Least Absolute Shrinkage and Selection Operator-based Prediction of Collision Cross Section Values for Ion Mobility Mass Spectrometric Analysis of Lipids

Jian-Ying Wang[a,b,c#], Ying-Hao Yin[a#], Jia-Yi Zheng[a], Li-Fang Liu[a*], Zhong-Ping Yao[b,c*], Gui-Zhong Xin[a*]

[a]State Key Laboratory of Natural Medicines, Department of Chinese Medicines Analysis, School of Traditional Chinese Pharmacy, China Pharmaceutical University, No. 24 Tongjia Lane, Nanjing, China.

[b]State Key Laboratory of Chinese Medicine and Molecular Pharmacology (Incubation) and Shenzhen Key Laboratory of Food Biological Safety Control, Shenzhen Research Institute of Hong Kong Polytechnic University, Shenzhen 518057, China.

[c]State Key Laboratory of Chemical Biology and Drug Discovery, Food Safety and Technology Research Centre and Department of Applied Biology and Chemical Technology, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR, China.

[#]These authors contributed equally to this work.

***Corresponding authors:**

Li-Fang Liu

State Key Laboratory of Natural Medicines, Department of Chinese Medicines Analysis, China Pharmaceutical University, No. 24 Tongjia Lane, Nanjing, China

E-mail address: liulifang69@126.com (L.F. Liu)

Gui-Zhong Xin

Department of Chinese Medicines Analysis, School of Traditional Chinese Pharmacy, China Pharmaceutical University, No. 24 Tongjia Lane, Nanjing, China

Tel: +86 13813884764; email address: xinguizhong1984@126.com.

Zhong-Ping Yao

Department of Applied Biology and Chemical Technology, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

Tel: +852 2358 7332; fax: +852 2358 1559, email address: zhongping.yao@polyu.edu.hk.

Abstract

Collision cross section (CCS) values generated from ion mobility mass spectrometry (IM-MS) have been commonly employed to facilitate lipid identification. However, this is hindered by the limited available lipid standards. Though recently CCS values were predicted by means of computational calculation, the prediction precision was generally not good and the predicted CCS values of lipid isomers were almost identical. To address this challenge, a least absolute shrinkage and selection operator (LASSO)-based prediction method was developed for the prediction of lipids' CCS values in this study. In this method, an array of molecular descriptors were screened and optimized to reflect the subtle difference in structures among different lipid isomers. The use of molecular descriptors together with a wealth of standard CCS values for lipids (365 in total) significantly improved the accuracy and precision of the LASSO model. Its accuracy was externally validated with median relative errors (MRE) of <1.1% using an independent data set. This approach was demonstrated to allow differentiation of *cis/trans* and sn-positional isomers. The results also indicated that the LASSO-based prediction method could practically reduce false-positive identification in IM-MS-based lipidomics.

Keywords: Lipid; Lipidomics; LASSO; Collision cross section; Ion mobility; Mass spectrometry.

1. Introduction

The field of lipidomics aims to detect and characterize lipids present in biological samples. The full characterization of lipids is beneficial for understanding their biological roles such as serving as energy source and signaling molecules for a wide variety of biological functions [1-4]. Currently, mass spectrometry (MS)-based lipidomics approaches are prevalent in providing detailed information on lipidome. However, the full characterization of lipids is still a challenging task due to the vast number of possible isomers that may exist [5, 6]. It is difficult to distinguish so many structural isomers due to the distinct lipid acyl chain positions, double bond locations and specific glycan types, which hinders the delineation and annotation of their biological roles [7-9].

Plenty of literature indicated that the positional distribution of saturated fatty acid on the glycerol backbone of triacylglyceride (TG) molecules, especially in the sn-2 fatty acyl chain, might result in atherosclerosis [10]. In addition, *trans* isomers of fatty acid increase the risk of sudden cardiac death induced by coronary artery disease while *cis* isomers are good for health [11]. Thus, the differentiation of lipids is of great significance. Although many structural isomers of lipids can be distinguished by LC/MS-based approaches using retention time and characteristic fragment ions, potential pitfalls arise when *cis/trans* (Z/E) and sn-2 positional isobaric lipid species yield the same fragment ions [7]. More recently, ion mobility mass spectrometry (IM-MS) has shown to be a powerful tool for lipidomics study [12, 13]. As a gas-phase technique, IM-MS separates ions according to their shapes and sizes [14-16]. A number of collisions occur between ions and inert buffer gas under an electric field, resulting in the differences in drift time, which can be used to calculate the collision cross section (CCS) [17]. The measurement of an ion's CCS value using IM-MS provides specific structural information

of the ion, especially when coupled with a high-resolution MS such as time-of-flight (TOF), Orbitrap [7]. The CCS value together with characteristic fragment ions facilitate the separation of structural isomers, stereoisomers, and tautomers.

However, in the current metabolite libraries, only a small number of standard CCS values are available for lipids due to the limited availability of lipid standards. More recently, theoretical CCS values calculated by computational chemistry tools are being employed in IM-MS-based lipidomics studies [18, 19]. Machine learning is also a splendid method for forecasting CCS values of lipids [20, 21], and has been demonstrated to be very effective in our previous study in combining MALDI-MS analysis with the established spectral database for rapid classification of edible oils [22]. More recently, Zhu *et al*. have utilized a support vector regression (SVR) model to predict the CCS values of lipids [23]. The high prediction accuracy of this method was externally validated using an independent set of lipids, indicating that machine learning models should be of promising perspectives in the prediction of theoretical CCS values of lipids. However, these methods cannot accurately predict the CCS values of isomeric lipids that differ in position or geometry (*cis/trans*) yet. Lack of isomeric lipid standards in the training data set should be the main reason; in addition, insufficient molecular descriptors (MDs) closely related to isomeric lipid should be another reason. Generally, the MDs of a target molecule can be acquired from databases (lipidmaps or HMDB) or be produced with lipids' SMILES formats using softwares (e.g. E-dragon software, R project and ChemAxon).

In this work, using MDs closely related to subtle structural differences of lipid isomers, we described a novel least absolute shrinkage and selection operator (LASSO)-based approach for predicting CCS values of lipids. As a machine learning model, LASSO is generally utilized in

regression analysis which is performed with variable selection and regularization to enhance the prediction accuracy and interpretability of the produced statistical model [24-26]. Firstly, a combination of MDs was screened to precisely describe the subtle structural differences for lipids. Especially, MDs related to isomeric subtle structure were specifically optimized by targeted comparative analysis. Subsequently, the screened combination of MDs together with the published experimentally measured lipids' CCS values (365 in total) were used to build and internally validate the LASSO-based prediction model. Then, our new prediction approach was externally validated using the CCS values produced by other laboratories (CCS values for the 81 lipids), which have a high precision with MRE <1.16%. The prediction accuracy of our LASSO-based prediction model was also validated in our laboratory using 19 lipid standards with MRE <0.51%. The established LASSO-based prediction method was further demonstrated to allow the prediction of CCS values for lipids and the differentiation of lipid isomers, and effectively reduced false positives identification of lipid in the IM-MS-based lipidomics study.

2. Experimental

2.1. Chemicals and materials

Lipid standards (Table S1) used in the external validation were purchased from Avanti polar lipid Co., Ltd (Sweden). LC-MS grade acetonitrile (ACN), isopropyl alcohol (IPA) and methanol (MeOH) were purchased from Thermo Fisher Scientific Ltd. Co (Boston, USA). Ultrapure water was obtained using an 18 Ω m Milli-Q system (Millipore Corporation, Billerica, MA). Poly-DL-alanine was purchased from Sigma-Aldrich (U.K.). Male zebrafish were purchased from Model Animal Research Center of Nanjing University (Nanjing, China). All animal procedures were performed in accordance with the Guidelines for Care and Use of

Laboratory Animals of Shenzhen Research Institute of The Hong Kong Polytechnic University and approved by the Animal Ethics Committee of Shenzhen Research Institute of The Hong Kong Polytechnic University.

2.2. Sample preparation

Standard mixtures. Phospholipids and triacylglycerides were prepared in MeOH and dichloromethane (DCM)/MeOH (50:50, v: v), respectively, before MS analysis.

Biological samples. Lipid extraction protocol was modified from the previously reported protocol [27]. Fresh liver samples were rapidly homogenized for 2 min in 1 mL ice-cold solvents mixture (DCM/MeOH, 2:1, v/v), then centrifuged at 13000 rpm for 10 min after 800 μL pure water was added. The lower organic phase was collected in a new vial and evaporated to dryness under nitrogen. Immediately prior to analysis the lipid extract was diluted with isopropanol/acetonitrile (2:1, 100 μL).

2.3. LC-IM-MS analysis

All experiments were conducted using a Synapt G2 HDMS traveling wave ion mobility mass spectrometer (TWIM-MS) coupled with a UHPLC system (Waters, Manchester, UK). UHPLC separation was performed on a BEH C18 column (particle size:1.7 μm; length: 50 mm; i.d.: 2.1 mm) maintained at 55 °C. The solvent A was H_2O / ACN (6:4, v/v) with 10 mM ammonium formate, and solvent B was IPA / ACN (9:1, v/v). The linear gradient condition was as follows: 0 min, 70% B; 10 min, 100% B; 10-20 min, 100% B; 20.5 min, 70% B; 30 min, 70% B. The total gradient time was 30 minutes for each analysis. The flow rate was 0.2 mL min^{-1} and the injection volume was 2 μL. All CCS values were measured using the single electric field

method with nitrogen as the drift gas. The wave velocity and wave height were 600 m s^{-1} and 40 V, respectively. The source parameters were set as follows: source temperature, 90 °C; desolvation temperature, 400 °C; core gas flow, 20 L h^{-1}; cone voltage, 40 V; capillary voltage, 3 kV and 2.5 kV in positive or negative ion modes, respectively. The TOF mass range was set as *m/z* 50 – 1200 Da. The collision energy was set as 40 V.

2.4. Optimization and development of LASSO-based prediction method

All data processes and calculations were performed in an open-source R programming environment. MDs are a series of numeric values that characterize the structural and physicochemical properties of molecules. To produce MDs that can reflect the difference between lipid isomers, the optimization and selection of MDs were conducted in the following steps. First, we downloaded SMILES formats of lipids from LIPID MAPS and then 256 MDs were obtained with the assistance of those SMILES structures utilizing the R package "rcdk" [23, 28]. Especially, another 15 MDs, produced by MarvinSketch from ChemAxon, were specifically selected by targeted comparative analysis manually for 15 MDs of many lipid isomers downloaded from LIPIDMAPS, these 15 MDs should be closely related to subtle structures of isomeric lipids. Subsequently, 95 out of 271 MDs were indiscriminate among the lipids in the training data set and thereby were removed. The remaining 176 MDs were stepwisely optimized using the training data set. Briefly, 80% of the CCS values of lipids for the modeling were randomly chosen as training data set to build a regression model using the 176 MDs. The MDs with the least contribution to the regression model were subsequently removed. As a result, 52 and 63 molecular descriptors were screened for the prediction of lipids' CCS values in the positive and negative ion modes, respectively (Tables S2, S3). Finally, the training data set and MDs were used to develop the LASSO prediction model, and 20% of the CCS values for the modeling were employed as internal validation dataset, followed by the

external validation including inter-lab validation and intra-lab validation. Same workflow (Figure 1) was employed for the prediction method in different ion modes.

A total of 314 and 132 experimental CCS values of lipid species acquired from other laboratories in positive and negative modes [12, 23, 29-32], respectively, were obtained and separated into two datasets, including dataset A (365 lipids) and dataset B (81 lipids). It should be noted that only those lipids annotated into the level of sn-1_sn-2 position, instead of those annotated as total carbon: total double bond, in the references mentioned above were recruited. 80% of dataset A was used for model building and the remaining 20% was for the internal validation, and dataset B, in which the CCS values were obtained from the laboratories different from that of dataset A, was employed for the external inter-lab validation. In dataset A, CCS values from 255 lipids consisting of 94 sphingolipids ($[M-H_2O+H]^+$/$[M+H]^+$), 98 protonated phospholipids and 62 glycerolipids ($[M+NH_4]^+$) were collected in positive ion mode; and 110 lipid CCS values composing of deprotonated fatty acids, phospholipids and sphingolipids were collected in negative ion mode. Firstly, using the training data set and the screened MDs, the LASSO-based prediction model was optimized and trained. Least-Angle Regression (LARS) was utilized to calculate the regularization parameter (Lambda). An appropriate Lambda was crucial in building a convincing model. Here, cross-validation of the lattice point of Lambda was proceeded to choose a Lambda with the smallest root mean square error (RMSE) (Figure S1). To validate the performance of the model, around 20% of lipids in the data set were then used for internal validation. Finally, a total of 100 lipids were utilized for external validation, including inter-lab validation (81 lipids) and intra-lab validation (19 lipids, calculated in our lab) to observe the accuracy and differentiation ability of this method.

2.5. Calculation of experimental CCS values of lipids

In the intra-lab validation, CCS values of lipid standards were measured. CCS values obtained in nitrogen were experimentally determined with poly-DL-alanine at a concentration of 10 mg L^{-1} in H_2O/ACN (50:50, v:v) as the calibrant species in both positive and negative ion modes. CCS values were derived using a procedure previously reported [33].

3. Results and discussion

3.1. Development of the LASSO-based method for prediction of CCS values with high accuracy

LASSO algorithm was used to develop the CCS prediction method due to its powerful capability to perform both variable selection and regularization, which could improve the prediction accuracy. Ultimately, 52 and 63 MDs that were concerned with subtle structural variation of lipid species and isomers were particularly optimized for prediction in positive and negative ion modes, respectively. It should be noted that MDs calculated from MarvinSketch were of great variation for lipid isomers (Table S3). The experimental CCS values of 255 and 110 lipids in positive and negative ion modes, respectively, were then collected as dataset A for predictive method development and internal validation. Besides, as shown in Figure S1, Lambda of LASSO was first optimized to achieve the best prediction performance. Then, the development of the LASSO-based prediction model was carried out using 80% lipids in the dataset A, the remaining 20% were randomly selected for internal validation. For both positive and negative ion modes, excellent fits with R^2 values more than 0.99 were obtained (Figure 2a, b). MREs of the comparison were 0.928% and 0.867%, for data sets of positive and negative ion modes, respectively. The result proved that the optimized Lambda value for the LASSO-based model performed well for CCS value prediction. Finally, using the optimized Lambda,

the LASSO prediction method was built based on the training data set. For all the lipids in positive ion mode, median relative error was 0.691%, and the R^2 value was 0.9914 (Figure 2c). In addition, 100-fold cross-validation was performed during method development to avoid overfitting. Similar parameter optimization and method development were performed in negative ion mode using MDs and CCS values from lipids, with a MRE of 0.796% and an R^2 value of 0.9970 (Figure 2d). These results supported the high prediction accuracy of the LASSO prediction method.

3.2. External validation

The wave velocity and wave height were optimized at the initial stages, and when the wave velocity was adjusted between 450 m s^{-1} and 600 m s^{-1} with wave height changed between 30 V and 40 V, the CCS error was around 2%, which was acceptable. Finally, when the wave velocity and the wave height were set at 600 m s^{-1} and 40 V, respectively, a best separation of isomer was obtained. We further evaluated the performance of the LASSO prediction method using an independent set of lipids (59 and 22 lipids in positive and negative ion modes, respectively) to serve as an external inter-lab validation data set. The CCS values of these lipids were measured in the labs different from that of training and internal validation dataset. As shown in Figure 3a and Figure 3b, the predicted CCS values from the LASSO method matched very well with the experimentally measured CCS values for these tested lipids. The R^2 values of the regression curves were 0.9925 and 0.9983 in positive and negative ion modes, respectively. MREs were 1.159% and 0.532%, respectively. This performance was similar to our intra-lab validation conducted using the same instrument (Figure 3c, d). The MREs of our intra-lab validation were 0.569% and 0.529% in positive and negative ion mode, respectively. To summarize the performance of the LASSO prediction method, in terms of external validation results, we discovered that about 96% of lipids had predicted CCS values within 1.8%

of relative errors, while more than 70% of lipids had predicted CCS values within 1% of relative errors (Figure 3e). Meanwhile, with the use of the present LASSO-based prediction method, we also compared the predicted CCS values of glycerolipids, phospholipids and sphingolipids, and we found that phospholipids generated the smallest predicted CCS values (205-338 $Å^2$), followed by sphingolipids (286-319 $Å^2$) and glycerolipids (287-332 $Å^2$) (Figure 3f). This trend was consistent with that of some experimental CCS values reported previously [33]. Besides, our result demonstrated that the prediction accuracy of the predicted CCS values for different lipid classes were distinct. Phosphate acid (PA), phosphoserine (PS) and phosphoinositol (PI) had the highest prediction accuracy with MREs of smaller than 0.5%, followed by sphingomyelin (SM), GL, phosphocholine (PC), phosphoglycerol (PG), phosphoethanolamine (PE) and ceramide with MREs of ~0.7%.

3.3. Good differentiation ability of lipid isomers

It is difficult to differentiate lipid isomers using the previously reported prediction methods or theoretical calculation methods [34, 35]. In order to test the separating capacity of the LASSO-based prediction method in differentiating lipid isomers, we further compared the LASSO predicted CCS values of 3 pairs of lipid isomers (2 pairs of TG isomer in positive and of 1 pair of fatty acid isomers in negative ion modes, respectively) with their experimental CCS values measured in our laboratory. As shown in Figure 4a, a marked change in ion mobility behavior was observed between a pair of *cis* and *trans* fatty acids, including 9Z-octadecenoic acid and 9E-octadecenoic acid with predicted CCS values of 183.3 $Å^2$ and 186.2 $Å^2$. Besides, a similar difference of predicted CCS values was found between a pair of sn-2 positional isomers, *i.e.* 1,2-linolein-3-olein and 1,3-linolein-2-olein, with predicted CCS values of 310.3 $Å^2$ and 312.8 $Å^2$ (Figure 4b). To test whether the LASSO prediction method can be used to distinguish ω-3/ ω-6 lipid isomers, we measured and predicted the CCS values of ω-3 and ω-6

docosapentaenoic acid (DPA). The result showed that not only their retention time, but also their experimental CCS values were identical, which means that they cannot be separated by means of LC and IMMS separation (Figure 4c and Figure S2). The prediction method also exhibited only 0.6 $Å^2$ differences between two isomers, enabling them to be unclearly distinguished (Figure 4c).

Compared with the previous prediction methods, a possible explanation for the improvement of our method in the differentiation of lipid isomers is that several isomeric lipid standards were imported into the training data set to reflect the subtle structures of isomers during the model establishment in the present study. Another novel aspect of our LASSO prediction approach is the selection of important MDs to effectively distinguish the structures of lipid isomers, especially for larger molecules. Taking dipole moment as an example, since the two substituents of the *trans* isomer are opposite in orientation and thus can be completely or partially offset, while the *cis* isomers generate dipole moment in the direction of the vector resultant force due to two symmetrically arranged groups. In general, the *cis* isomer has a larger dipole moment than the *trans* isomer. Therefore, the dipole moment should be one of the most dominant factors to discriminate *cis/trans* isomers [36, 37]. It was reported that differential mobility spectrometer system (DMS) (SelexIon™, AB SCIEX, Concord, Ontario, Canada) with higher resolving power can be used to separate lipid isomers based on dipole moment [40, 41], so we utilized DMS to analyze a pair of sn-positional isomers and *cis/trans* isomers separately. Ultimately, the results demonstrated that the dipole moments of PC(14:0/16:0) and PC(16:0/14:0) were distinct and different dipole moments were also obtained from PC (18:1△6-*cis*) and PC (18:1△9-*cis*). (Figure S3). Therefore, the results suggested that dipole moment that used in our LASSO method should be one of important MDs to reflect the structural difference of isomer and then help to improve the separation of isomers in our method.

The energy of lipids varies with the alteration of conformation, therefore, some important MDs, such as Dreiding and MMFF94 [42], yielded by calculation of energy of molecule were chosen to allow differentiation of both *cis/trans* and sn-2 positional lipid isomers. Although this method cannot be utilized to differentiate all types of lipid isomers, our investigation indicated that if more topological MDs, *e.g.* molecular connectivity indices which are related to the three-dimensional structures of isomers, are employed in the development of LASSO model, the structural difference will be amplified in CCS values [43]. However, this method can only be applied in the prediction of CCS values produced from the TWIM instrument since the training data set was mainly based on the TWIM-derived CCS values. Furthermore, although lipid calibrants could yield CCS values with higher accuracy, currently a variety of CCS values calibrated by lipid calibrant were mainly generated for lipid subclasses (e.g., PC(36:2)) instead of lipid species (e.g., PC(18:1/18:1) or PC(18:0/18:2)). Therefore, the number of CCS values for lipid species produced based on lipid calibrant was quite limited in our prediction model. This method is thus applicable in TW instrument when polyalanine was applied as calibrant.

3.4. Application for identification of biomarkers in lipidomics

The LASSO prediction method was employed to improve the confidence of lipid identification in our study for the profiling of lipid alterations occurred after obese zebrafish development. Raw UPLC-ESI-MS data was directly imported to the Progenesis QI software (Waters-Nonlinear) for data processing, which includes peak picking, alignment and data normalization. The processed data matrices were imported to the IBM SPSS Statistics software (Version 11.0, SPSS Inc., Chicago, IL, USA) for t-test. The *m/z* values of ions with p-value < 0.05 were further exported to the SIMCA 14.1 software for Orthogonal Partial Least Squares Discriminant Analysis (OPLS-DA). Ions with fold change $\geqslant 1.5$ and Variable Importance for Projection

(VIP) ⩾1 were considered as potential biomarkers and were subjected to identification. After the data analysis, searching with the accurate masses of lipids and mass error < 5 ppm against HMDB and LIPID MAPS resulted in 39 potential lipid ions. These 39 lipid ions have many candidates with the combinations of various isomers. Although MS/MS spectra are further used to confirm and rule out the searching results, the majority of these lipid candidates possessed several isomers that have identical *m/z* values and similar fragmentation patterns. Herein, a cutoff of 5 ppm and experimental CCS ± 2% search criteria was used to remove false positive identifications of the potential biomarkers [12]. By this way, we found 17 mismatched lipid identifications (false positives). The remaining 22 ions and their candidates were listed in Table S4. The use of our predicted CCS values could effectively remove the lipid candidates with lower confidences. For example, ion at *m/z* 806.7206 had three potential lipid candidates matched with parent ion and fragment ions, but one mismatch was found by applying CCS filter and it was ruled out, thus resulting in two candidates. (Table S4). Finally, a total of 22 lipid ions including 78 lipid candidates were annotated in this study. These results indicated that the use of the predicted CCS values could effectively reduce false positive identifications and improve the identification confidence for untargeted lipidomics.

4. Conclusions

Lipids play crucial roles in biological processes such as energy storage and signaling. IM-MS can facilitate the identification of lipids, but this is restricted by the unavailability of lipid standards. In our study, a LASSO-based prediction method with good accuracy was developed, which could allow the prediction of CCS value for lipids, enable the differentiation of lipid isomers, including *cis/trans* and sn-2 lipid isomers. An application of this prediction method in untargeted lipidomics demonstrated that the method could be used to effectively reduce false

positive identifications of lipid biomarkers. Due to the important roles of lipids in health and diseases, this LASSO-based prediction method is expected to be used in a wide range of studies, including differentiation of important lipid isomers, exploration of the mechanism of diseases and drug development. But there were some limitations in this work, for example, some fatty acid species in the training dataset were collected from DTIMS instead of TWIMS, that might be responsible for the largest MRE (~2%) for fatty acid. In addition, CCS values calculated by different calibrants also contributed to the error of this prediction method. Although these errors were not larger than that of our method, this model should be improved in the future when more TWIMS-based CCS values calculated by lipid calibrant are published.

Author contributions

Jianying Wang designed the experiment and conducted the lipid extraction, LC-IMMS analysis, data processing and analysis. YingHao Yin collected the CCS values, build the LASSO model and validate the model. Jiayi Zheng and Lifang Liu helped to check the results in this manuscript. Guizhong Xin and Zhongping Yao provided the guidance about the LC-MS analysis and helped to proofread throughout the manuscript.

Conflict of interest

There is no conflict of interest to declare.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Nos. 82074089, 81874306 and 81973577), Hong Kong Research Grants Council (Grant Nos. 15304020, R5013-19, R4005-18 and C4002-20WF), Postgraduate Research & Practice Innovation

Program of Jiangsu Province (Nos. KYCX19_0654) and priority academic program development of Jiangsu higher education institutions (PAPD). The supports from the University Research Facility in Chemical and Environmental Analysis (UCEA) and the University Research Facility in Life Sciences (ULS) of Hong Kong Polytechnic University are also acknowledged.

References

[1] F. Spener, M. Lagarde, A. Géloën and M. Record, *Euro. Lipid. Sci. Technol.*, 2003, **105**, 481-482.

[2] C. R. Santos and A. Schulze, *FEBS J.,* 2012, **279**, 2610-2623.

[3] M. P. Wymann and R. Schneiter, *Nat. Rev. Mol. Cell Biol.,* 2008, **9,** 162-176.

[4] X. Han, *Nat. Rev. Endocrinol.,* 2016, **12,** 668-679.

[5] R. Wojcik, I. K. Webb, L. Deng, S. V. Garimella, S. A. Prost, Y. M. Ibrahim, E. S. Baker and R. D. Smith, *Int. J. Mol. Sci.,* 2017, **18,** 183.

[6] A. P. Bowman, R. R. Abzalimov and A. A. Shvartsburg, *J. Am. Soc. Mass. Spectrom.,* 2017, **28,** 1552-1561.

[7] J. E. Kyle, X. Zhang, K. K. Weitz, M. E. Monroe, Y. M. Ibrahim, R. J. Moore, J. Cha, X. Sun, E. S. Lovelace, J. Wagoner, S. J. Polyak, T. O. Metz, S. K. Dey, R. D. Smith, K. E. Burnum-Johnson and E. S. Baker, *Analyst,* 2016, **141,** 1649-1659.

[8] B. L. J. Poad, A. T. Maccarone, H. Yu, T. W. Mitchell, E. M. Saied, C. Arenz, T. Hornemann, J. N. Bull, E. J. Bieske and S. J. Blanksby, *Anal. Chem.,* 2018, **90,** 5343-5351.

[9] X. Ma, L. Chong, R. Tian, R. Shi, T. Y. Hu, Z. Ouyang and Y. Xia, *Proc. Natl. Acad. Sci. U S A,* 2016, **113,** 2573-2578.

[10] S. Kubow, *J. Nutr. Biochem.,* 1996, **7,** 530-541.

[11] R. J. de Souza, A. Mente, A. Maroleanu, A. I. Cozma, V. Ha, T. Kishibe, E. Uleryk, P. Budylowski, H. Schünemann, J. Beyene and S. S. Anand, *BMJ,* 2015, **351,** h3978.

[12] G. Paglia, P. Angel, J. P. Williams, K. Richardson, H. J. Olivos, J. W. Thompson, L. Menikarachchi, S. Lai, C. Walsh, A. Moseley, R. S. Plumb, D. F. Grant, B. O. Palsson, J. Langridge, S. Geromanos and G. Astarita, *Anal. Chem.,* 2015, **87,** 1137-1144.

[13] G. Paglia, M. Kliman, E. Claude, S. Geromanos and G. Astarita, *Anal. Bioanal. Chem.,* 2015, **407,** 4995-5007.

[14] C. B. Lietz, Q. Yu and L. Li, J. *Am. Soc. Mass Spectrom.,* 2014, **25,** 2009-2019.

[15] M. F. Bush, Z. Hall, K. Giles, J. Hoyes, C. V. Robinson and B. T. Ruotolo, *Anal. Chem.,* 2010, **82,** 9557-9565.

[16] C. Lapthorn, F. Pullen and B. Z. Chowdhry, *Mass Spectrom. Rev.,* 2013, **32,** 43-71.

[17] J. C. May and J. A. McLean, *Anal. Chem.,* 2015, **87,** 1422-1436.

[18] H. I. Kim, H. Kim, E. S. Pang, E. K. Ryu, L. W. Beegle, J. A. Loo, W. A. Goddard and I. Kanik, *Anal. Chem.,* 2009, **81,** 8289-8297.

[19] C. Lapthorn, F. S. Pullen, B. Z. Chowdhry, P. Wright, G. L. Perkins, Y. Heredia, *Analyst,* 2015, **140,** 6814-6823.

[20] M. T. Soper-Hopper, A. S. Petrov, J. N. Howard, S. S. Yu, J. G. Forsythe, M. A. Grover and F. M. Fernandez, *Chem. Commun. (Camb.),* 2017, **53,** 7624-7627.

[21] Z. Zhou, J. Tu and Z. Zhu, *Curr. Opin. Chem. Biol.,* 2018, **42,** 34-41.

[22] T. Ng, S. Li, C. Ng, P. So, T. Wong, Z. Li, S. Chan and Z. Yao, *Food Chem.,* 2018, **252,** 335-342.

[23] Z. Zhou, J. Tu, X. Xiong, X. Shen and Z. Zhu, *Anal. Chem.,* 2017, **89,** 9559-9566.

[24] R. Tibshirani, *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 2011, **73,** 267-288.

[25] S. Datta, J. Le-Rademacher and S. Datta, *Biometrics,* 2007, **63,** 259-271.

[26] S. Kopprasch, S. Dheban, K. Schuhmann, A. Xu, K. M. Schulte, C. J. Simeonovic, P. E. H. Schwarz, S. R. Bornstein, A. Shevchenko and J. Graessler, *PLOS ONE,* 2016, **11,** e0164173.

[27] W. Chang, J. Zhou, Y. Li, Z. Shi, L. Wang, J. Yang, P. Li, L. Liu and G. Xin, *Anal. Chim. Acta,* 2016, **950,** 138-146.

[28] T. Serino, Y. Takigawa, S. Nakamura, S. Nakamura, M. Huang, N. Ono, A. U. Amin and S. Kanay,. *Journal of Computer Aided Chemistry,* 2019, **20,** 92-103.

[29] F. Zhang, S. Guo, M. Zhang, Z. Zhang and Y. Guo, *J. Mass Spectrom.,* 2015, **50,** 906-913.

[30] K. M. Hines, J. C. May, J. A. Mclean and L. Xu, *Anal. Chem.,* 2016, **88,** 7329-7336.

[31] L. Zhang and A. Vertes, *Anal. Chem.,* 2015, **87,** 10397-10405.

[32] P. Deng, D. Zhong, X. Wang, Y. Dai, L. Zhou, Y. Leng and X. Chen, *Anal. Chim. Acta,* 2016, **925,** 23-33.

[33] D. P. Smith, T. W. Knapman, I. Campuzano, R. W. Malham, J. T. Berryman, S. E. Radford and A. E. Ashcroft, *Eur. J. Mass Spectrom. (Chichester),* 2009, **15,** 113-130.

[34] I. Blaženović, T. Shen, S. S. Mehta, T. Kind, J. Ji, M. Piparo, F. Caccioa, L. Mondello, and O. Fiehn, *Anal. Chem.,* 2018, **90,** 10758-10764.

[35] D. H. Ross, J. H. Cho, R. Zhang, K. M. Hines, and L. Xu, *Anal. Chem*., 2020, **92,** 14967-14975.

[36] C. Brand, W. L. Meerts and M. Schmitt, *J. Phys. Chem.* A, 2011, **115,** 9612-9619.

[37] A. Muždalo, P. Saalfrank, J. Vreede and M. Santer, *J. Chem. Theory Comput.,* 2018, **14,** 2042-2051.

[38] J. Boschmans, S. Jacobs, J. P. Williams, M. Palmer, K. Richardson, K. Giles, C. Lapthorn, W. A. Herrebout, F. Lemiere and F. Sobott, *Analyst,* 2016, **141,** 4044-4054.

[39] P. M. Lalli, Y. E. Corilo, M. Fasciotti, M. F. Riccio, G. F. de Sa, R. J. Daroda, G. H. M. F. Souza, M. McCullagh, M. D. Bartberger, M. N. Eberlin and I. D. G. Campuzano, *J. Mass Spectrom.,* 2013, **48,** 989-997.

[40] M. Šala, M. Lísa, J. L. Campbell and M. Holčapek, *Rapid Communications in Mass Spectrometry,* 2016, **30,** 256-264.

[41] A.T. Maccarone, J. Duldig, T. W. Mitchell, S. J. Blanksby, E. Duchoslav and J. L. Campbell, *J. Lipid Res.,* 2014, 55, 1668-1677.

[42] J. W. Lee, H. H. L. Lee, K. L. Davidson, M. F. Bush and H. I. Kim, *Analyst,* 2018, **143,** 1786-1796.

[43] A. Golbraikh and A. Tropsha, *J. Chem. Inf. Comput. Sci.,* 2003, **43,** 144-154.

Figure captions

Figure 1. Schematic diagram of CCS value prediction using LASSO.

Figure 2. (a, b) Regression curves between LASSO-based predicted and measured CCS values in internal validation for positive (a) and negative (b) ion mode; (c, d) validation of the prediction performance for positive (c) and negative (d).

Figure 3. External validations of the CCS prediction: (a, b) Regression curves between the predicted and measured CCS values in inter-lab validations in positive (a) and negative (b) ion modes; (c, d) regression curves between the predicted and measured CCS values in intra-lab validations in positive (c) and negative (d) ion modes; (e, f) the percentages of CCS values within certain relative errors in positive (e); and (f) ordering of the predicted CCS values among different lipid categories.

Figure 4. Comparisons of the experimental and predicted CCS value of *cis/trans* isomers 9Z-octadecenoic acid/9E-octadecenoic acid in negative ion mode (a), sn-positional isomers 1,2-linolein-3-olein /1,3-linolein-2-olein (b) and ω-3/ω-6 DPA (c) in positive ion mode.

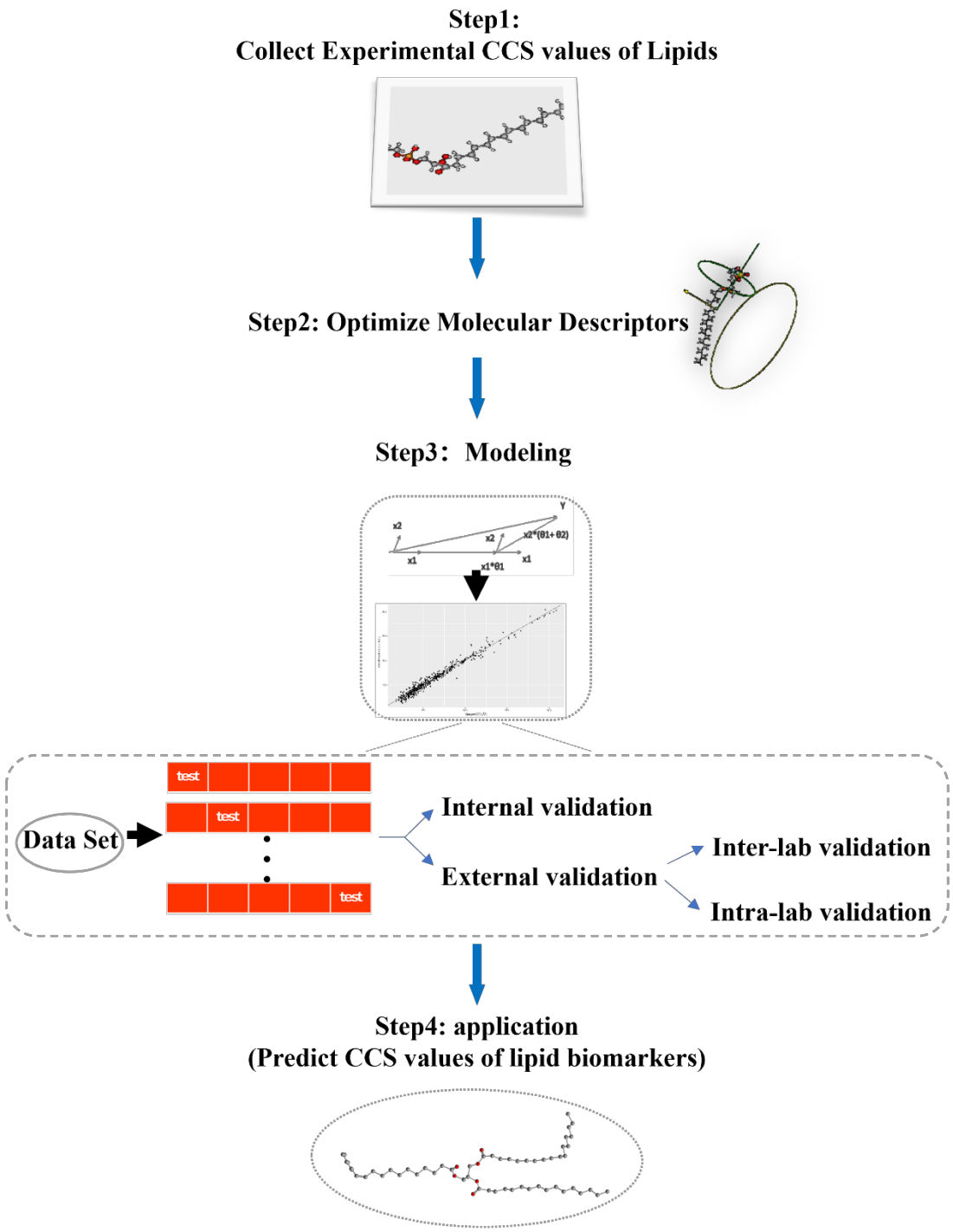
Step1:
Collect Experimental CCS values of Lipids
Step2: Optimize Molecular Descriptors
Step3: Modeling
x2
x1
x1*θ1
x2*(θ1+θ2)
Y
test
Data Set
Internal validation
External validation
Inter-lab validation
Intra-lab validation
Step4: application
(Predict CCS values of lipid biomarkers)

Figure 1

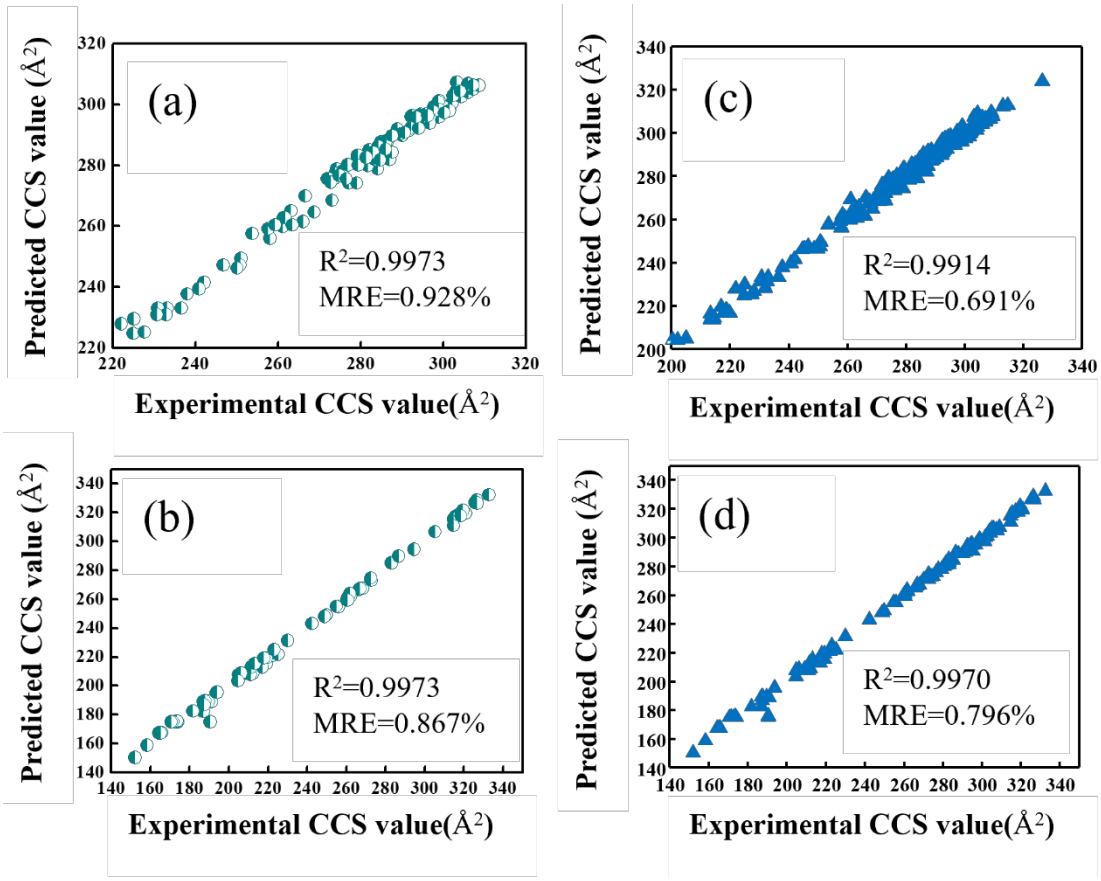
(a)
R²=0.9973
MRE=0.928%
Predicted CCS value (Å²)
Experimental CCS value(Å²)
(b)
R²=0.9973
MRE=0.867%
Predicted CCS value (Å²)
Experimental CCS value(Å²)
(c)
R²=0.9914
MRE=0.691%
Predicted CCS value (Å²)
Experimental CCS value(Å²)
(d)
R²=0.9970
MRE=0.796%
Predicted CCS value (Å²)
Experimental CCS value(Å²)

Figure 2

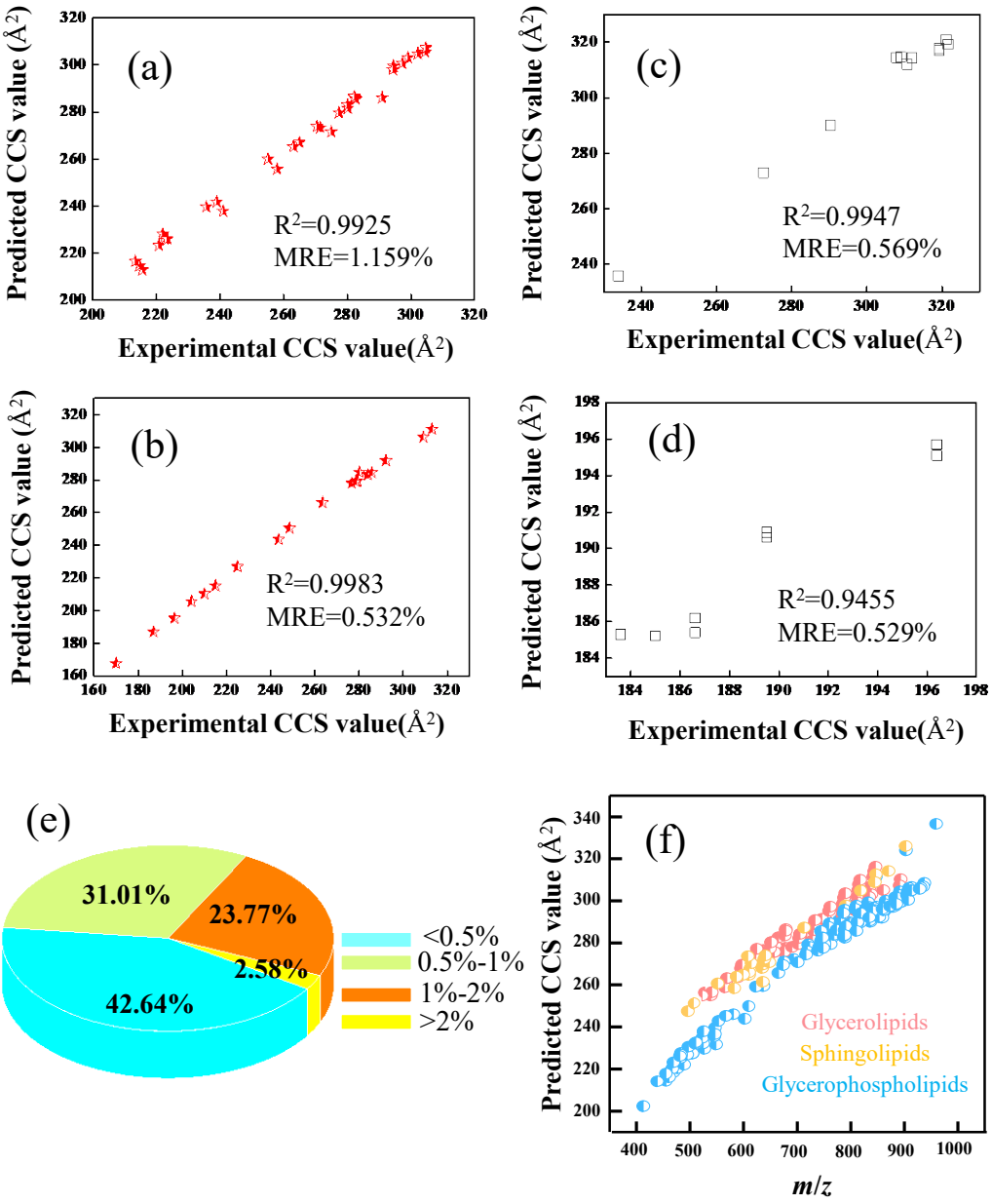
(a)
Predicted CCS value (Å²)
Experimental CCS value(Å²)
R²=0.9925
MRE=1.159%
(c)
R²=0.9947
MRE=0.569%
(b)
R²=0.9983
MRE=0.532%
(d)
R²=0.9455
MRE=0.529%
(e)
31.01%
23.77%
2.58%
42.64%
<0.5%
0.5%-1%
1%-2%
>2%
(f)
Glycerolipids
Sphingolipids
Glycerophospholipids
m/z

Figure 3

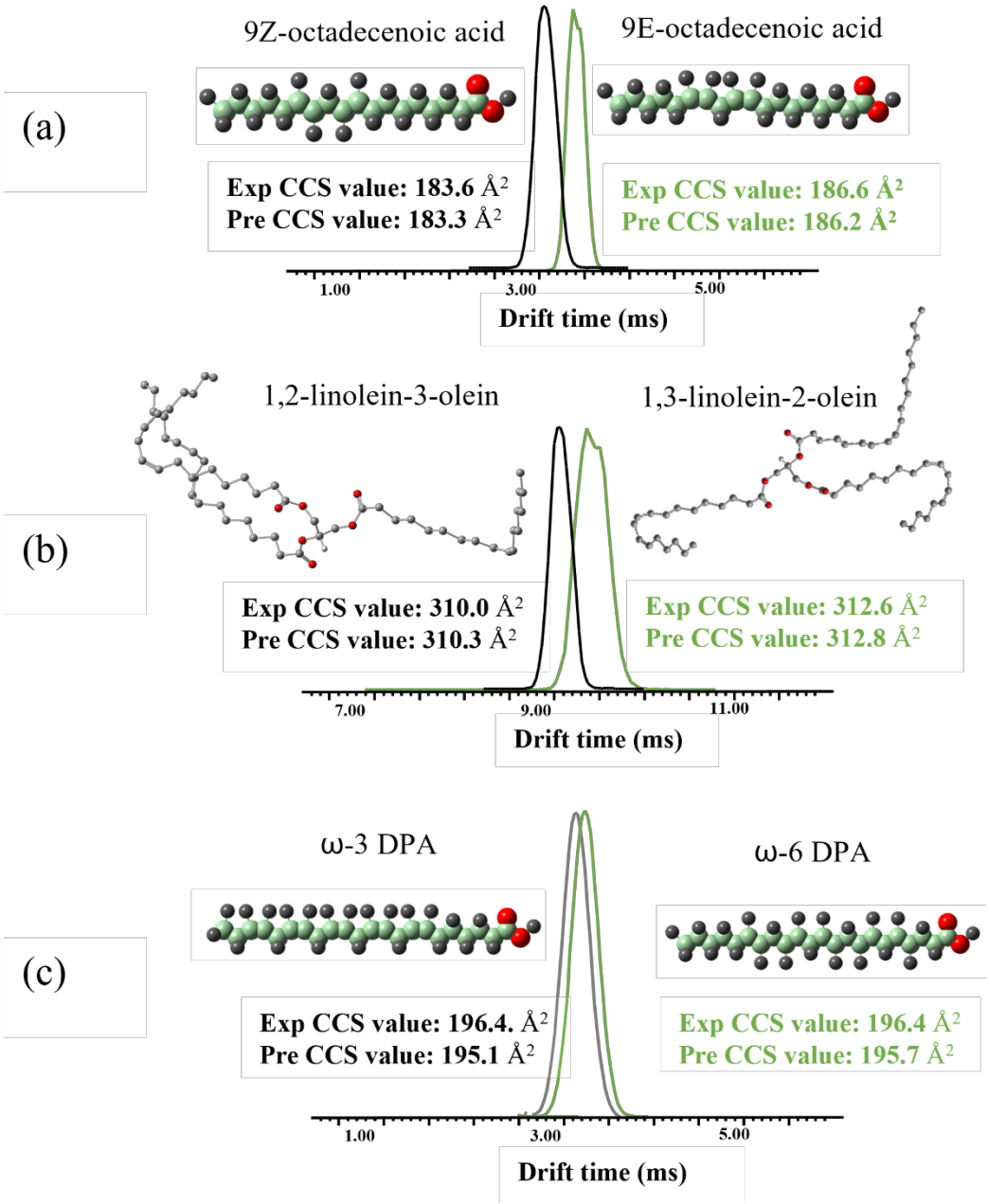
(a)
9Z-octadecenoic acid
9E-octadecenoic acid
Exp CCS value: 183.6 Å²
Pre CCS value: 183.3 Å²
Exp CCS value: 186.6 Å²
Pre CCS value: 186.2 Å²
1.00
3.00
5.00
Drift time (ms)
(b)
1,2-linolein-3-olein
1,3-linolein-2-olein
Exp CCS value: 310.0 Å²
Pre CCS value: 310.3 Å²
Exp CCS value: 312.6 Å²
Pre CCS value: 312.8 Å²
7.00
9.00
11.00
Drift time (ms)
(c)
ω-3 DPA
ω-6 DPA
Exp CCS value: 196.4. Å²
Pre CCS value: 195.1 Å²
Exp CCS value: 196.4 Å²
Pre CCS value: 195.7 Å²
1.00
3.00
5.00
Drift time (ms)

Figure 4