analytical chemistry

Article

pubs.acs.org/ac

PredRet: Prediction of Retention Time by Direct Mapping between Multiple Chromatographic Systems

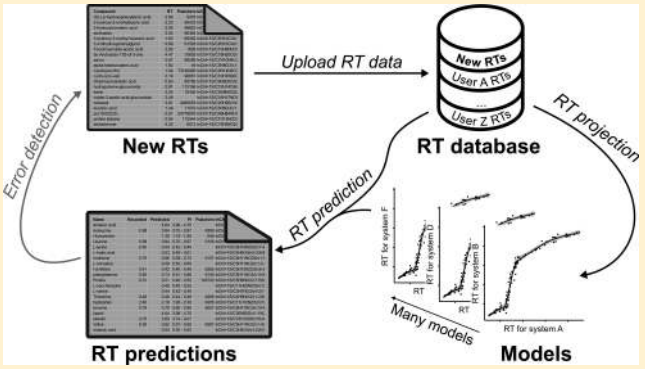
Jan Stanstrup,*[,†] Steffen Neumann,[‡] and Urška Vrhovšek[†]

[†]Department of Food Quality and Nutrition, Research and Innovation Centre, Fondazione Edmund Mach (FEM), Via E. Mach 1, 38010 San Michele all'Adige, Trentiono, Italy

[‡]Department of Stress and Developmental Biology, Leibniz Institute of Plant Biochemistry, Weinberg 3, 06120 Halle, Germany

Supporting Information

ABSTRACT: Demands in research investigating small molecules by applying untargeted approaches have been a key motivator for the development of repositories for mass spectrometry spectra and automated tools to aid compound identification. Comparatively little attention has been afforded to using retention times (RTs) to distinguish compounds and for liquid chromatography there are currently no coordinated efforts to share and exploit RT information. We therefore present PredRet; the first tool that makes community sharing of RT information possible across laboratories and chromatographic systems (CSs). At http://predret.org, a database of RTs from different CSs is available and users can upload their own experimental RTs and download predicted RTs for compounds which they have not experimentally determined in their own experiments. For each possible pair of CSs in the database, the RTs are used to construct a projection model between the RTs in the two CSs. The number of compounds for which RTs can be predicted and the accuracy of the predictions are dependent upon the compound coverage overlap between the CSs used for construction of projection models. At the moment, it is possible to predict up to 400 RTs with a median error between 0.01 and 0.28 min depending on the CS and the median width of the prediction interval ranging from 0.08 to 1.86 min. By comparing experimental and predicted RTs, the user can thus prioritize which isomers to target for further characterization and potentially exclude some structures completely. As the database grows, the number and accuracy of predictions will increase.



In the untargeted analysis of complex mixtures, one of the major bottlenecks is the identification of compounds. In untargeted analyses, and in analytical chemistry in general, liquid chromatography (LC) plays a central role due to its high separation power and versatility. LC is often coupled to mass spectrometry (MS), as MS offers superior sensitivity and selectivity, such that even compounds present at very low concentrations can be measured.

Especially in metabolomics there has been a drive to create resources to aid rapid compound identification. One such resource is data repositories for MS spectra that have been developed with great success to the benefit of the scientific community.[1,2] In addition, recently efforts have been made to create automated tools, often assisted by these databases of experimental data, to aid compound identification.[3−5] However, these tools and databases only focus on one aspect of the experimental data: the fragmentation of the compounds formed in the mass spectrometers. However, utilizing only the fragmentation is ignoring half of the available information. The retention time (RT) is often neglected.

Because the mass spectra is often the only information considered, it is common for researchers working with LC−MS-based metabolomics to encounter many compounds where mass and fragmentation are ambiguous and could match multiple molecular structures. Confirming the structure thus becomes a process of elimination where authentic standards have to be purchased or synthesized to compare to the experimental MS spectra and RT. Prior knowledge of the RT of the plausible structures would allow further reduction of the number of compound structures that need to be investigated.

Several different strategies for RT prediction have therefore been developed. For peptides, strategies based on summing the effect of each amino acid are common.[6,7] This is, however, not feasible as a general approach in metabolomics where molecules have more diverse structures. Therefore, approaches have been developed that rely on complex models based on physicochemical descriptors of compounds[8−13] or in the most simple form model log P or log D to RT.[14] These quantitative structure−retention relationship (QSRR) models share the characteristic that they require a large number of training

Received: June 17, 2015
Accepted: August 20, 2015
Published: August 20, 2015

DOI: 10.1021/acs.analchem.5b02287
Anal. Chem. 2015, 87, 9421−9428

compounds and the more complex models risk severe overfitting, making them less generally applicable. While these models can be applied to any molecular structure, they currently have limited accuracy, in part due to the limited accuracy of the underlying physicochemical descriptors.[14]

In gas chromatography, retention indexes are routinely used to make different systems comparable, but for LC there are currently no coordinated efforts to share and exploit information regarding the RT of compounds. The reason RT information has been neglected in LC systems is that the RT is specific to a specific chromatographic system (CS) and there is no general agreement on RT references.

We have therefore sought to rectify this by building a database of compound RTs. With this database, we are able to map the RT of compounds between CS if they reasonably similar. Experimentally determined RTs of a number of compounds in two different systems, is used to build a model between the RTs in the two systems. If the RT of a compound is then known in only one of the CSs, this model can be used to predict the RT of the compound in the other CS. Building these models between all CSs in the database thus allows predicting the RT of a high number of compounds in CSs where they have not been experimentally determined.

For the first time, the developed tool makes community sharing of RT information possible across laboratories and CSs. The free, open source and Web-based tool is available at www.predret.org while an R package for querying the database is available at https://github.com/stanstrup/PredRet.

EXPERIMENTAL SECTION

Retention Time Mapping. The basis of the prediction system is a database consisting of compound retention times (RTs) recorded on different chromatographic systems (CSs). The mapping, i.e., "conversion", of RTs from one CS to another is done pairwise for each possible pair of CSs. Since retention order is not conserved nor predictable between very different types of chromatographic columns (e.g., hydrophilic interaction liquid chromatography (HILIC) and reversed-phase liquid chromatography), only pairs of CS with the same type of chromatography are used.

For each pair of CSs in the database, the RTs of compounds measured in both systems are used to construct a monotonically constrained generalized additive model (GAM) between the RTs using the R package mgcv.[15] To make the models more robust against outliers, the residuals of an initial GAM are used to weigh each data point in the penalized constrained least-squares fitting (PCLS) step that follows the application of monotonic constraints. Instead of using the residuals directly, a sigmoidal function is applied to the residuals normalized to the total CS runtime such that

$$w_i = \frac{1}{1 + e^{-\alpha\left(\frac{|res_i|}{\max(RT_y)} - \beta\right)}} \quad (1)$$

where w_i is the weight applied to the ith data point, res_i is the residual of the ith data point, and α and β are model tuning parameters. The tuning parameters were set to $\alpha = -30$ and $\beta = 0.1$. The longest RT in the CS to be predicted, $\max(RT_y)$, was used as a surrogate for the total runtime in the CS.

Bootstrapping with 1 000-fold resampling was used to construct 99% prediction intervals (PIs) for the models. For each combination of CSs, the experimental RT needs to have been determined for at least 10 compounds in both systems, otherwise the CS combination is skipped.

The RT prediction process is triggered once the models have been constructed between all pairs of CSs with new data.

Retention Time Prediction. The database is periodically checked for updated models and new predictions are made where models have been updated. The prediction process is then run for each CS with corresponding updated models. All compounds known in systems with models to the CS under consideration are collected. Then for each of these compounds, the GAMs are used to predict the RT in the CS under consideration. In cases where predictions for one compound can be made based on several GAMs, the prediction with the smallest PI is used. Predictions with PI width larger than 2 min or more than 20% of the predicted RT are discarded as unreliable. Predictions are also discarded where the density of observations is low.

Finally, the predictions are also used to discover user reported (experimental) RTs that seem to be implausible. A RT is considered potentially incorrect if the difference between the recorded and the predicted RT is more than twice the distance from the predicted RT to the outer limits of the PI. If a reported RT falls outside these limits, the value is marked as "suspicious" and ignored in subsequent modeling and prediction iterations. The "suspicious" entries are also listed in the Web application for the user to inspect.

Web Interface. A Web interface was constructed such that the user can easily upload RT data and download the new RT predictions. On the Web site, the user is first prompted to define a new CS. Each system will have (1) a name, (2) a column type (example, "Reversed-phase", "HILIC"), (3) a column description (example, "Waters ACQUITY UPLC BEH C18"), (4) an eluent system (example, "(95:5 Water/ACN)/(ACN)"), (5) the eluent pH (example, "acidic", "alkaline"), (6) eluent additives (example, "0.1% Formic acid").

In the next step the user will upload a CSV file containing RTs for compounds measured in his or her own system. The CSV file needs to contain the name of each compound, the measured RT, either PubChem CID or InChI, and the name of the CS, which the user defined in the system. The system then automatically converts PubChem CIDs to InChIs and for each structure stereochemistry, charges, and salts are removed to allow for unambiguous comparison of molecular structures.

The upload of new data to the database triggers recalculation of models between the new CS and existing CSs that have sufficient overlap. The Web interface was built using R v3.1.0[16] and the R package Shiny v0.10.2.9003 running on Shiny-server v1.2.1.362 (www.rstudio.com/shiny). The shiny application was integrated with the login system of the content management system WordPress v4.1 (https://wordpress.org) running on Apache HTTP Server v2.4.7 (http://httpd.apache.org). The database of RTs is stored in a Mongodb v2.4.10 (https://www.mongodb.org) database and queried using Rmongodb v1.7.3 (https://github.com/mongosoup/rmongodb).

Retention Time Database. The current database used for this project was constructed from many sources: (1) Two systems (LIFE_old,[17] LIFE_new[17]) recorded at the Department of Nutrition, Exercise and Sports, Faculty of Science, University of Copenhagen (NEXS), Denmark. (2) Five systems (FEM_long,[18] FEM_short,[19] FEM_orbitrap_plasma, FEM_orbitrap_urine, FEM_lipids[20]) recorded at the Department of Food Quality and Nutrition, Research and Innovation Centre,

DOI: 10.1021/acs.analchem.5b02287
Anal. Chem. 2015, 87, 9421−9428

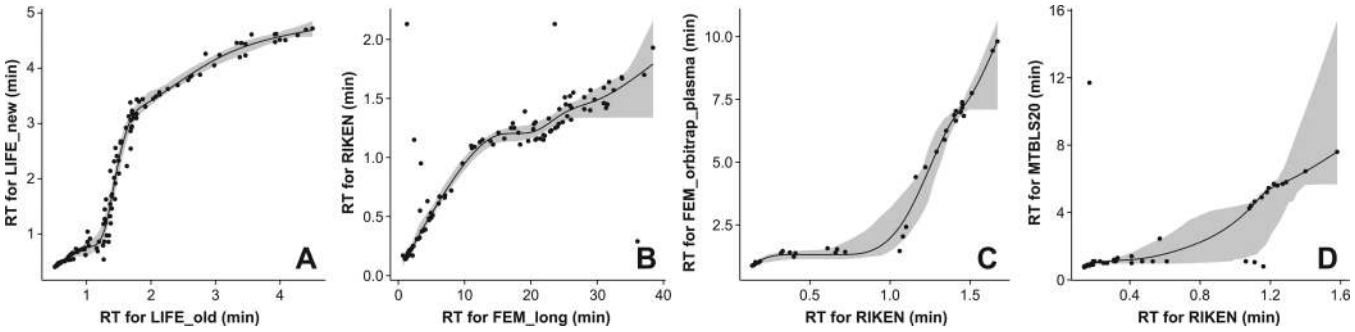


Figure 1. Examples of robust monotonically constrained generalized additive models between the retention times of compounds in two different chromatographic systems. Examples are given of a "good" model (A), a model with many outliers (B), a model where there are only enough data points to get predictions in a small RT interval (C), and a model where there are not enough data points to establish a model with reasonable prediction accuracy (D). All examples are of the initial model that will be refined further as erroneous entries are discarded.

Fondazione Edmund Mach (FEM), Italy. (3) One system (IPB_Halle[14,21]) recorded at the Department of Stress and Developmental Biology, IPB Halle, Germany. (4) One system (RIKEN[22]) recorded at and made available by the RIKEN Plant Science Center, Yokohama, Kanagawa, Japan. (5) Three systems available in the published literature (Cao_HILIC,[12] Eawag_XBridgeC18,[23,24] MPI_Symmetry[25]). (6) Nine systems (MTBLS4,[26] MTBLS17,[27] MTBLS19,[28] MTBLS20,[29] MTBLS36,[30] MTBLS36,[30] MTBLS38,[31] MTBLS39,[32] MTBLS87,[33] MTBLS52) publicly available at MetaboLights.[34] (7) Two systems extracted from MassBank[1,24] (UFZ_Phenomenex, UniToyama_Atlantis).

From these 23 systems, two are HILIC based systems (MTBLS87, Cao_HILIC) while the rest are acidic C18-based reversed-phase (RP) systems. A range of C18 columns were used, with the Waters ACQUITY UPLC HSS T3 C18 column being the most popular and used in 40% of the systems. Different gradients using different combinations of water together with acetonitrile, methanol, acetone, and isopropanol were used. A description of each system is available in the Supporting Information.

RESULTS AND DISCUSSION

We present PredRet, a Web-based service that allows easy sharing of retention time (RT) data and unsupervised mapping of RTs between different chromatographic systems (CSs). The user can upload RT data from their own CS. Then the user will be able to view or download the predicted RTs.

Modeling Retention Times. In the PredRet system the RTs of compounds measured in two different CSs are used to build a generalized additive model (GAM) that describe the relationship between RTs in the two systems. It is thus a projection system that projects, or maps, RTs in one system to RTs in another system. Examples of models that map the RT from one system to another can be seen in Figure 1.

The hypothesis behind this mapping is that the elution order is *largely* conserved and the models can therefore be monotonically constrained which increases robustness considerably, especially when there are few data points (i.e., compounds where the RT is known in both systems). This assumption is only valid for similar CSs. For example it is not possible to project RTs from a RP system to a HILIC systems. Therefore, models are only constructed between all *similar* CSs (currently only RP/HILIC are separated). Since all the RP CSs in the database are currently acidic RP C18-based systems, we have not tested the limits of how similar CSs need to be for it to be possible to create sensible models. We presume that systems should also be separated based on acidic/alkaline eluent characteristics and very different column stationary phases (for example, C8-based) while we found that differences in the organic eluents, gradient, and specific C18 columns do not invalidate the assumption of largely conserved elution order judging by the CSs currently in the database. PredRet, therefore, support adding RT data from any CS.

Current Database. Currently the database contains 3 300 RTs entries across 23 CSs. Overall, the database covers 1 700 unique compounds. In Figure 2 it can be observed how the

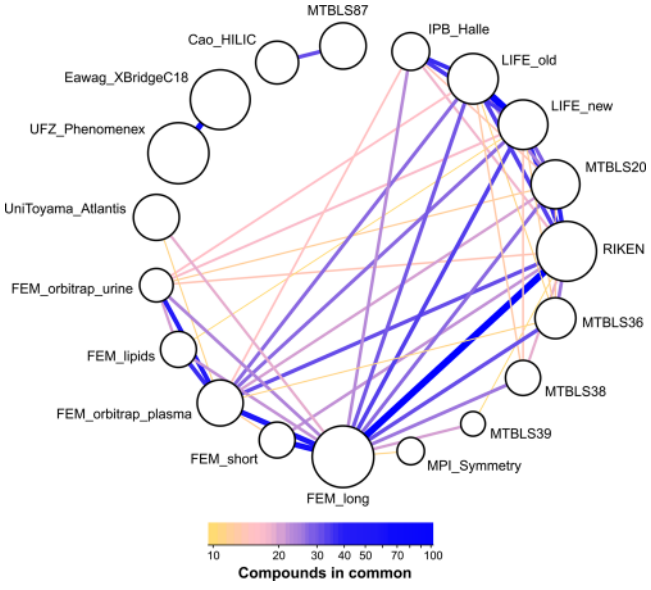


Figure 2. Network showing compound coverage overlap between chromatographic systems (CS). Only CSs that have at least one connection are shown. The lines connecting the CSs show the extent of overlap and go from thin orange to thick blue.

difference CSs are connected in terms of coverage overlap. The compounds are small (<1000 Da) molecules ranging from compounds found in humans, including some lipids, environmental contaminants, plants, and foods as well as wine.

The number of entries for each CS in the database are reported as green bars in Figure 3A. These systems cover a range of typical chromatographic conditions from different laboratories and each CS comprises tens to hundreds of entries.

DOI: 10.1021/acs.analchem.5b02287
Anal. Chem. 2015, 87, 9421−9428

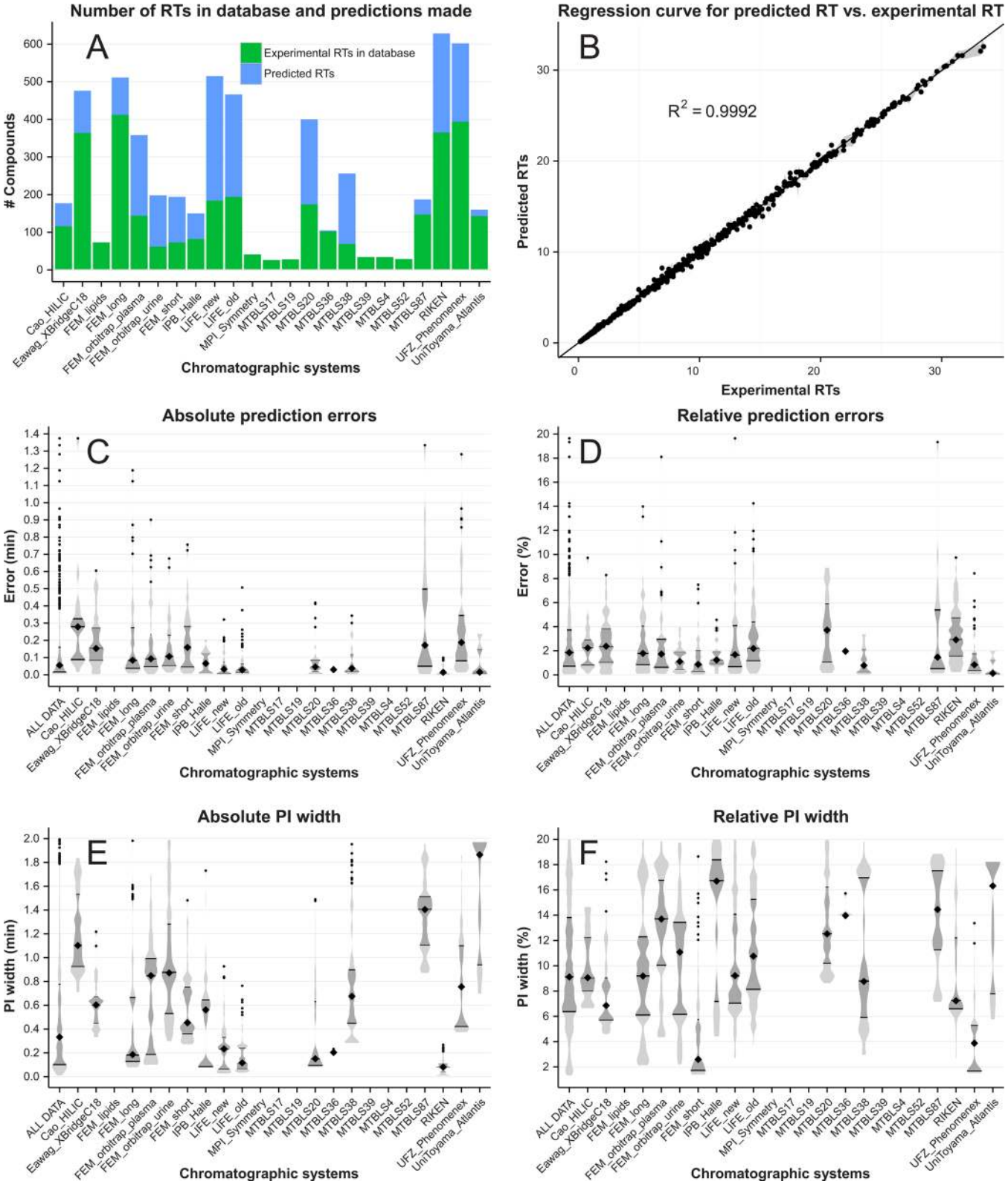


Figure 3. In part A, the number of entries for each system in the database are reported (green) as well as the number of predicted retention times that can be made for each system (blue). In part B, a regression curve between predicted and experimental retention time is shown for all predictions where the experimental RT is known. The prediction intervals are in gray. In parts C−F, the distribution of the absolute and relative prediction errors (C,D) and width of the prediction intervals (E,F) are shown. The width of the bars represent the density of the data. The quartiles have been marked. Data points have been marked explicitly if they exceed 1.5 times the interquartile range.

Overlap in the compound coverage is required to construct models between CSs.

In the upper part of Figure 3A, the corresponding number of predictions for each system are reported in blue. The number

DOI: 10.1021/acs.analchem.5b02287
Anal. Chem. 2015, 87, 9421−9428

of compounds for which RTs can be predicted depends mainly on the number of compounds where the RT is known also in other CSs. In general, this means that the more known RTs the user supplies, the more RTs can be predicted. However, relatively few known RTs can be sufficient to achieve a large number of predicted RTs given suitable compound coverage overlap between CSs. This is the case for, e.g., MTBLS38 that has overlap with four of the CSs that have the most data (see Figure 2) and therefore 69 known RTs were enough to predict 187 new RTs. In the current database there are 7 CSs that do not have sufficient overlap with other CSs to make predictions. Expansion of the database will lead to better coverage of different chromatographic conditions such that different column types and pH environments can be supported.

Prediction Intervals. When RT predictions are used to annotate untargeted data sets, it is useful to know not just the predicted RT but also a prediction interval (PI). PIs are unfortunately not easily established for all but the simplest models, and for this reason we found that of the previously developed RT prediction methods only the models based on a linear relationship between log *P* and RT provide PIs and not for individual predictions as such.[35] Likewise, we are currently not aware of methods for building PIs for GAM models with constraints. We therefore used bootstrapping to establish empirical PIs.

When the PIs for each prediction are joined we get prediction bands. Correctly defined prediction bands would ensure that the fraction of predictions where the true RTs fall outside the prediction band is no more than the chosen confidence level. In our case, however, because of multiplicity issues this cannot be ensured. PIs should therefore be regarded as anticonservative and not be used as strict filters to exclude a possible compound annotation but rather as an indication of the likelihood of a match.

The accuracy of the predictions and width of the PIs depend on the accuracy of the projection model associated with each prediction. The accuracy and precision of this model depends mainly on the number of compounds that have known RTs in both CSs; or more precisely, it depends on the number of known RTs in the RT range of the compound that is to be translated from one CS to the other. Since the RT of one compound can potentially be predicted from several CSs, the model that provides the prediction with the narrowest PI is used.

Another potential source of inaccuracies, that all prediction systems suffer from, is the potential for small changes in RTs over time. This can be caused by column aging but also by small changes in the eluent composition such as slight pH differences. These changes are not necessarily systematic and can therefore increase the width of the PIs if the set of RTs used for modeling was accumulated over time. In addition the accuracy of the predictions might be lower than expected based on the model if the "current" system differs sufficiently from the system the original RTs were recorded on.

In Figure 1A the model used to predict RTs from the CS "LIFE_old" to the CS "LIFE_new" can be seen. This model is very accurate since there are no major outliers and many compounds with known RT in both CSs. This leads to very narrow PIs for the predictions. It should be noted, however, that due to the lack of methods to construct accurate prediction bands as explained above, the very narrow PIs can be misleading. In Figure 1A, for the interval around 1.5 min (for LIFE_old) the PI is narrow in part because the slope is steep, which makes it easier to establish the fitted curve. Unfortunately a steep slope also means that relatively small inaccuracies (instrumental variation) in the independent variable (LIFE_old RT) leads to relatively large errors in the prediction of the dependent variable (LIFE_new RT) which the PI does not reflect.

We will thus reiterate that the PIs are anticonservative and should be approached as such. Nevertheless, we believe that providing anticonservative PIs are better than having no PIs at all.

To avoid reporting RTs based on models of clearly insufficient quality, predictions are skipped where the density of observations is low (i.e., in the RT region of the prediction there are few experimentally determined RTs). Likewise for predictions with very wide PIs (see Figure 1C,D).

Prediction Accuracy. For compounds where both the experimental and predicted RT are known, a regression curve between the two can be constructed. The regression curve for all predictions is shown in Figure 3B and has R^2 = 0.9992. While this is a common model quality parameter, it relates poorly to more intuitive measures of accuracy as it does not tell how closely experimental and predicted RTs can be expected to match. Therefore, in Figure 3C, the distribution of the absolute prediction error is reported for each CS. It can be seen that the median error for a CS can be as low as 0.01 min and is in all cases below 0.28 min. In these plots, the width of the bars represent the density distribution of the predicted data.

Some CSs, e.g., RIKEN, have exceptionally low *absolute* errors. In the case of the RIKEN system, this is because it is an extremely short gradient system with a run time of only 3 min. It can therefore be more instructive to compare different systems by looking at the errors *relative* to the true RTs. The relative error in the prediction is shown in Figure 3D, and it can be seen that the median error is in all cases below 3.7%. Figure 3C,D shows that a few percent of compounds exhibit much larger errors. While it cannot be verified in all cases (since we cannot verify the RT in published data), we suspect that this is the "natural rate" of erroneous entries across typical data sets. This further highlights the need for RT to become a parameter utilized more in compound annotation.

Perhaps even more important than knowing what accuracy can *normally* be expected, are the PIs of the predictions. The PI of each prediction determines how close a RT match needs to be to be "*close enough*" or how different is "*different enough*" to exclude a possible compound structure. In Figure 3E it can be seen that the median width of the PI for the prediction can be as low as 0.1 min with the current database, while in all cases lower than 1.86 min. As above, the relative width of the PIs are given in Figure 3F.

Projection vs Structure-Based Prediction. Most developments in RT prediction have been made in construction of quantitative structure−retention relationship (QSRR) models. It is difficult to compare the accuracy of these models since the same prediction error statistics are not always reported. In the published literature, we typically found models with mean predictions errors about 0.5−2 min equivalent to about 5−15% relative error.[8,10−14,35−40] Better accuracy has been achieved for specific compound groups such as peptides[41,42] and polybrominated diphenyl ethers.[43] A different approach has been published that projects RTs from a system using isocratic conditions to a gradient system.[44,45] This approach is extremely accurate but can only predict the RT of compounds previously characterized in the isocratic system. Solvents, column, and

DOI: 10.1021/acs.analchem.5b02287
Anal. Chem. 2015, 87, 9421−9428

column temperature must also be the same as originally used which severely limit general applicability. In a similar way, "porting" of RTs between similar columns has been shown in ion chromatography systems.[46]

The QSRR models have the advantage of being able to predict the RT of any structure while PredRet can only predict the RT for compounds already in the database. On the other hand, since in the PredRet system we use direct projection between CSs, we are able to achieve much higher accuracy than in QSRR prediction systems.

A method that models classical QSRR parameters in an artificial neural network simultaneously for multiple CSs has also been developed for gas chromatography of polybrominated diphenyl ethers with promising results.[43] The PredRet database could potentially be used for development of such a "hybrid" method for liquid chromatography (LC).

With the current PredRet database, the mean and median prediction error across all predictions were 0.13 and 0.06 min, respectively, equivalent to 2.6 and 1.8%. For the CSs where predictions can be made most precisely, the mean accuracy approaches the batch to batch analytical variance. We therefore believe that PredRet predictions are valuable evidence in the assignment of putatively annotated compounds. PredRet predictions can serve as additional evidence for level 2 and 3 identifications as defined by the metabolomics standards initiative.[47]

Detection of Erroneous Data. Even if there are sufficient data points to build good models, the quality of the model can be compromised by outliers that typically indicate that the RT was reported incorrectly for one of the two CSs in the given model. We found that almost all larger collections of RTs contain such errors and their influence on models can thus compromise the validity of the predictions (see Figure 1B for an example). To eliminate the deleterious influence of such data points, we sought to make the model more robust while keeping the process completely unsupervised. Even though procedures for robust GAM have been proposed,[48] there is to our knowledge no method that allows robust GAM to be combined with a monotonic increasing constraint that is crucial to building sensible models. We therefore used the residuals of an initial GAM to weigh each data point in the penalized constrained least-squares fitting (PCLS) step that succeeds the application of monotonic constraints.

Because there is inherent variation in the data unrelated to outliers, it is preferable to penalize very high residuals (probable outliers) in a nonlinear way such that very aberrant values are disproportionally penalized and we therefore applied a sigmoidal function to the residuals before using them as weights in the subsequent modeling step. The tuning parameters in practical terms result in residuals below 10% not being penalized while residuals above 10% are penalized harshly.

Erroneous data is not only problematic for the modeling step but also in the prediction step if the erroneous entry is used as the basis for a prediction. We therefore implemented a process that tries to detect erroneous entries. In this step we go back to each model and detect compounds for which the data point lies more than twice the PI width from the predicted value. However, this approach cannot determine if it is the entry from one or the other CS used in the model that is erroneous.

Therefore, in cases where the same compound has been recorded in more than two systems, all the models are used to pinpoint which of the entries are wrong. If the compound is only used in a single model (i.e., exists in only two CSs), both entries are assumed to be wrong and the compound will be excluded from predictions. This process insures that no grossly erroneous data is used for predictions if avoidable. If a RT is only reported in one CS, an erroneous entry cannot be detected.

On the PredRet Web site the user will be able to review the list of entries that have been marked as "suspicious". Therefore, the step of detecting "suspicious" entries can also serve as a quality control check for the user.

Ability to Discriminate Isomers. In a liquid chromatography−mass spectrometry (LC−MS) analysis, compounds with different masses can usually be discriminated. However for isomers, masses are identical and RT information is where PredRet, as a tool complementary and orthogonal to mass spectral analysis, provides an added value from its ability to distinguish isomers. In Figure 4, two examples of the prediction output are given for sets of isomers that can be distinguished based on nonoverlapping PIs.

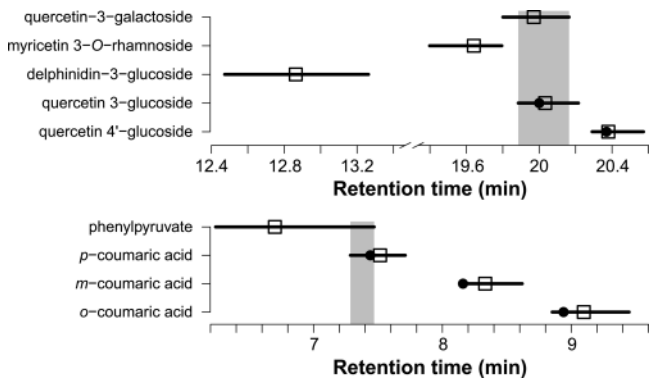


Figure 4. Predicted retention times of two sets of isomers (predicted to different systems). The black lines indicate the prediction intervals. The squares indicate the predicted retention time while the circle indicates the experimental retention time (when available). Gray areas indicate overlapping prediction internals. The predictions indicate that some structurally very similar isomers can be distinguished solely based on the predicted retention times.

In the first example, delphinidin-3-glucoside can be clearly distinguished from the other isomers because the PI of the predicted RT does not overlap with the PI of any of the other isomers. However, for an unknown isomer with an experimental RT in the interval from 19.89 to 20.16 min, it cannot be determined if this unknown is quercetin-3-galactoside or quercetin 3-glucoside since the predicted RT PI in both cases cover this interval. In the second example, all positional isomers of coumaric acid can be distinguished while the more structurally dissimilar phenylpyruvate cannot be distinguished from *p*-coumaric acid in the RT range 7.30−7.47 min.

PredRet can of course only give predictions for isomers for which RT information has been added to the database, and it is therefore always up to the user to consider other plausible isomers. Despite the ability to report PIs and the potential shown above, PredRet should be seen as an exclusion tool, not a tool for confirmation. It serves to funnel the elucidation efforts away from implausible isomers. Classical comparison to authentic standards or in depth structural analysis is still required for confirmation.

The PredRet user should also be aware that some database entries contain ambiguous molecular structures. A common

DOI: 10.1021/acs.analchem.5b02287
Anal. Chem. 2015, 87, 9421−9428

example could be a compound like lysophosphatidylcholine (18:1). There are three structural aspect that are not defined by this name and often unknown when analyzing untargeted or even targeted data: (1) The lipid chain can be in two different positions on the glycerol, sn1 (i.e., LysoPC(18:1/0:0)) and sn2 (i.e., LysoPC(0:0/18:1)). (2) The position of the double bond is not specified. (3) The relative stereochemistry around the double bond is not specified (i.e., cis/trans).

Differences in any of these aspects can lead to different RTs. This is not a limitation of PredRet as such but a limitation of current reporting standards.

Since stereochemistry in general cannot be determined in LC, it is ignored by PredRet. In general this has no influence on the predicted RT since enantiomers have the same RT on nonchiral columns. Diastereoisomers, however, do not necessarily have the same RT. An example is hydroxy-methyl-butyric acid for which there exist a number of structural isomers (PubChem CIDs: 95433, 99823, 160471, 69362, 14081034, 131760, 188979). Most of them have at least one chiral center and therefore exists in enantiomeric forms. 3-Hydroxy-2-methylbutyric acid on the other hand has two chiral centers and thus exists in four forms that are enantiomers and diastereoisomers in pairs of two (PubChem CIDs: 12313369, 11966260 and 11815846, 12313370). The diastereoisomers can have different RTs while the enantiomers cannot. The PI of a prediction for 3-hydroxy-2-methylbutyric acid might therefore not match the experimental RT if the database entry and the experimental data was obtained from different diastereoisomers.

CONCLUSION

The user-friendly Web site http://predret.org allows users to easily upload retention times (RTs) recorded in their chromatographic system (CS) and download predicted RTs for potentially hundreds of compounds in their own system. The prediction system is a novel approach based on direct projection of experimental RTs between many CSs simultaneously. The number of predicted RTs is typically in the hundreds and highly accurate. For prediction systems that do not only model different gradients on the same equipment this allows the prediction of RTs with unprecedented accuracy; on average with an error of 0.13 min equivalent to 2.6% relative error. This is accurate enough to discriminate some structurally very similar isomers. The user can thus prioritize which compounds to run in their own CS and potentially exclude some structures completely with reasonable confidence. In contrast to previous prediction systems, PredRet also provides prediction intervals for each prediction which gives a direct way to decide if an unknown is a likely match or not.

Quantitative structure−retention relationship models do not need the RT of a compound to have ever been experimentally determined to predict the RT and can in many cases therefore predict the RT of any structure once built. On the other hand, these models are typically very complex and require manual optimization and the universality comes at the price of precision. PredRet on the other hand can be used to build precise models without having the expertise to build projection or prediction models. PredRet is, however, limited to compounds for which the RT have previously been determined in a comparable CS. As the database grows it can not only be valuable in compound identification but can also serve as the foundation for further research into the specific effects of different solvents, modifiers, and columns. In addition, the database could be used as a large training data set for improved quantitative structure−retention models.

By using the system, users will add more data from their own CSs to the database. The coverage of different chromatographic conditions and the total number of predictions that can be made will therefore increase significantly as the result of more users. We hope that with support from the scientific community in general, and the metabolomics community in particular, PredRet will be able to grow to cover a large enough number of compounds to be used in the study of many different matrixes. We believe that this tool will greatly help the identification process since compounds that are not compatible with the observed RT can be disregarded. Confirmatory experiments can then be reserved for compounds that could have the observed RT. This will allow researchers to complete the feature annotation and compound identification process in a faster and more rational manner and thus save time and resources, both monetary and environmental.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.5b02287.

Descriptions of all the chromatographic systems in the current database (XLSX)

AUTHOR INFORMATION

Corresponding Author

*Phone: +39 0461-615565. Fax: +39 0461-615200. E-mail: jan.stanstrup@fmach.it, stanstrup@gmail.com.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Dr. Marco Giordan for vital discussions and input on the modelling algorithm used. We thank Dr. Emma Schymanski for extraction of data from MassBank and valuable feedback and testing. We thank Professor Lars Ove Dragsted at the Department of Nutrition, Exercise and Sports at the University of Copenhagen for giving access to his compound library. We thank Dr. Yuji Sawada and Dr. Masami Yokota Hirai at the Metabolomics Research Group at the RIKEN Center for Sustainable Resource Science for providing compound identifiers for their extensive compound library. We thank all the scientists that have contributed to the PredRet database by depositing their data on MetaboLights and MassBank. The work by J.S. was supported by a research grant (Grant VKR023371) from Villum Fonden.

REFERENCES

(1) 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* (7), 703−714.

(2) Wishart, D. S.; Knox, C.; Guo, A. C.; Eisner, R.; Young, N.; Gautam, B.; Hau, D. D.; Psychogios, N.; Dong, E.; Bouatra, S.; Mandal, R.; Sinelnikov, I.; Xia, J.; Jia, L.; Cruz, J. A.; Lim, E.; Sobsey, C. A.; Shrivastava, S.; Huang, P.; Liu, P.; Fang, L.; Peng, J.; Fradette, R.; Cheng, D.; Tzur, D.; Clements, M.; Lewis, A.; De Souza, A.;

DOI: 10.1021/acs.analchem.5b02287
Anal. Chem. 2015, 87, 9421−9428

Zuniga, A.; Dawe, M.; Xiong, Y.; Clive, D.; Greiner, R.; Nazyrova, A.; Shaykhutdinov, R.; Li, L.; Vogel, H. J.; Forsythe, I. *Nucleic Acids Res.* **2009**, *37* (Database), D603−D610.
(3) Gerlich, M.; Neumann, S. *J. Mass Spectrom.* **2013**, *48* (3), 291−298.
(4) Wolf, S.; Schmidt, S.; Müller-Hannemann, M.; Neumann, S. *BMC Bioinf.* **2010**, *11* (1), 148.
(5) Peironcely, J. E.; Rojas-Chertó, M.; Tas, A.; Vreeken, R.; Reijmers, T.; Coulier, L.; Hankemeier, T. *Anal. Chem.* **2013**, *85* (7), 3576−3583.
(6) Krokhin, O. V.; Craig, R.; Spicer, V.; Ens, W.; Standing, K. G.; Beavis, R. C.; Wilkins, J. A. *Mol. Cell. Proteomics* **2004**, *3* (9), 908−919.
(7) Gillet, L. C.; Navarro, P.; Tate, S.; Röst, H.; Selevsek, N.; Reiter, L.; Bonner, R.; Aebersold, R. *Mol. Cell. Proteomics* **2012**, *11* (6), O111.016717.
(8) Menikarachchi, L. C.; Cawley, S.; Hill, D. W.; Hall, L. M.; Hall, L.; Lai, S.; Wilder, J.; Grant, D. F. *Anal. Chem.* **2012**, *84* (21), 9388−9394.
(9) Hall, L. M.; Hall, L. H.; Kertesz, T. M.; Hill, D. W.; Sharp, T. R.; Oblak, E. Z.; Dong, Y. W.; Wishart, D. S.; Chen, M.-H.; Grant, D. F. *J. Chem. Inf. Model.* **2012**, *52* (5), 1222−1237.
(10) Creek, D. J.; Jankevics, A.; Breitling, R.; Watson, D. G.; Barrett, M. P.; Burgess, K. E. V. *Anal. Chem.* **2011**, *83* (22), 8703−8710.
(11) Eugster, P. J.; Boccard, J.; Debrus, B.; Bréant, L.; Wolfender, J.-L.; Martel, S.; Carrupt, P.-A. *Phytochemistry* **2014**, *108*, 196−207.
(12) Cao, M.; Fraser, K.; Huege, J.; Featonby, T.; Rasmussen, S.; Jones, C. *Metabolomics* **2015**, *11* (3), 696−706.
(13) Goryński, K.; Bojko, B.; Nowaczyk, A.; Buciński, A.; Pawliszyn, J.; Kaliszan, R. *Anal. Chim. Acta* **2013**, *797*, 13−19.
(14) Stanstrup, J.; Gerlich, M.; Dragsted, L. O.; Neumann, S. *Anal. Bioanal. Chem.* **2013**, *405* (15), 5037−5048.
(15) Wood, S. N. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2011**, *73* (1), 3−36.
(16) R Development Core Team. *R: A Language and Environment for Statistical Computing*; The R Foundation for Statistical Computing: Vienna, Austria, 2014.
(17) Barri, T.; Holmer-Jensen, J.; Hermansen, K.; Dragsted, L. O. *Anal. Chim. Acta* **2012**, *718*, 47−57.
(18) Theodoridis, G.; Gika, H.; Franceschi, P.; Caputi, L.; Arapitsas, P.; Scholz, M.; Masuero, D.; Wehrens, R.; Vrhovsek, U.; Mattivi, F. *Metabolomics* **2012**, *8* (2), 175−185.
(19) Arapitsas, P.; Speri, G.; Angeli, A.; Perenzoni, D.; Mattivi, F. *Metabolomics* **2014**, *10* (5), 816−832.
(20) Della Corte, A.; Chitarrini, G.; Di Gangi, I. M.; Masuero, D.; Soini, E.; Mattivi, F.; Vrhovsek, U. *Talanta* **2015**, *140*, 52−61.
(21) Strehmel, N.; Böttcher, C.; Schmidt, S.; Scheel, D. *Phytochemistry* **2014**, *108*, 35−46.
(22) Sawada, Y.; Akiyama, K.; Sakata, A.; Kuwahara, A.; Otsuki, H.; Sakurai, T.; Saito, K.; Hirai, M. Y. *Plant Cell Physiol.* **2009**, *50* (1), 37−47.
(23) Schymanski, E. L.; Singer, H. P.; Longrée, P.; Loos, M.; Ruff, M.; Stravs, M. A.; Ripollés Vidal, C.; Hollender, J. *Environ. Sci. Technol.* **2014**, *48* (3), 1811−1818.
(24) Stravs, M. A.; Schymanski, E. L.; Singer, H. P.; Hollender, J. *J. Mass Spectrom.* **2013**, *48* (1), 89−99.
(25) Rasche, F.; Svatoš, A.; Maddula, R. K.; Böttcher, C.; Böcker, S. *Anal. Chem.* **2011**, *83* (4), 1243−1251.
(26) Koulman, A.; Woffendin, G.; Narayana, V. K.; Welchman, H.; Crone, C.; Volmer, D. A. *Rapid Commun. Mass Spectrom.* **2009**, *23* (10), 1411−1418.
(27) Ressom, H. W.; Xiao, J. F.; Tuli, L.; Varghese, R. S.; Zhou, B.; Tsai, T.-H.; Nezami Ranjbar, M. R.; Zhao, Y.; Wang, J.; Di Poto, C.; Cheema, A. K.; Tadesse, M. G.; Goldman, R.; Shetty, K. *Anal. Chim. Acta* **2012**, *743*, 90−100.
(28) Xiao, J. F.; Varghese, R. S.; Zhou, B.; Nezami Ranjbar, M. R.; Zhao, Y.; Tsai, T.-H.; Di Poto, C.; Wang, J.; Goerlitz, D.; Luo, Y.; Cheema, A. K.; Sarhan, N.; Soliman, H.; Tadesse, M. G.; Ziada, D. H.; Ressom, H. W. *J. Proteome Res.* **2012**, *11* (12), 5914−5923.
(29) Roux, A.; Xu, Y.; Heilier, J.-F.; Olivier, M.-F.; Ezan, E.; Tabet, J.-C.; Junot, C. *Anal. Chem.* **2012**, *84* (15), 6429−6437.
(30) Beisken, S.; Earll, M.; Baxter, C.; Portwood, D.; Ament, Z.; Kende, A.; Hodgman, C.; Seymour, G.; Smith, R.; Fraser, P.; Seymour, M.; Salek, R. M.; Steinbeck, C. *Sci. Data* **2014**, *1*, 140029.
(31) Beisken, S.; Earll, M.; Portwood, D.; Seymour, M.; Steinbeck, C. *Mol. Inf.* **2014**, *33* (4), 307−310.
(32) Dal Santo, S.; Tornielli, G. B.; Zenoni, S.; Fasoli, M.; Farina, L.; Anesi, A.; Guzzo, F.; Delledonne, M.; Pezzotti, M. *Genome Biol.* **2013**, *14* (6), r54.
(33) Chaleckis, R.; Ebe, M.; Pluskal, T.; Murakami, I.; Kondoh, H.; Yanagida, M. *Mol. BioSyst.* **2014**, *10* (10), 2538−2551.
(34) Haug, K.; Salek, R. M.; Conesa, P.; Hastings, J.; de Matos, P.; Rijnbeek, M.; Mahendraker, T.; Williams, M.; Neumann, S.; Rocca-Serra, P.; Maguire, E.; Gonzalez-Beltran, A.; Sansone, S.-A.; Griffin, J. L.; Steinbeck, C. *Nucleic Acids Res.* **2013**, *41* (D1), D781−D786.
(35) Munro, K.; Miller, T. H.; Martins, C. P. B.; Edge, A. M.; Cowan, D. A.; Barron, L. P. *J. Chromatogr. A* **2015**, *1396*, 34−44.
(36) D'Archivio, A. A.; Maggi, M. A.; Ruggieri, F. *J. Sep. Sci.* **2010**, *33* (2), 155−166.
(37) Miller, T. H.; Musenga, A.; Cowan, D. A.; Barron, L. P. *Anal. Chem.* **2013**, *85* (21), 10330−10337.
(38) Aicheler, F.; Li, J.; Hoene, M.; Lehmann, R.; Xu, G.; Kohlbacher, O. *Anal. Chem.* **2015**, *87* (15), 7698−7704.
(39) Bade, R.; Bijlsma, L.; Sancho, J. V.; Hernández, F. *Talanta* **2015**, *139*, 143−149.
(40) Kern, S.; Fenner, K.; Singer, H. P.; Schwarzenbach, R. P.; Hollender, J. *Environ. Sci. Technol.* **2009**, *43* (18), 7039−7046.
(41) Escher, C.; Reiter, L.; MacLean, B.; Ossola, R.; Herzog, F.; Chilton, J.; MacCoss, M. J.; Rinner, O. *Proteomics* **2012**, *12* (8), 1111−1121.
(42) Spicer, V.; Yamchuk, A.; Cortens, J.; Sousa, S.; Ens, W.; Standing, K. G.; Wilkins, J. A.; Krokhin, O. V. *Anal. Chem.* **2007**, *79* (22), 8762−8768.
(43) D'Archivio, A. A.; Giannitto, A.; Maggi, M. A. *J. Chromatogr. A* **2013**, *1298*, 118−131.
(44) Boswell, P. G.; Schellenberg, J. R.; Carr, P. W.; Cohen, J. D.; Hegeman, A. D. *J. Chromatogr. A* **2011**, *1218* (38), 6742−6749.
(45) Boswell, P. G.; Schellenberg, J. R.; Carr, P. W.; Cohen, J. D.; Hegeman, A. D. *J. Chromatogr. A* **2011**, *1218* (38), 6732−6741.
(46) Ng, B. K.; Shellie, R. A.; Dicinoski, G. W.; Bloomfield, C.; Liu, Y.; Pohl, C. A.; Haddad, P. R. *J. Chromatogr. A* **2011**, *1218* (32), 5512−5519.
(47) Sumner, L.; Amberg, A.; Barrett, D.; Beale, M.; Beger, R.; Daykin, C.; Fan, T.; Fiehn, O.; Goodacre, R.; Griffin, J.; Hankemeier, T.; Hardy, N.; Harnly, J.; Higashi, R.; Kopka, J.; Lane, A.; Lindon, J.; Marriott, P.; Nicholls, A.; Reily, M.; Thaden, J.; Viant, M. *Metabolomics* **2007**, *3* (3), 211−221.
(48) Croux, C.; Gijbels, I.; Prosdocimi, I. *Biometrics* **2012**, *68* (1), 31−44.

DOI: 10.1021/acs.analchem.5b02287
Anal. Chem. 2015, 87, 9421−9428