its amino acids. Models then use information concerning the amino acids sequence alone. For DNA oligomers, these simple approaches do no longer hold, and secondary structure information has to be included.[32,33]

In this paper, we present a t_R prediction model for lipids trained on MS data following reversed-phase ultra high pressure liquid chromatography (UHPLC). Modern nonlinear regression methods are typically more accurate than linear models. One of these methods are ANNs mentioned above. However, ANNs are prone to converging to suboptimal solutions. We decided to use SVR instead, which lead to globally optimal solutions for nonlinear machine learning problems.[34−36] Data representation can be versatile due to the modular use of kernels in SVR. Example representations include DNA and peptides using secondary structure information,[32,37] combinations of physicochemical properties with string kernels,[38] or amino acid compositions.[29,31]

In contrast to mixtures of standard compounds commonly used for t_R predictor analysis, we chose to evaluate a more relevant complex lipid profile obtained from mouse adipose tissue. This ensures that the method can deliver proof of principle for a real, practical application. Besides demonstrating the high accuracy of the SVR-based model, we will also demonstrate its ability to significantly enhance lipid identification when used in conjunction with accurate mass searches.

■ EXPERIMENTAL SECTION

Data Acquisition. Lipidomics data used for t_R prediction model validation in this study was acquired from murine adipose tissue.[39] Briefly, methyl *tert*-butyl ether (MTBE) extraction was applied to isolate the lipidome from mouse adipose tissue according to the method described by Matyash et al.,[40] followed by global lipidomics profiling. Lipids were separated by reversed-phase UHPLC (RPLC) with ACQUITY Ultra Performance Liquid Chromatography (UPLC, Waters, USA) and ACQUITY UPLC BEH C8 column (2.1 × 100 mm, 1.7 μm, Waters, USA). Solvent A consisted of acetonitrile−water (6:4), solvent B of isopropyl alcohol−acetonitrile (9:1), with both solvents containing 10 mM ammonium acetate. Gradient elution was initiated with 32% solvent B at a flow rate of 0.26 mL/min and a column temperature of 55 °C. This composition was held for 1.5 min. Then it was linearly increased to 85% B in the next 14 min followed by a further increase to 97% B over 0.1 min, at which point the composition was kept for 2.4 min. Afterward, equilibration was done at 32% A for 2 min to prepare for the next injection. Lipidomics data was acquired with a high-resolution AB Sciex triple-TOF 5600 plus (AB Sciex, USA) mass spectrometer equipped with an electrospray ionization (ESI) ion source in negative mode. Ion spray voltage setting was set to −4,500 V and the declustering potential to −100 V. MS/MS analysis was performed using information-dependent acquisition (IDA) triggered by signal intensity, with collision energy settings at −45 V with a collision energy spread of ±15 V. Scan range for acquisition was 90−1,000 m/z for the negative mode. Mass resolution was 30,000, resulting mass accuracies were <3 ppm. Standards mixtures were used as calibration solutions for the external mass calibration every five batch samples. Lipid internal standard (IS) acquired in negative mode included d4-palmitic acid, lysophospatidylcholines (LPC) (19:0), phosphatidylcholine (PC) (17:0/17:0) and (19:0/19:0), phosphatidylethanolamine (PE) (17:0/17:0), and sphingomyelin (SM) (d18:1/12:0) which were spiked into adipose tissue sample prior to lipid extraction. ISs were purchased from Avanti Polar Lipids, Inc. (Alabaster, Alabama, USA) or Sigma-Aldrich (Taufkirchen, Germany).

Lipid Identification. The lipid identification was performed manually based on accurate mass measurement by high-resolution mass spectrometry, MS/MS fragmentation based on the "building-block" feature of lipids, and also considering the elution order of lipids in RPLC. Our experimental data is at the fatty acyl/alkyl identification level as defined by Liebisch et al.[41] Note that the data set we used for model building differs from the supplemental data set on murine adipose tissue published previously.[39] Rationale for data set modifications, selection of data points and steps for the creation of the data set as used here are detailed in Supporting Information S1. Our resultant data set contains 201 data points measured in negative mode and distributed over ten lipid classes: Apart from 32 free fatty acids (FFA) and 13 sphingomyelins (SM, including one IS), the data contained phosphoglycerides and lysophosphoglycerides. Head group moieties of observed phosphoglycerides were spread over 40 cholines (PC, including two ISs), 41 ethanolamines (PE, including one IS), 11 glycerols (PG), 13 inositols (PI), and 11 phopshatidylserines (PS). Lysophosphoglyceride moieties comprised nine cholines (LPC, including one IS) and seven ethanolamines (LPE). Additionally, plasmalogen substituents occurred in 24 plasmenylethanolamines (PE-P). Four internal standard (IS) data points included those of LPC 19:0, PC 34:0, PC 38:0, and PE 34:0.

In order to use physicochemical and other molecular descriptors for the learning task, we used structures from the LIPID Metabolites And Pathways Strategy (LIPID MAPS, LM) Structure Database.[42] LIPID MAPS structures were manually assigned to the individual lipids, using the online LIPID MAPS database. If online entries with chain lengths and number of double bonds corresponding to an experimentally determined lipid could not be found, a molecular structure was derived from similar lipids. For coeluting isomers differing only in individual side chain lengths and double bond number per chain, one was chosen for structure assignment. When unable to distinguish isobaric variants differing only in double-bond locations inside a chain, one representative structure was decided on. Using this approach, we annotated the 201 negative-mode lipids with molecular structures. Of these, eight lipids could not be labeled with a LIPID MAPS identifier (ID). Structure notations for the lipids were constructed manually according to the simplified molecular-input line-entry system (SMILES[43]).

Molecular Descriptors. The descriptors representing our lipids were chosen manually as follows: First, we aimed to include descriptors related to elution behavior, for example molecular mass and partition-coefficient logP. Second, we focused on publicly available, easily computable features. We chose 11 descriptors from RDKit,[44] an open-source cheminformatics toolkit that includes methods for the computation of a wide range of molecular descriptors. The descriptors were all computable from SMILES structures. They include an estimate of the octanol−water partition coefficient, the average molecular weight, and the approximate surface area. Multiple descriptors each characterize electrostatic interactions, hydrophobic and hydrophilic effects, and polarization properties of the molecules (see Supporting Table S3 for more details).

Descriptor values of each lipid were replaced by their normalized z-scores. This normalization improves numerical stability during the learning phase. It also prevents features with large value ranges from dominating other features during the training. To minimize the possibility of data set bias during normalization, true descriptor distributions were approximated using the whole LIPID MAPS database. In theory, using LIPID MAPS for normalization can interfere with our objective of equal feature importance in the training data. Observed prediction results however hint at sufficiently similar feature distributions of LIPID MAPS and our lipidomics data set.

Prediction Model. Support vector regression is a supervised machine learning method developed and popularized to a large extent by Vapnik.[45,46] Prediction errors are minimized with a hinge loss criterion; errors below a given threshold are ignored. Inner products play an important role as similarity measures between data points in SVR: During training, they define the optimal regressing hyperplane. The use of inner products in SVR allows mapping of nonlinear data into other inner product spaces for linear regression there. Efficient mapping and inner product calculation is done implicitly by kernel functions.

For the implementation of our regression models we used ν-SVR,[47] which is comprehensively explained in ref 48. We used the Gaussian radial basis function (RBF) kernel,[49] which allows implicit data projection into nonlinear, infinite-dimensional feature spaces.

We use LIBSVM,[50] a popular open-source library of support vector methods, including ν-SVR. Computationally less demanding analyses in an integrated lipid identification pipeline were done in KNIME,[51] a workflow-based data analytics framework. To speed up computation by parallelization, some of the evaluations were done with Python scripts and the scikit-learn machine learning package,[52] which also supports LIBSVM.

Predictor Evaluation. Predictor performances were solely evaluated on respective test sets. Evaluation was done via the squared correlation coefficients (R^2), absolute errors, and absolute percentage errors. We constructed an initial SVR predictor using the RBF kernel and trained it on 50 analytes of our data set. The eight lipids without assigned LM IDs were placed in the training set. This way, all test lipids could later be found during AMS. The remaining 42 training and 151 test points were chosen in a stratified manner to prevent test lipid classes from being underrepresented in the training set: Matching the distribution of lipid classes in training and test sets reduces the danger of reporting under- or overestimated performances for individual lipid classes.

Where specified, we evaluated models jointly using cross-validation based on repeated random subsampling to increase confidence in observed predictor characteristics. In this case, sampling was done without stratification. To evaluate lipid class-specific prediction errors for 50 training points, we constructed SVR models for 100 randomly sampled training sets of this size. Effects of the training set size on predictor performance were additionally analyzed for fixed sizes between 30 and 190, in increments of 10 training points.

As optimal model parameters are dependent on the specific data in question, all training sets were subject to nested 10-fold cross-validations. Optimal parameters were determined with nested folds and used for learning on the particular whole training data. The nested folds used for parameter determination underwent no stratification.

Lipid Identification Pipeline. We applied a lipid identification pipeline using the training set, test set, and model of our initial predictor. All steps of the pipeline, including data preparation, t_R model training and prediction, AMS and t_R-based filtering were implemented as KNIME workflows. Accurate mass search was performed using TOPP tools,[53] which are part of the OpenMS[54] package. A mass tolerance of 5 ppm was assumed to simulate common AMS candidate numbers. As library for the mass-based database search, a downloaded version of the LIPID MAPS structure database was used (compiled on March 18th, 2014, containing 37,752 entries). Allowed analyte modifications were restricted to negative ionization mode adducts. For each input compound, all candidate ions within the mass tolerance were reported. For AMS filtering, we used filtering thresholds based on the relative error between predicted and experimental t_Rs. Instead of fixed thresholds, compound specific dynamic thresholds were used for all filtering in this work. We automatically adapted the threshold for different experimental t_R ranges. Using the previously obtained optimal parameters for the training set, we retrained 200 models. For every model, we split the training data randomly into 40 training and 10 test data points each. This allowed us to get out-of-bag estimates for the relative errors of the training points. For each of our original 50 training points, the sampled errors of all training points within the three smallest experimental t_R distances were combined. The respective 95% quantiles were chosen as thresholds. In the filtering phase, given a candidate lipid, the nearest training points regarding experimental t_R were used to average their thresholds. Candidates whose relative error was larger than the respective averaged threshold were automatically rejected. A description of the pipeline including parameters of employed OpenMS tools can be found in the Supporting Information (Supporting data S6).

Data Availability and Study Reproducibility. To aid interested readers in the understanding of our used methods and to enable replication of our results, we supply various data in the Supporting Information: We explained the data set creation in more detail in Supporting data S1. We created a comma-separated file of the lipid data set, with mass-to-charge ratio, t_R, LIPID MAPS ID where available, structures and further information, as well as files representing the training and test data splits (Supporting data S2). We also supply database and adduct files needed for LIPID MAPS-based AMS, as well as the input data set for the predictor analysis on our cluster. All input or output files of our prediction and AMS pipeline workflows are included as part of the compressed workflows (Supporting data S5). We provide all Python scripts used to generate analysis results on clusters together with their results as a .zip file (Supporting data S4).

RESULTS AND DISCUSSION

The lipid retention time model described above has been assessed with respect to its accuracy, its dependence on the data set size, and its ability to support lipid identification based on accurate mass.

Model Accuracy. We first evaluated the performance of our initial predictor trained on 50 of the 201 data points. The measured and predicted t_Rs of the 151 remaining test analytes plotted against each other are shown in Figure 1. A very good overall performance with an R^2 of 0.989 was observed. Experimental retention times could be predicted within 0.5 min for 136 out of 151 test analytes. Average and median prediction error across the test compounds was 11.1 and 8.0 s, respectively. The average error corresponds to a relative deviation of 1.0% (for a gradient time of 1,080 s). Of the 151 test analytes, 121 deviated within 6% from their respective experimental t_R.

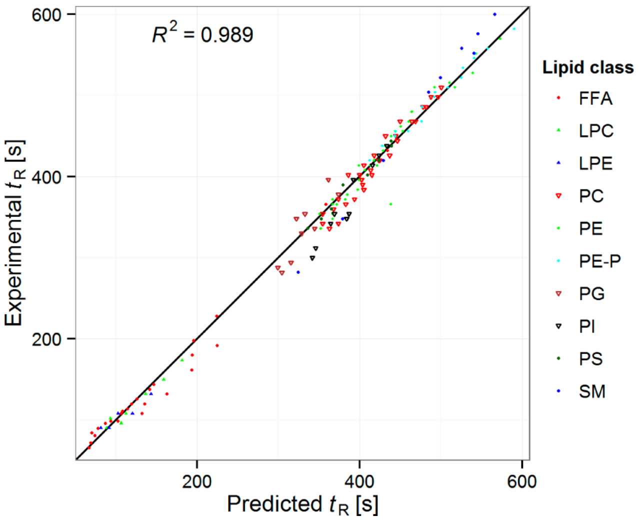


Figure 1. Comparison of predicted and experimental t_Rs of the initial model. The model was trained on 50 data points and evaluated for 151 validation analytes. The depicted test lipids are distinguished by lipid class. The listed R^2 was computed from the test lipids.

Lipids are mostly separated into two larger groups: The first group eluting in the lower t_R ranges contains FFA and lysophosphoglycerides. The second group is concentrated in elution ranges above 300 s and comprises the remaining phospholipids of our data, as well as sphingomyelins. In general, differences between predicted and observed t_Rs are larger at the extremes of elution times as well as in the sparse middle region, which separates the two main groups. Besides the sparseness of training data in these areas, the smaller sample size of SM combined with chemical differences between SM and other phospholipids might be the cause. A plot of the absolute errors versus experimental times, emphasizing inaccuracies observed in Figure 1, is available (Supporting data S7).

Model Robustness. The excellent overall prediction performance prompts the question whether these results are due to overfitting on the data set. We thus evaluated the model including its class-specific performance in more detail. Using our 201 lipids, we aggregated predictions for 100 different SVR models trained on 50 randomly sampled analytes each. Errors were evaluated on the corresponding 151 test analytes of the respective models. For each lipid class, errors were integrated across all 100 models. Apart from LPEs, FFAs, and LPCs, lipid classes had mean deviations below 13.3% in relation to experimental t_Rs of respective compounds. Over 90% of the lipids excluding these three classes had relative errors below 13.5%. If taken into account, median deviation relative to analyte t_Rs was 5.2%. 75% of all lipids deviated less than 10.4% from the experimental t_Rs. The median and mean error across all lipids were 18 and 25.6 s. This mean error corresponds to 2.4% of the gradient time. 75% of the lipids had error times below 34 s. This includes lipid classes with as few as seven experimental measurements (LPE), signifying the model's ability to leverage similarities of different lipid classes. Compared to the other lipid classes, FFA, LPC, and LPE performance varied more. LPC and LPE make up the groups with the fewest measurements. This explains a larger observed standard deviation of their relative errors. Larger relative errors of FFA are possibly due to two reasons: FFA elute over a wide range which overlaps with many other, structurally and chemically less similar, lipid classes. In other words, the similarity-based SVR might use structurally similar FFA with dissimilar t_Rs for prediction. Second, our model tolerates small absolute deviations. For the fast eluting FFA, LPC, and LPE, relative errors are proportionally larger than those of later eluting lipids with similar absolute errors. Thus, the absolute errors observed in Figure 1 for the lipid classes are in accord with measured relative error ranges. However, these deviations are not necessarily of much consequence: We do not aim to predict exact t_R but rather to discard improbable lipid suggestions.

Summarily, the behavior of our initial predictor for specific lipid classes was shown to be consistent with multiple models. While prediction performance differs between lipid classes, the differences are overall modest.

Effect of Training Set Size. A common problem with retention time models is their dependence on a specific separation system (column, gradient, etc.). This limits the usefulness of such a predictor unless the predictor can be easily retrained on a specific setup (e.g., by running a standard mix of lipids). Ideally, the model should be retrainable with as few reference lipids as possible.

To assess this issue, we repeatedly trained models using different training set sizes. Corresponding leftover analytes of our 201 lipids were used in validation sets. For each specific set size, 50 models were trained on a differently subsampled training set each to obtain performance distributions. To better estimate how well our models generalize to independent data, sampling was done randomly, without stratification by lipid class. Training set sizes ranged from 30 to 190 lipids.

The results of this evaluation are shown in Figure 2 (more detailed version in Supporting data S8). Unsurprisingly, model performance increases with increasing training set size. However, the models perform on a high level for all but the sparsest training sets. Median R^2 increases from 0.860 for models trained on 30 points to 0.992 for models trained on 190 points. This demonstrates that the models achieve excellent performance with very few training data points. Overall, 50 data points proved enough for stable models, with a median R^2 of 0.947. A plot of the training size dependent average error is available in Supporting data S9.

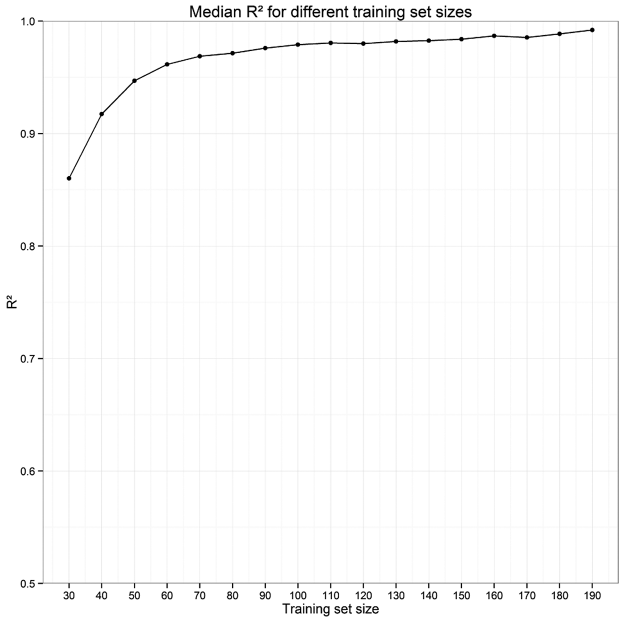


Figure 2. Predictor performance for different training set sizes. We repeatedly split our data set of 201 lipids into nonoverlapping training and test sets. Tested training set sizes were limited to the shown range and increased in steps of ten. For each size, models were trained on 50 different training sets. The depicted R^2 are the median R^2 of the respective models.

Application to Lipid Identification. Given the strong overall performance of our predictor, we wanted to evaluate its use in a practical problem setting encountered during experiment analysis. Here, t_R provides lipid information orthogonal to mass. We tested whether putative, mass-based identifications profit from our t_R prediction approach. Specifically, we applied a filter using the relative error between observed and predicted t_Rs of candidate lipids. This also provides insights whether typical experimental sample sizes suffice to train and predict against larger databases: Compared to our experimental data, the lipid space covered by the LIPID MAPS library used for accurate mass search is not only vastly larger but also more diverse and differs in its lipid composition.

We applied the same initial predictor we evaluated for model accuracy in an AMS workflow using the OpenMS accurate mass search. Common negative-mode adducts were used for putative molecule ions. Mass-to-charge ratio of the 151 compounds of the test set were matched to 6,873 candidates by AMS. On average, individual compounds matched to 46 candidates. Candidate numbers differed significantly between compounds, ranging between two (for LMGP01050133 and LMSP03010004) and 249 candidates for linoleic acid. Due to the employed training data and descriptors, our t_R predictor is not designed to differentiate between lipids of the same type which disagree in the placement of double bonds in the side chains but not in their number, or which have the same summed side chain length. In addition, we used lipids as representatives of their isobaric variants during ID assignment. We grouped such isobaric lipids together in an automated manner: Candidate lipids of a given feature were aggregated if they contained the same LIPID MAPS lipid class, chemical formula, and predicted t_Rs differing only after rounding to the third decimal place. This resulted in 1,209 candidate groups. The median number of candidates per group was three, 75% of the groups contained ten or less candidates, and the largest group of highly similar candidates contained 56 lipids.

To analyze the performance of our AMS filtering, we first looked at the behavior of our adaptive filtering threshold. For this we compared the experimental and predicted t_Rs of all candidates returned by AMS (Figure 3 a)). In agreement with the observed performance in Figure 1, the adaptive threshold is smaller for the higher t_R ranges, where predicted t_R deviates less from the observed t_R. In contrast, the sparser middle t_R region shows larger errors and wider thresholds. The same is the case for the lowest t_R regions, where underrepresented lipids classes and FFA lead to larger error and threshold windows. Overall, adaptive filtering performs well for the more diverse and populated t_R ranges while keeping the allowed error range small.

Assessing the filtering performance that is visually obvious in Figure 3 a) more quantitatively, we find filtering decreases the number of matching lipid groups per analyte drastically. Figure 3 b) shows the median number of lipid candidates and groups assigned to a feature via accurate mass search before and after retention time filtering. The median number of candidates

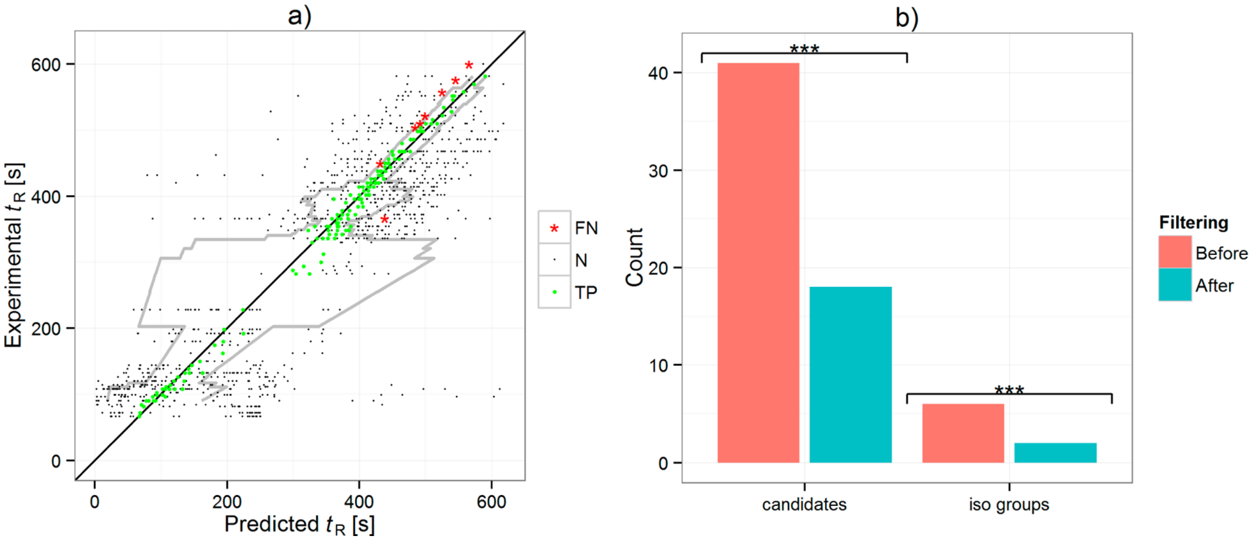


Figure 3. Results of the t_R-assisted AMS filtering pipeline. a) Predicted and measured t_Rs of grouped AMS candidate lipids. Only lipid candidates inside a sensible range of predicted t_R are shown, as defined by the experimental elution. The filtering threshold for the different t_R ranges is depicted in gray. The candidate groups fall into three categories: Groups containing the manually assigned LM ID below the filtering threshold (true positives, TP) are shown as green dots. Wrongly filtered groups containing the LM ID (false negatives, FN) are depicted as red stars. Black dots signify candidate groups that do not contain an assigned ID (negatives, N). b) Filtering efficiency of the AMS pipeline using the t_R predictor. The median number of candidates or candidate groups (iso groups) per test analyte was determined. Medians before and after filtering were compared and evaluated using the Wilcoxon signed-rank-test statistical test.

decreases from 41 to 18, while the median number of groups decreases from six down to two. Both median changes are significant as determined by Wilcoxon signed-rank test: P-value for the candidate distributions is 2.15e-24, whereas the p-value for the group distributions is 1.76e-24. Overall, unfiltered AMS output contained 1,209 groups, of which 151 (12.5%) included the manually assigned LIPID MAPS IDs. Filtering resulted in 542 groups, with 143 (26.4%) groups containing assigned IDs retained. The number of candidates was also more than halved, with 3,423 of 6,873 candidates remaining. Thus, filtering almost doubled the proportion of correct candidates among all candidates while simultaneously retaining 94.7% of the correct candidates.

■ CONCLUSION

In this work we show that support vector regression based on molecular descriptors permits a robust and highly accurate prediction of lipid retention time even from as few as 50 different lipids. The resulting model contains sufficient information to cut the number of candidate lipids in half. It is thus a valuable tool in assigning lipid structures and for the automated annotation of lipidomics data. We believe that t_R-based ranking is a powerful tool for annotation and can assist in the prioritization of additional measurements for structure elucidation: A common lipidomics workflow includes (high-resolution) nontargeted LC-MS analysis followed by software screening against MS and MS/MS databases using e.g. LipidView (AB Sciex) and manual and time-consuming refinement of screening results. Automatic exclusion of false candidates that fall in unlikely t_Rs facilitates final manual identification. This could be of even more use for low-resolution MS, where the lower accuracy of mass measurements results in even more false candidates and would benefit a high number of laboratories that cannot afford high-resolution instruments. The strategy illustrates the advantage of nonlinear regression approaches for t_R prediction of metabolites. With regards to the presented predictor and filtering pipeline, their nature as proof of principle methods allows for further refinement in the future. Our results show a high potential for the application of SVR-based t_R prediction in the large-scale experiments analytical chemists are facing nowadays in MS-based lipidomics research. While our method can be applied to other types of experimental parameters suitable for lipid separation (e.g., other stationary phase, column and particle dimensions, mobile phase selectivity, etc.) by retraining the model, the model quality would need to be validated in each case.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Additional information as noted in text. Information on LC reproducibility in Supporting Material S10. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.5b01139.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: 49-7071-29-70461. Fax: 49-7071-29-5152. E-mail: aicheler@informatik.uni-tuebingen.de.

Author Contributions

The manuscript was written through contributions of all authors. Fabian Aicheler contributed with the computational method design and implementation. Jia Li and Miriam Hoene contributed the experimental data. Rainer Lehmann, Guowang Xu, and Oliver Kohlbacher contributed to the idea conception of the study. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was funded in part by a grant from the Sino-German Center for Research Promotion (GZ 753 by DFG and NSFC to G.X. and R.L. and LE 1391/1-1 by DFG to R.L.); by a grant from the German Federal Ministry of Education and Research (BMBF) to the German Centre for Diabetes Research (DZD e.V., Grant 01GI0925); by BMBF grant BIOMARKERS - FKZ 01GI1104A; by the European Commission FP7MARINA (Grant No. 236215); by DFG grant KO 2313/6-1; and by the foundation and creative research group project (No. 21175132 and No. 21321064 by NSFC to G.X.).

■ REFERENCES

(1) Han, X.; Gross, R. W. *J. Lipid Res.* **2003**, *44*, 1071−1079.

(2) Wenk, M. R. *Nat. Rev. Drug Discovery* **2005**, *4*, 594−610.

(3) Puri, R.; Duong, M.; Uno, K.; Kataoka, Y.; Nicholls, S. J. *Expert Opin. Drug Discovery* **2012**, *7*, 63−72.

(4) Hu, C.; van der Heijden, R.; Wang, M.; van der Greef, J.; Hankemeier, T.; Xu, G. *J. Chromatogr. B: Anal. Technol. Biomed. Life Sci.* **2009**, *877*, 2836−2846.

(5) Kind, T.; Fiehn, O. *BMC Bioinf.* **2006**, *7*, 234.

(6) Smith, C. A.; O'Maille, G.; Want, E. J.; Qin, C.; Trauger, S. A.; Brandon, T. R.; Custodio, D. E.; Abagyan, R.; Siuzdak, G. *Ther. Drug Monit.* **2005**, *27*, 747−751.

(7) Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; Oda, Y.; Kakazu, Y.; Kusano, M.; Tohge, T.; Matsuda, F.; Sawada, Y.; Hirai, M. Y.; Nakanishi, H.; Ikeda, K.; Akimoto, N.; Maoka, T.; Takahashi, H.; Ara, T.; Sakurai, N.; Suzuki, H.; Shibata, D.; Neumann, S.; Iida, T.; Tanaka, K.; Funatsu, K.; Matsuura, F.; Soga, T.; Taguchi, R.; Saito, K.; Nishioka, T. *J. Mass Spectrom.* **2010**, *45*, 703−714.

(8) Fahy, E.; Sud, M.; Cotter, D.; Subramaniam, S. *Nucleic Acids Res.* **2007**, *35*, W606−612.

(9) Kind, T.; Liu, K.-H.; Lee, D. Y.; DeFelice, B.; Meissen, J. K.; Fiehn, O. *Nat. Methods* **2013**, *10*, 755−758.

(10) Herzog, R.; Schwudke, D.; Shevchenko, A. In *Current Protocols in Bioinformatics*; John Wiley & Sons, Inc., 2013.

(11) Kováts, E. *Helv. Chim. Acta* **1958**, *41*, 1915−1932.

(12) van Den Dool, H.; Dec. Kratz, P. *J. Chromatogr. A* **1963**, *11*, 463−471.

(13) Kind, T.; Wohlgemuth, G.; Lee, D. Y.; Lu, Y.; Palazoglu, M.; Shahbaz, S.; Fiehn, O. *Anal. Chem.* **2009**, *81*, 10038−10048.

(14) Seppänen-Laakso, T.; Laakso, I.; Hiltunen, R. *Anal. Chim. Acta* **2002**, *465*, 39−62.

(15) Tranchida, P. Q.; Costa, R.; Donato, P.; Sciarrone, D.; Ragonese, C.; Dugo, P.; Dugo, G.; Mondello, L. *J. Sep. Sci.* **2008**, *31*, 3347−3351.

(16) Wahjudi, P. N.; Yee, J. K.; Martinez, S. R.; Zhang, J.; Teitell, M.; Nikolaenko, L.; Swerdloff, R.; Wang, C.; Lee, W. N. *J. Lipid Res.* **2011**, *52*, 2226−2233.

(17) Bogusz, S., Jr.; Hantao, L. W.; Braga, S. C.; de Matos Franca Vde, C.; da Costa, M. F.; Hamer, R. D.; Ventura, D. F.; Augusto, F. *J. Sep. Sci.* **2012**, *35*, 2438−2444.

(18) Payeur, A. L.; Lorenz, M. A.; Kennedy, R. T. *J. Chromatogr. B: Anal. Technol. Biomed. Life Sci.* **2012**, *893*−*894*, 187−192.

(19) Cacas, J. L.; Melser, S.; Domergue, F.; Joubes, J.; Bourdenx, B.; Schmitter, J. M.; Mongrand, S. *Anal. Bioanal. Chem.* **2012**, *403*, 2745−2755.

(20) Shen, Y.; Xu, Z. *J. Chromatogr. B: Anal. Technol. Biomed. Life Sci.* **2013**, *930*, 36−40.

(21) Héberger, K. *J. Chromatogr. A* **2007**, *1158*, 273−305.

(22) Kaliszan, R. *Chem. Rev.* **2007**, *107*, 3212−3246.

(23) Yamada, T.; Uchikata, T.; Sakamoto, S.; Yokoi, Y.; Fukusaki, E.; Bamba, T. *J. Chromatogr. A* **2013**, *1292*, 211−218.

(24) Sandra, K.; dos Santos Pereira, A.; Vanhoenacker, G.; David, F.; Sandra, P. *J. Chromatogr. A* **2010**, *1217*, 4087−4099.

(25) Creek, D. J.; Jankevics, A.; Breitling, R.; Watson, D. G.; Barrett, M. P.; Burgess, K. E. V. *Anal. Chem.* **2011**, *83*, 8703−8710.

(26) Stanstrup, J.; Gerlich, M.; Dragsted, L. O.; Neumann, S. *Anal. Bioanal. Chem.* **2013**, *405*, 5037−5048.

(27) Miller, T. H.; Musenga, A.; Cowan, D. A.; Barron, L. P. *Anal. Chem.* **2013**, *85*, 10330−10337.

(28) Moruz, L.; Tomazela, D.; Kall, L. *J. Proteome Res.* **2010**, *9*, 5209−5216.

(29) Pfeifer, N.; Leinenbach, A.; Huber, C. G.; Kohlbacher, O. *BMC Bioinf.* **2007**, *8*, 468.

(30) Schulz-Trieglaff, O.; Pfeifer, N.; Gröpl, C.; Kohlbacher, O.; Reinert, K. *BMC Bioinf.* **2008**, *9*, 423.

(31) Pfeifer, N.; Leinenbach, A.; Huber, C. G.; Kohlbacher, O. *J. Proteome Res.* **2009**, *8*, 4109−4115.

(32) Sturm, M.; Quinten, S.; Huber, C. G.; Kohlbacher, O. *Internat. Begegnungs-und Forschungszentrum für Informatik*; 2006.

(33) Kohlbacher, O.; Quinten, S.; Sturm, M.; Mayr, B. M.; Huber, C. G. *Angew. Chem., Int. Ed.* **2006**, *45*, 7009−7012.

(34) Suykens, J. A.; Van Gestel, T.; De Moor, B.; Vandewalle, J. *Least Squares Support Vector Machines*; World Scientific: 2002; pp 29−70.

(35) Shawe-Taylor, J.; Cristianini, N. *Kernel methods for pattern analysis*; Cambridge University Press: 2004; pp 195−251.

(36) Suykens, J. A.; Vandewalle, J. *Neural Process. Lett.* **1999**, *9*, 293−300.

(37) Sturm, M.; Quinten, S.; Huber, C. G.; Kohlbacher, O. *Nucleic Acids Res.* **2007**, *35*, 4195−4202.

(38) Toussaint, N. C.; Widmer, C.; Kohlbacher, O.; Rätsch, G. *BMC Bioinf.* **2010**, *11*, S7.

(39) Hoene, M.; Li, J.; Häring, H.-U.; Weigert, C.; Xu, G.; Lehmann, R. *Biochim. Biophys. Acta, Mol. Cell Biol. Lipids* **2014**, *1841*, 1563−1570.

(40) Matyash, V.; Liebisch, G.; Kurzchalia, T. V.; Shevchenko, A.; Schwudke, D. *J. Lipid Res.* **2008**, *49*, 1137−1146.

(41) Liebisch, G.; Vizcaíno, J. A.; Köfeler, H.; Trötzmüller, M.; Griffiths, W. J.; Schmitz, G.; Spener, F.; Wakelam, M. J. *J. Lipid Res.* **2013**, *54*, 1523.

(42) Sud, M.; Fahy, E.; Cotter, D.; Brown, A.; Dennis, E. A.; Glass, C. K.; Merrill, A. H.; Murphy, R. C.; Raetz, C. R. H.; Russell, D. W.; Subramaniam, S. *Nucleic Acids Res.* **2007**, *35*, D527−D532.

(43) Weininger, D. *J. Chem. Inf. Model.* **1988**, *28*, 31−36.

(44) Landrum, G. [Online]. 2006. http://www.rdkit.org/ (accessed 20/01/2015).

(45) Cortes, C.; Vapnik, V. *Machine Learning* **1995**, *20*, 273−297.

(46) Drucker, H.; Burges, C. J.; Kaufman, L.; Smola, A.; Vapnik, V. *Adv. Neural Inf. Process. Syst.* **1997**, *9*, 155−161.

(47) Schölkopf, B.; Smola, A. J.; Williamson, R. C.; Bartlett, P. L. *Neural Comput.* **2000**, *12*, 1207−1245.

(48) Smola, A.; Schölkopf, B. *Statistics and Computing* **2004**, *14*, 199−222.

(49) *Kernel Methods in Computational Biology*; Vert, J.-P., Tsuda, K., Schölkopf, B., Eds.; 2004; pp 35−70.

(50) Chang, C.-C.; Lin, C.-J. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1−27.

(51) Berthold, M.; Cebron, N.; Dill, F.; Gabriel, T.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. In *Data Analysis, Machine Learning and Applications*; Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R., Eds.; Springer: Berlin, Heidelberg, 2008; pp 319−326.

(52) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(53) Kohlbacher, O.; Reinert, K.; Gröpl, C.; Lange, E.; Pfeifer, N.; Schulz-Trieglaff, O.; Sturm, M. *Bioinformatics* **2007**, *23*, e191−197.

(54) Sturm, M.; Bertsch, A.; Gröpl, C.; Hildebrandt, A.; Hussong, R.; Lange, E.; Pfeifer, N.; Schulz-Trieglaff, O.; Zerck, A.; Reinert, K.; Kohlbacher, O. *BMC Bioinf.* **2008**, *9*, 163.

DOI: 10.1021/acs.analchem.5b01139