analytical chemistry

Article

pubs.acs.org/ac

LipidCCS: Prediction of Collision Cross-Section Values for Lipids with High Precision To Support Ion Mobility−Mass Spectrometry-Based Lipidomics

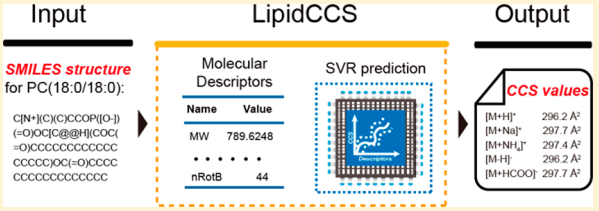
Zhiwei Zhou,[†,‡] Jia Tu,[†,‡] Xin Xiong,[†] Xiaotao Shen,[†,‡] and Zheng-Jiang Zhu*[,†]

[†]Interdisciplinary Research Center on Biology and Chemistry, and Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai 200032, P. R. China

[‡]University of Chinese Academy of Sciences, Beijing 100049, P. R. China

Supporting Information

ABSTRACT: The use of collision cross-section (CCS) values derived from ion mobility−mass spectrometry (IM−MS) has been proven to facilitate lipid identifications. Its utility is restricted by the limited availability of CCS values. Recently, the machine-learning algorithm-based prediction (e.g., MetCCS) is reported to generate CCS values in a large-scale. However, the prediction precision is not sufficient to differentiate lipids due to their high structural similarities and subtle differences on CCS values. To address this challenge, we developed a new approach, namely, LipidCCS, to precisely predict lipid CCS values. In LipidCCS, a set of molecular descriptors were optimized using bioinformatic approaches to comprehensively describe the subtle structure differences for lipids. The use of optimized molecular descriptors together with a large set of standard CCS values for lipids (458 in total) to build the prediction model significantly improved the precision. The prediction precision of LipidCCS was externally validated with median relative errors (MRE) of ~1% using independent data sets across different instruments (Agilent DTIM-MS and Waters TWIM-MS) and laboratories. We also demonstrated that the improved precision in the predicted LipidCCS database (15 646 lipids and 63 434 CCS values in total) could effectively reduce false-positive identifications of lipids. Common users can freely access our LipidCCS web server for the following: (1) the prediction of lipid CCS values directly from SMILES structure; (2) database search; and (3) lipid match and identification. We believe LipidCCS will be a valuable tool to support IM−MS-based lipidomics. The web server is freely available on the Internet (http://www.metabolomics-shanghai.org/LipidCCS/).



Lipids are the major components of cell membranes and play many vital roles in cells such as energy storage, cell signaling, and interactions with proteins.[1−3] Dysresgulation of lipid homeostasis has been validated to associate with several major human diseases including diabetes, obesity, cardiovascular disease, and Alzherimer's disease.[4−6] Lipids have the enormous structural diversity varying from the head groups, the position and number of double bonds, and the composition of acyl chain. In total, about 180 000 lipid species were estimated in the lipidome.[7,8] Therefore, the comprehensive analysis of lipids requires powerful analytical techniques to separate and identify them. Mass spectrometry (MS) is one of the most important techniques for lipid analysis with high sensitivity and specificity, such as shotgun lipidomics and LC-MS-based lipidomics.[9−11]

Recently, ion mobility−mass spectrometry (IM−MS) has showed great application potential for lipidomics.[12−16] In IM−MS, structurally different lipid ions can be separated rapidly in gas phase as a result of many collisions occurring between lipid ions and inert buffer gas (e.g., nitrogen) under an electric field.[17,18] The previous studies have proven that the use of IM−MS provided enhanced peak capacity,[19,20] improvement of signal-to-noise,[21] and separation of lipid isomers.[22−26] Lipids separated by ion mobility can be further identified using collision cross-section (CCS) values.[27−29] The CCS value derived from IM−MS is a unique physicochemical property of the lipid, and shows high reproducibility across different laboratories and instruments.[29] The addition of CCS values to lipidomics workflow significantly improved the accuracy of lipid analysis.[25,27,28] However, the limited number of available CCS values for lipids largely restricts the application of IM−MS in lipidomics.

Two common strategies to generate CCS values are the experimental measurement of chemical standards and the theoretical calculation.[24,27,30] For lipids, a very limited number of standards are available to obtain the experimental CCS values. Theoretical calculation of CCS values was widely employed for small molecules[31] and peptides.[32] However, the conformations of lipid ions in the gas phase are very challenging to be simulated due to the complexity of the lipid structures.[13] So far, very few theoretical CCS values of lipids were reported.

Received: July 6, 2017
Accepted: July 31, 2017
Published: August 1, 2017

DOI: 10.1021/acs.analchem.7b02625
Anal. Chem. 2017, 89, 9559−9566

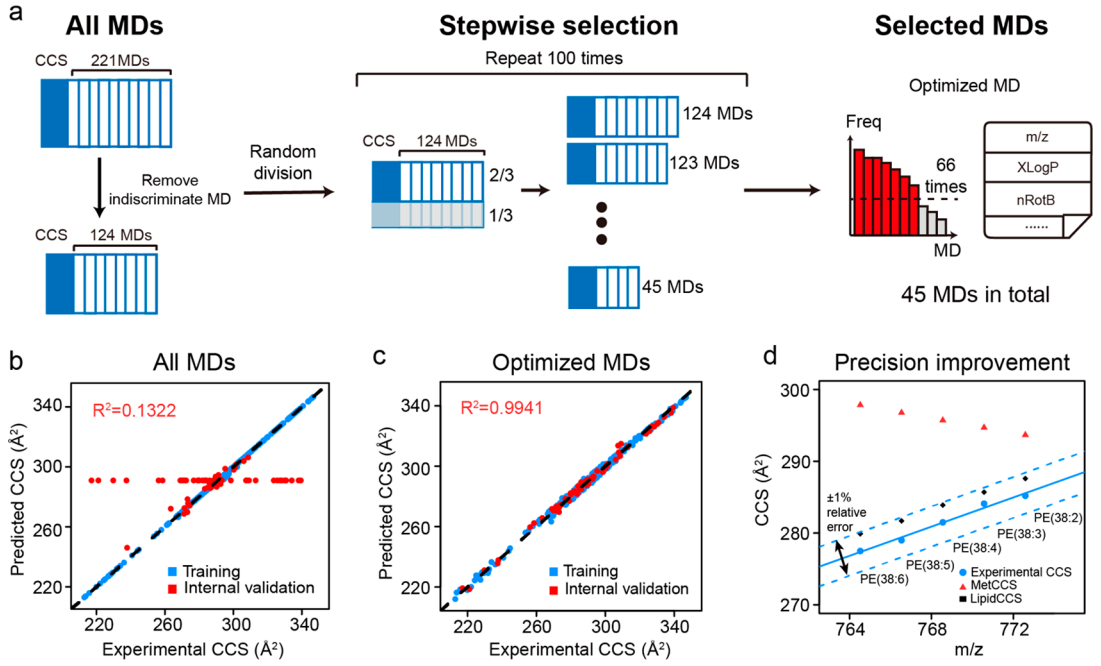


Figure 1. (a) Optimization and selection molecular descriptors to effectively characterize lipid structures. MD represents molecular descriptor; (b,c) prediction performance evaluation using all 221 molecular descriptors (b) or 45 optimized molecular descriptors (c); (d) the comparison of the prediction precision between LipidCCS and MetCCS.

Instead, our previous work successfully demonstrated the use of a machine-learning algorithm-based approach, namely, MetCCS, to predict the CCS values of metabolites using 14 molecular descriptors (MDs).[33,34] The prediction precision has been externally validated with a median relative error (MRE) of ~3%, and the use of predicted CCS values improved metabolite identification accuracy. However, the prediction precision of MetCCS is not enough for lipid identifications since most lipids have very similar structures and thus very subtle differences on CCS value. For example, isomeric lipids such as LPE(18:0) (CCS, 221.5 Å^2) and LPC(15:0) (CCS, 225.0 Å^2) have a small difference of 1.6% on CCS values. In addition, MetCCS requires the users to self-calculate and manually input 14 different molecular descriptors for prediction.[34] This process is tedious and low efficient, especially for the compounds which are not contained in the MetCCS database.

To address these challenges, we have to develop a new prediction approach that is suitable for lipids with a significantly improved prediction precision. To do so, we first developed a bioinformatic approach to optimize the selection of molecular descriptors for prediction. A combination of molecular descriptors was generated to precisely describe the subtle structure differences for lipids, therefore significantly improving the prediction precision. Specifically, we optimized 45 and 66 out of the 221 calculated molecular descriptors to build the support vector regression (SVR)-based prediction models for positive and negative ions, respectively. Second, we experimentally measured a large set of standard CCS values for lipids (458 in total) to serve as the training data set. Then, our new prediction approach was externally validated using the independent external data sets, which have a precision as high as 1% (MRE). The predicted lipid CCS values from our new approach fit very well with experimental measurements from different IM−MS instruments (both Agilent DTIM-MS and Waters TWIM-MS). We also demonstrated that the improved precision in predicted CCS values could effectively remove false positives in lipid identification and facilitate the IM−MS-based lipidomics workflow. Finally, to facilitate the use of CCS values for lipidomics, we created the large-scale database, namely, LipidCCS, and web server with a user-friendly graphical interface. Common users can use the web server for the following: (1) the prediction of lipid CCS values directly from SMILES structure; (2) LipidCCS database search; and (3) lipid match and identification. The web server is freely available on the Internet (http://www.metabolomics-shanghai.org/LipidCCS/).

EXPERIMENTAL SECTION

All experiments were performed using Agilent DTIM-QTOF-MS 6560 coupled with an Agilent UHPLC 1290 (Agilent Technologies, U.S.A.). All CCS values were obtained using nitrogen gas and the single-field calibration method as previously reported.[33] Experimental details about the lipid standards, lipid extraction protocol, and instrument parameters are provided in Supporting Information.

Measurements of Experimental CCS Values of Lipids. Lipid standard mixtures (Table S1) were purchased from Avanti, and they were used to measure experimental CCS values of glycerophospholipids (PC, PE, PG, PS, PI, PA, LPC, LPE, LPI, LPS) and sphingolipids (SM, Cermide, GlcCer, ST). Since no glycerolipid standard mixtures were commercially available, lipid extraction from mouse heart tissue was used to measure the experimental CCS values of triacylglycerol (TG) and diacylglycerol (DG). In order to obtain reliable CCS values, each lipid sample was independently analyzed three times over 4 months, and CCS values then were averaged as the experimental CCS values. Before each measurement, both Agilent tuning mix solution and quality control (QC) sample were injected and used for CCS calibration and instrument reproducibility evaluation, respectively. The QC sample consists of three lipid standards: PE(17:0/17:0) (10 μg/mL), lysoPC(16:0/0:0) (0.5 μg/mL), and TG(17:1/17:1/17:1) (0.5

DOI: 10.1021/acs.analchem.7b02625
Anal. Chem. 2017, 89, 9559−9566

μg/mL) in MeOH. The relative standard deviation (RSD) and absolute deviation of lipids in QC samples was less than 0.2% and 0.3% over 4 months, respectively, demonstrating the excellent stability and accuracy of IM−MS instrument (Figure S1, Tables S2 and S3 in Supporting Information). Similarly, the averaged RSD of experimental lipid CCS values were 0.12% and 0.16% in positive and negative modes, respectively. The raw data was processed using Agilent IM−MS Browser for peak detection, calculation of CCS values and MS/MS extraction.[35] Please note that CCS values were calculated using 6 calibrants in Agilent tune mix. Both LipidCCS and MetCCS used B.07.01 version of IM−MS Browser. All detected lipids were first matched against with the LIPID MAPS Structure Database (LMSD) with a tolerance of 10 ppm. Then, the lipid identifications were further manually assigned and validated according to their MS/MS spectra and retention time. Although the previous publications have reported that the formation of adducts helps to distinguish the lipid isomers (e.g., cis/trans isomers),[24,25] the difference of their CCS values is too small (usually less than 1%) to be resolved by most commercial IM−MS instruments (Figure S2). Therefore, in this work, lipid isomers such as PC(18:1(11Z)/16:0) and PC(16:0/18:1(6E)) were not differentiated in lipid identification.

Development of LipidCCS Predictor. LipidCCS Predictor is developed in R programming environment (version 3.3.2). LipidCCS Predictor accepts the input of the SMILES structure for prediction. The program first calculates the molecular descriptors of the lipid using the R package "rcdk" (version 3.3.8) from its SMILES structure. Then, a set of molecular descriptors are selected and input into the SVR-based prediction model. For each lipid, CCS values from 5 common adducts are finally reported.

Here, support vector regression (via the R package "e1071") is used to build the prediction model (Scheme S1). To build the model, 329 and 129 experimental CCS values of lipids in positive and negative modes, respectively, were used as a training data set. Similar to MetCCS, two parameters in SVR, cost of constraints violation (C) and gamma (γ), were first optimized to achieve the best prediction performance. Then, the LipidCCS prediction models were built using the training data set and molecular descriptors (see details in Supporting Information). The prediction methods were separately developed with the same workflow for different ionization modes.

Optimization of Molecular Descriptors. Molecular descriptors are a series of numeric values to characterize the structural and physiochemical properties of one molecule.[36] For each SMILES structure of lipids, a total of 221 molecular descriptors can be calculated by the "rcdk" package. The optimization and selection of MDs were performed in the following steps (Figure 1a). First, 97 out of 221 MDs were indiscriminate among the lipids in the training data set and thereby removed. The remaining 124 MDs were stepwisely optimized using the training data set. Briefly, 2/3 of the CCS values in the training data set were randomly chosen and used to build a regression model using the 124 MDs. Then the MD with the least contribution to the regression model was stepwisely removed, and the Akaike information criterion (AIC) value was calculated to evaluate the model quality relative to the previous models. The process to remove the MD one by one continues until a minimal AIC value was obtained (Scheme S2 in Supporting Information). The final combination of MDs for the prediction model was recorded. More details about the stepwise optimization were provided in Supporting Information. To ensure the reliability on MD optimization process, this process was repeated 100 times. Then MDs with frequencies higher than 2/3 were selected (Figure S3). Using this strategy, 45 and 66 molecular descriptors were finally selected for the prediction of lipid CCS values in positive and negative modes, respectively (see Table S4 and S5 in Supporting Information for the list of MDs).

RESULTS AND DISCUSSION

Development of the Prediction Method with High Precision. In this work, a new approach, namely, LipidCCS Predictor, was developed for the precise prediction of lipid CCS values. One can simply input the SMILES structure of the lipid to LipidCCS Predictor, and the software automatically calculates and outputs five CCS values in positive and negative modes. The prediction takes only several seconds to complete. The SMILES structures of lipids can be obtained from the LIPID MAPS Structure Database (LMSD) or other similar databases.[37] To develop the LipidCCS Predictor, we first measured 329 and 129 lipid CCS values in positive and negative modes, respectively, to serve as a training data set (Excel Files 01 and 02 in Supporting Information). Here, all identified lipids were singly charged, and therefore, LipidCCS only supports the prediction of CCS values for the singly charged lipids. The whole experimental data set has a broad coverage and diversity, and includes three major categories of lipids (glycerophospholipids (GP), sphingolipids (SP), and glycerolipids (GL)), and five common adducts (Figure S4). To the best of our knowledge, it is the largest experimental data set of lipid CCS values in IM−MS. Most CCS values of glycerolipids (TG and DG) were first reported.

Another novel aspect of LipidCCS compared to our previous reported MetCCS is the optimization of molecular descriptors to effectively differentiate lipid structures (Figure 1a). Most lipids have very similar physicochemical properties, for example, glycerophospholipids having subtle differences on the structure of head groups, the length of acyl chain, and the number of double bonds. The selection of suitable molecular descriptors to effectively distinguish different lipids presents a significant challenge. We first constructed an SVR prediction model using all the calculated 221 molecular descriptors. The results demonstrated that the generated model has a very poor prediction precision with a fitting value R^2 of 0.1322 in the internal validation (red square in the Figure 1b). Therefore, a stepwise optimization of molecular descriptors was performed. Finally, 45 and 66 molecular descriptors were selected to build SVR-based prediction model in positive and negative modes, respectively. Among these molecular descriptors, the m/z value has the best correlation with the CCS value and is the most important molecular descriptor. In addition, many topological descriptors derived from molecular graph were observed to be important for the prediction (Excel file 09 in Supporting Information). Using the optimized 45 molecular descriptors in positive mode, a good prediction precision was obtained with an R^2 value of 0.9941 (Figure 1c). As a comparison, we randomly selected 45 molecular descriptors to build the prediction model, which also demonstrated an overfitting effect similar to the use of all 221 molecular descriptors (Figure S5). Additionally, taking five PE lipids as examples, the prediction precision of LipidCCS is significantly improved compared to our previously reported MetCCS (Figure 1d). Similar results

DOI: 10.1021/acs.analchem.7b02625
Anal. Chem. 2017, 89, 9559−9566

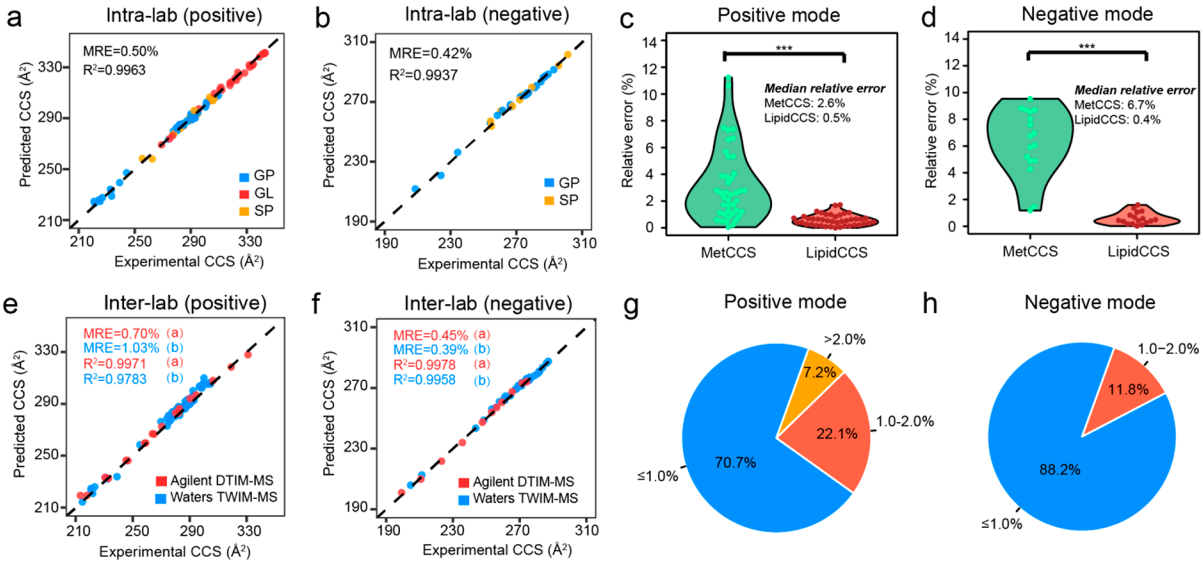


Figure 2. External validations of prediction precision. (a,b) Correlation between predicted and experimental CCS values from the intralab validations in positive (a) and negative ionization modes (b); (c,d) the comparison of prediction precision between LipidCCS and MetCCS in positive (c) and negative modes (d). Wilcoxon rank sum test was used for the statistical tests; (e,f) correlation between predicted and experimental CCS values from the interlab validations in positive (e) and negative ionization modes (f). Letters "a" and "b" in panels e and f indicate these values calculated from Agilent DTIM-MS and Water TWIM-MS data sets, respectively; (g,h) the percentages of CCS values within certain relative errors in positive (g) and negative (h) modes.

were also obtained for selected molecular descriptors in negative mode (Figure S6).

Finally, with the optimized molecular descriptors, the SVR-based prediction method was built based on the whole training data set. The median relative errors were 0.49% and 0.47%, and R^2 values were 0.9959 and 0.9964 in positive and negative modes, respectively (Table S6 in Supporting Information). In addition, the q^2 values from 10-fold cross-validation in method development are as high as 0.9950 and 0.9957 in positive and negative modes, respectively, indicating that the prediction method was not overfitting. Overall, these results demonstrated that the newly developed LipidCCS Predictor has an excellent capability for the precise prediction of lipid CCS values. However, LipidCCS cannot accurately predict the CCS values of isomeric lipids that differ in position or geometry (cis/trans) yet, because there are no isomeric lipid standards in the training data set and the used IM−MS instrument has insufficient resolution to separate those isomers.

External Validation of High Prediction Precision Using Independent Data Sets. We further externally validated the prediction precision of LipidCCS Predictor using four independent experimental data sets of lipid CCS values. These data sets were from different laboratories and instrument platforms (Agilent DTIM-MS and Waters TWIM-MS), and they were generated using different measurement methods (Excel Files 03−06 in Supporting Information).[27,28,38] Please note that none of these lipid CCS values were included in our training data set, or participated in the development of LipidCCS Predictor. The first external validation data set is the intralab data set, including 82 and 32 lipid CCS values in positive and negative modes, respectively. The data was generated from the same Agilent DTIM-MS instrument in our lab using the single-field method. LipidCCS Predictor gives an excellent prediction precision in both positive and negative ionization modes (Figure 2a,b). The R^2 values of regression curves were 0.9963 and 0.9937, and the median relative errors were 0.50% and 0.42% in positive and negative modes, respectively. Using the intralab data set, we systematically compared the prediction precision between LipidCCS and MetCCS. The results demonstrated that MetCCS has a worse prediction precision, and the MRE values were 2.6% and 6.7% in positive and negative modes, respectively (Figure 2c,d, and Excel Files 07−08 in Supporting Information).

Next, we performed the external interlab validation using three interlab data sets. Recently, Hines et al. reported the use of the multifield method in Agilent DTIM-MS to measure the CCS values of 10 PCs and 14 PEs with the acyl side chain ranging from 6 to 24 carbons.[38] We compared these CCS values that were not included in our training data set to ones generated from our LipidCCS Predictor (Figure 2e,f, red dots). The predicted CCS values fitted well with the experiment CCS values, with an R^2 value larger than 0.99 and MRE value less than 1%. The results proved that LipidCCS Predictor gives reliable and precise prediction of lipid CCS values with different acyl lengths, and these predicted values are consistent with the experimental ones derived from the multifield method in Agilent DTIM-MS.

Waters TWIM-MS is another widely used IM−MS platform for lipidomics. In the TWIM-MS platform, CCS values of small molecules are commonly measured using polyalanine (polyAla) as the CCS calibrants.[39] Recently, Hines et al. found that the use of structurally similar lipids as the calibrants in TWIM-MS generates more precise experimental CCS values for lipids.[38] Here, we first compared the lipid CCS values (65 in total) derived from TWIM-MS instrument using the lipid calibrants to ones generated from LipidCCS Predictor (Figure 2e,f, blue dots).[28] A good prediction precision was observed that the R^2 values of regression curves were 0.9783 and 0.9958 in positive and negative modes, respectively. Median relative errors were 1.03% and 0.39% in positive and negative modes, respectively. In addition, we also compared the lipid CCS values derived from TWIM-MS instrument[27] using polyAla calibrants to our prediction results (Figure S7). In contrast, system errors were clearly observed between the two data sets. Taken together, this

DOI: 10.1021/acs.analchem.7b02625
Anal. Chem. 2017, 89, 9559−9566

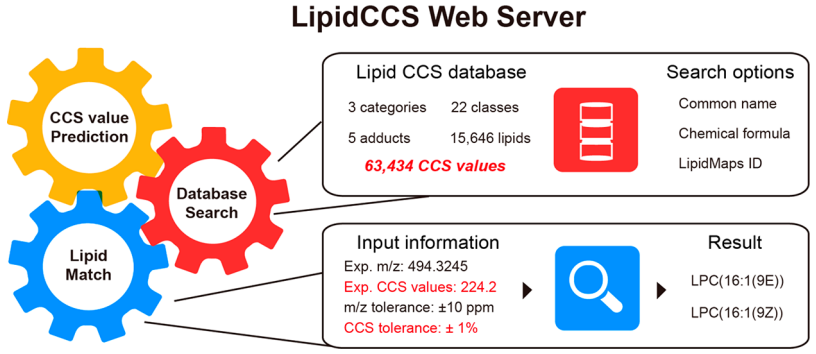


Figure 3. LipidCCS web server including three major functions: (1) LipidCCS database search, (2) prediction of lipid CCS values using SMILES structure, and (3) lipid match and identification.

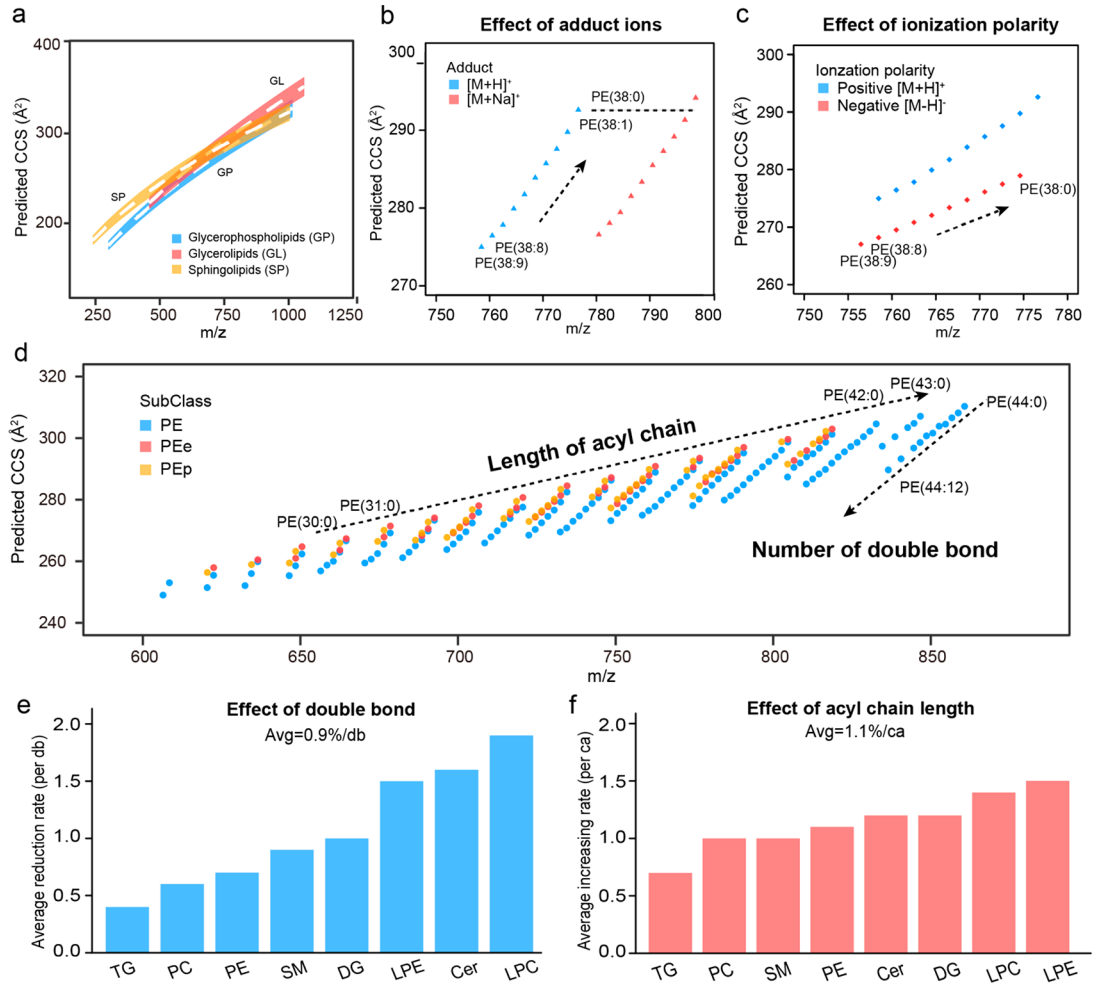


Figure 4. Investigating the relationship between lipid structures and CCS values. (a) Conformational ordering of different lipid categories fitted by power functions. The inclusion band represents deviation (±3%) from the best fit line; (b–d) 2D plot (*m/z vs* CCS) to investigate the effect of adduct ions (b), ionization polarity (c), double bond number and acyl chain length (d) to the CCS values for PE lipids; (e,f) quantitative evaluation for the effect of the number double bond and acyl chain length to the CCS values for different lipid classes. The abbreviations in legend: "db" refers to the number of double bond; "ca" refers to the number of carbon atom. All quantitative results were calculated from $[M+H]^+$ ions.

confirmed that lipid calibrants in TWIM-MS can generate more precise lipid CCS values, and these values are consistent with our LipidCCS Predictor.

In summary, the prediction precision of our newly developed LipidCCS Predictor has remarkably improved to 1% (MRE). It generates lipid CCS values matched very well with multiple external validation data sets derived from different ion mobility techniques (DTIM-MS and TWIM-MS) and different laboratories. More than 92% of predicted lipid CCS values had less than 2% of relative errors compared to external experimental ones in both positive and negative ionization modes (Figure 2g,h). The results also demonstrated the precision improvement of lipid CCS value prediction compared to our previously reported MetCCS Predictor.

Creation of Large-Scale LipidCCS Database and Web Server. The addition of IM separation into lipidomics

DOI: 10.1021/acs.analchem.7b02625
Anal. Chem. 2017, 89, 9559–9566

effectively improves the selectivity, but this workflow suffers from the limit number of available CCS values.[12,13,18] Here, to facilitate the use of lipid CCS values for lipidomics, we generated a large-scale LipidCCS database and an easy-to-use web server, and three major functions were designed in LipidCCS (Figure 3):

Prediction of Lipid CCS Values. In the LipidCCS Predictor page, users can rapidly predict lipid CCS values using lipid SMILES structures within several seconds. The SMILES structure can be obtained from ChemDraw or LipidMaps Database. This function enables the convenient prediction of the CCS values for novel lipids that are not included in the LipidCCS database.

LipidCCS Database Search. We predicted 15 646 common lipids from LMSD, the largest publicly available lipid database, covering 3 major categories and 22 common lipid classes, including glycerophospholipids, glycerolipids and sphingolipids (Figure S8). For each lipid, five common ion adducts were predicted, including $[M+H]^+$, $[M+Na]^+$, and $[M+NH_4]^+$ for positive ionization, and $[M-H]^-$, $[M+HCOO]^-$ for negative ionization. Finally, a total of 63 434 predicted CCS values of lipids were deposited into the LipidCCS database, dramatically improved compared to MetCCS (Figure S9). To the best of our knowledge, this is the only available large-scale CCS database for lipids. It should be noted that CCS values of glycerolipids were predicted only in positive mode due to their poor ionization efficiency in negative mode. The LipidCCS allows users to search the lipid CCS values using LipidMaps ID, chemical formula, and common name. It also supports batch search function with a maximum of 100 query lines per request.

Lipid Match and Identification. In LipidCCS, lipid match function was designed to help users to identify lipids through matching experimentally measured m/z and CCS values with the defined tolerances. For example, one can import the experimental m/z value of 494.3245 and CCS value of 224.2, and define the tolerances as 10 ppm and 1% for m/z and CCS match, respectively. Two lipid hits, LPC(16:1(9E)) and LPC(16:1(9Z)), are returned immediately through the match with our LipidCCS database.

Relationship between Lipid Structure and CCS Values. The CCS values derived from IM−MS can reflect the structural information on lipids.[29] Several previous studies had reported that CCS values of different lipid classes have clear trend lines in the 2D plots of m/z and CCS values.[27,40,41] Here, we systematically investigated the relationship between lipid structures and CCS values using precisely predicted CCS values of lipids. First, the trend lines of each lipid category were obtained by the power fitting (Table S7).[42] As shown in Figure 4a, the CCS values of sphingolipids were generally larger than glycerophospholipids in lower mass range (below 700 Da). For example, Cer(d18:1/24:0) and PE(14:0/15:0) have similar exact mass (649.6373 Da vs 649.4683 Da), but the CCS value of Cer(d18:1/24:0) is obviously larger than PE(14:0/15:0) (290.0 $Å^2$ vs 262.4 $Å^2$, 10.5% relative error). It may be caused by the presence of longer acyl groups in sphingolipids than glycerophospholipids. Triglycerolipids generally have larger CCS values than glycerophospholipids and sphingolipids especially due to the higher number of acyl groups. More details about CCS values differences between lipid subclasses were provided in Supporting Information (Figure S10 and Table S8 in Supporting Information). For the effect of headgroup to CCS values of glycerophospholipids, the slopes of trend lines decrease in the order as PC > PA > PE ≈ PG > PS > PI. Similarly, the decreased order of trend lines for lysoglycerophospholipids was observed as LPC > LPA > LPG > LPI > LPE > LPS. Finally, the decreased orders of trend lines, Cer > SM > GlcCer > ST > Sphingosine and TG > DG, were observed for sphingolipids and glycerolipids, respectively.

With the precisely predicted CCS values for a variety of lipid species, it is feasible to evaluate the effect of lipid structures to their CCS values. Taking PE lipids as examples, we discovered that ionization adducts (e.g., $[M+H]^+$, $[M+Na]^+$, $[M-H]^-$) influence the CCS values in different degrees, and the addition of double bonds into the lipid structure causes the reduction of CCS values (Figure 4b−d). Then, we further quantitatively evaluated the effect of the double bonds in acyl chain to lipid CCS values (Figure 4e). For different lipids, the CCS values were reduced at a rate from 0.4% to 1.9% for one addition of a double bond in acyl chain. This phenomenon may be explained by unsaturated double bonds caused the chain to bend to shorten the molecules.[40,41] Interestingly, the decrement of CCS values in TG (0.4%) was obviously lower than other lipids, which may be caused by complex conformations of TG in gas phase lessening the effect of double bond to CCS values (Figure S10). The length of acyl chain is another major factor to affect CCS values. For most lipids, the CCS values increase with the length of acyl chain increasing but in different degrees. Generally, the extending of one carbon in acyl chains causes the increase of 0.7−1.5% of CCS values for different lipid classes (Figure 4f). In summary, the effect of lipid structure to CCS values (or trend lines) varies from lipid classes. Most importantly, our qualitative analysis results from LipidCCS predictor were well consisted with the previous reports,[27,28,40,41] which further validated the high precision of our new approach for lipid CCS value prediction.

Use of LipidCCS to Support IM−MS-Based Lipidomoics. The CCS values had been validated as an additional physicochemical property to improve the confidence of lipid identification.[27,28] To validate, we first analyzed the purified mixture of PC standards using LC−IM−MS. A total of 100 PC lipids were identified through m/z match against LipidCCS. Using the combination of both m/z and CCS matches, only 81 PC lipids were identified (Figure 5a). Here, m/z and CCS match tolerances were set as 10 ppm and 1%, respectively. Then, we manually checked the 19 filtered lipids, and we found that 17 lipids (89.5%) were correctly filtered with only two false negatives (Figure 5b). Therefore, the addition of the predicted CCS values to lipid identification can effectively remove the false positive identifications.

The predicted LipidCCS database was further used to improve the lipid identification accuracy in complex biological samples including human plasma, human 239T cell, mouse brain, and heart tissues (Figure 5c and Figure S11). Taking human plasma sample as an example, a total of 2284 features were detected, and 954 features were identified through using m/z match using a tolerance of 10 ppm. Similarly, only 496 features were remained after the addition of the CCS match with a tolerance of 1%. In addition, 74.8% features had less lipid candidates using both m/z and CCS matches compared to m/z match only, supporting that the use of CCS values effectively reduced the candidate numbers for untargeted lipidomics (Figure 5d). More specifically, an average of 12.5 potential candidates was reduced after the addition of the CCS value match with 1% tolerance (Figure 5e). For example, feature M544T73CCS230 (m/z 544.3407 Da; RT 73 s; CCS 230.1 $Å^2$) had 13 potential lipid candidates using m/z match only (Figure

DOI: 10.1021/acs.analchem.7b02625
Anal. Chem. 2017, 89, 9559−9566

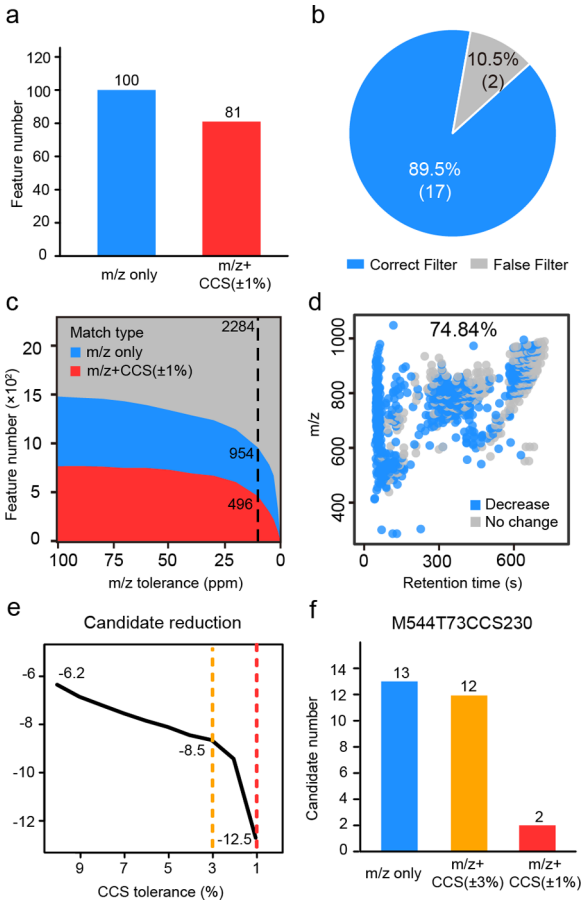


Figure 5. (a,b) Use of LipidCCS to reduce the false positive identifications of lipids in the purified mixture of PC standards: (a) bar plot of identified PC lipids using two match methods; (b) manual verification of the accuracy of the filtered lipids. (c–f) Use of LipidCCS to support IM–MS-based untargeted lipidomics: (c) in human plasma sample, statistics of identified lipid features using two match methods; (d) dot plot for the feature distribution with the decreased candidates after the additional CCS match; (e) statistics of the averaged numbers of decreased candidates after the additional CCS match; (f) number of lipid hits for the feature (M544T73CCS230).

5f). After the addition of the CCS match, only 2 candidates remained, namely, PC(20:4(5Z,8Z,11Z,14Z)/0:0) and PC-(20:4(8Z,11Z,14Z,17Z)/0:0). This feature can be further assigned as LPC(20:4), because two isomers are difficult to be resolved by IM–MS. The identification result was also confirmed using MS/MS spectrum against with in-house predicted MS/MS database of lipids (Figure S12). As a comparison, 12 potential candidate lipids were given using a CCS match tolerance of 3% (similar to MetCCS prediction precision). Overall, these results demonstrated that the high precision predicted LipidCCS database facilitated lipid identification for IM–MS-based untargeted lipidomics.

CONCLUSION

In conclusion, we developed a new approach, namely, LipidCCS Predictor, to precisely predict CCS values of lipids. A large-scale and diverse data set including 458 experimental lipid CCS values was used as the training data set. A novel method was developed to optimize molecular descriptors and significantly improved the prediction precision. As a result, the prediction precision of LipidCCS Predictor was externally validated with a median relative error (MRE) of ~1% across different instruments and laboratories using independent external data sets. The predicted CCS values from LipidCCS Predictor match very well with ones derived from Agilent DTIM-MS using both the single-field and multifield methods, as well as Waters TWIM-MS using the lipid calibrants, but they do not match well with the ones derived from TWIM-MS using polyAla as the calibrants. Then, the prediction approach is used to generate the large-scale CCS value database, namely, LipidCCS database, which contains 15 646 lipids and 63 434 CCS values in total. To facilitate the use of LipidCCS for lipidomics, a web server with user-friendly interface was created. Utilizing the LipidCCS database, we further systematically investigated the relationship between lipid structure and their CCS values, and we validated the use of LipidCCS database to effectively reduce false positive identifications of lipids in untargeted lipidomics. In the future, the combination of the predicted CCS values with other common properties (e.g., MS/MS, retention time) could further increase the accuracy of lipid identification. Therefore, we believe that the LipidCCS will be a valuable tool to support IM–MS-based lipidomics.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.7b02625.

Supplemental experiment details; Tables S1–S8, Schemes S1–S2, and Figures S1–S12 (PDF)

Excel files for experimental CCS values of lipids (ZIP)

AUTHOR INFORMATION

Corresponding Author

*E-mail: jiangzhu@sioc.ac.cn. Phone: 86-21-68582296.

ORCID

Zheng-Jiang Zhu: 0000-0002-3272-3567

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank John Fjeldsted and Man-Yu Zhang from Agilent Technologies for the help on IM–MS instrumentation. The work is financially supported by National Natural Science Foundation of China (Grant No. 21575151) and Agilent Technologies Thought Leader Award. Z.-J.Z. is supported by Thousand Youth Talents Program from Government of China.

REFERENCES

(1) Wenk, M. R. *Cell* **2010**, *143*, 888–895.

(2) Shevchenko, A.; Simons, K. *Nat. Rev. Mol. Cell Biol.* **2010**, *11*, 593–598.

(3) Han, X. *Nat. Rev. Endocrinol.* **2016**, *12*, 668–679.

(4) Quehenberger, O.; Dennis, E. A. *N. Engl. J. Med.* **2011**, *365*, 1812–1823.

(5) Mapstone, M.; Cheema, A. K.; Fiandaca, M. S.; Zhong, X.; Mhyre, T. R.; MacArthur, L. H.; Hall, W. J.; Fisher, S. G.; Peterson, D. R.; Haley, J. M.; Nazar, M. D.; Rich, S. A.; Berlau, D. J.; Peltz, C. B.; Tan, M. T.; Kawas, C. H.; Federoff, H. J. *Nat. Med.* **2014**, *20*, 415–418.

(6) Perry, R. J.; Samuel, V. T.; Petersen, K. F.; Shulman, G. I. *Nature* **2014**, *510*, 84–91.

DOI: 10.1021/acs.analchem.7b02625
Anal. Chem. 2017, 89, 9559–9566

(7) Yetukuri, L.; Ekroos, K.; Vidal-Puig, A.; Oresic, M. *Mol. BioSyst.* **2008**, *4*, 121−127.
(8) Lam, S. M.; Shui, G. *J. Genet. Genomics* **2013**, *40*, 375−390.
(9) Han, X.; Yang, K.; Gross, R. W. *Mass Spectrom. Rev.* **2012**, *31*, 134−178.
(10) Cajka, T.; Fiehn, O. *TrAC, Trends Anal. Chem.* **2014**, *61*, 192−206.
(11) Lam, S. M.; Tian, H.; Shui, G. *Biochim. Biophys. Acta* **2017**, *1862*, 752−761.
(12) Kliman, M.; May, J. C.; McLean, J. A. *Biochim. Biophys. Acta, Mol. Cell Biol. Lipids* **2011**, *1811*, 935−945.
(13) Paglia, G.; Kliman, M.; Claude, E.; Geromanos, S.; Astarita, G. *Anal. Bioanal. Chem.* **2015**, *407*, 4995−5007.
(14) Paglia, G.; Astarita, G. *Nat. Protoc.* **2017**, *12*, 797−813.
(15) Hankin, J. A.; Barkley, R. M.; Zemski-Berry, K.; Deng, Y.; Murphy, R. C. *Anal. Chem.* **2016**, *88*, 6274−6282.
(16) Berry, K. A.; Barkley, R. M.; Berry, J. J.; Hankin, J. A.; Hoyes, E.; Brown, J. M.; Murphy, R. C. *Anal. Chem.* **2017**, *89*, 916−921.
(17) May, J. C.; McLean, J. A. *Anal. Chem.* **2015**, *87*, 1422−1436.
(18) Zheng, X.; Wojcik, R.; Zhang, X.; Ibrahim, Y. M.; Burnum-Johnson, K. E.; Orton, D. J.; Monroe, M. E.; Moore, R. J.; Smith, R. D.; Baker, E. S. *Annu. Rev. Anal. Chem.* **2017**, *10*, 71−92.
(19) Stephan, S.; Jakob, C.; Hippler, J.; Schmitz, O. J. *Anal. Bioanal. Chem.* **2016**, *408*, 3751−3759.
(20) Stephan, S.; Hippler, J.; Kohler, T.; Deeb, A. A.; Schmidt, T. C.; Schmitz, O. J. *Anal. Bioanal. Chem.* **2016**, *408*, 6545−6555.
(21) Baker, P. R.; Armando, A. M.; Campbell, J. L.; Quehenberger, O.; Dennis, E. A. *J. Lipid Res.* **2014**, *55*, 2432−2442.
(22) Maccarone, A. T.; Duldig, J.; Mitchell, T. W.; Blanksby, S. J.; Duchoslav, E.; Campbell, J. L. *J. Lipid Res.* **2014**, *55*, 1668−1677.
(23) Damen, C. W.; Isaac, G.; Langridge, J.; Hankemeier, T.; Vreeken, R. J. *J. Lipid Res.* **2014**, *55*, 1772−1783.
(24) Groessl, M.; Graf, S.; Knochenmuss, R. *Analyst* **2015**, *140*, 6904−6911.
(25) Kyle, J. E.; Zhang, X.; Weitz, K. K.; Monroe, M. E.; Ibrahim, Y. M.; Moore, R. J.; Cha, J.; Sun, X.; Lovelace, E. S.; Wagoner, J.; Polyak, S. J.; Metz, T. O.; Dey, S. K.; Smith, R. D.; Burnum-Johnson, K. E.; Baker, E. S. *Analyst* **2016**, *141*, 1649−1659.
(26) Wojcik, R.; Webb, I. K.; Deng, L.; Garimella, S. V.; Prost, S. A.; Ibrahim, Y. M.; Baker, E. S.; Smith, R. D. *Int. J. Mol. Sci.* **2017**, *18*, E183.
(27) Paglia, G.; Angel, P.; Williams, J. P.; Richardson, K.; Olivos, H. J.; Thompson, J. W.; Menikarachchi, L.; Lai, S.; Walsh, C.; Moseley, A.; Plumb, R. S.; Grant, D. F.; Palsson, B. O.; Langridge, J.; Geromanos, S.; Astarita, G. *Anal. Chem.* **2015**, *87*, 1137−1144.
(28) Hines, K. M.; Herron, J.; Xu, L. *J. Lipid Res.* **2017**, *58*, 809−819.
(29) May, J. C.; Morris, C. B.; McLean, J. A. *Anal. Chem.* **2017**, *89*, 1032−1044.
(30) Metz, T. O.; Baker, E. S.; Schymanski, E. L.; Renslow, R. S.; Thomas, D. G.; Causon, T. J.; Webb, I. K.; Hann, S.; Smith, R. D.; Teeguarden, J. G. *Bioanalysis* **2017**, *9*, 81−98.
(31) Campuzano, I.; Bush, M. F.; Robinson, C. V.; Beaumont, C.; Richardson, K.; Kim, H.; Kim, H. I. *Anal. Chem.* **2012**, *84*, 1026−1033.
(32) Shvartsburg, A. A.; Siu, K. W.; Clemmer, D. E. *J. Am. Soc. Mass Spectrom.* **2001**, *12*, 885−888.
(33) Zhou, Z.; Shen, X.; Tu, J.; Zhu, Z. J. *Anal. Chem.* **2016**, *88*, 11084−11091.
(34) Zhou, Z.; Xiong, X.; Zhu, Z.-J. *Bioinformatics* **2017**, *33*, 2235−2237.
(35) Ma, X.; Liu, J.; Zhang, Z.; Bo, T.; Bai, Y.; Liu, H. *Rapid Commun. Mass Spectrom.* **2017**, *31*, 33−38.
(36) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; Wiley-VCH: Weinheim, 2009.
(37) Sud, M.; Fahy, E.; Cotter, D.; Brown, A.; Dennis, E. A.; Glass, C. K.; Merrill, A. H., Jr.; Murphy, R. C.; Raetz, C. R.; Russell, D. W.; Subramaniam, S. *Nucleic Acids Res.* **2007**, *35*, D527−532.
(38) Hines, K. M.; May, J. C.; McLean, J. A.; Xu, L. *Anal. Chem.* **2016**, *88*, 7329−7336.
(39) Bush, M. F.; Campuzano, I. D.; Robinson, C. V. *Anal. Chem.* **2012**, *84*, 7124−7130.
(40) Jackson, S. N.; Ugarov, M.; Post, J. D.; Egan, T.; Langlais, D.; Schultz, J. A.; Woods, A. S. *J. Am. Soc. Mass Spectrom.* **2008**, *19*, 1655−1662.
(41) Kim, H. I.; Kim, H.; Pang, E. S.; Ryu, E. K.; Beegle, L. W.; Loo, J. A.; Goddard, W. A.; Kanik, I. *Anal. Chem.* **2009**, *81*, 8289−8297.
(42) May, J. C.; Goodwin, C. R.; Lareau, N. M.; Leaptrot, K. L.; Morris, C. B.; Kurulugama, R. T.; Mordehai, A.; Klein, C.; Barry, W.; Darland, E.; Overney, G.; Imatani, K.; Stafford, G. C.; Fjeldsted, J. C.; McLean, J. A. *Anal. Chem.* **2014**, *86*, 2107−2116.

DOI: 10.1021/acs.analchem.7b02625
Anal. Chem. 2017, 89, 9559−9566