analytical chemistry

pubs.acs.org/ac

Article

LiPydomics: A Python Package for Comprehensive Prediction of Lipid Collision Cross Sections and Retention Times and Analysis of Ion Mobility-Mass Spectrometry-Based Lipidomics Data

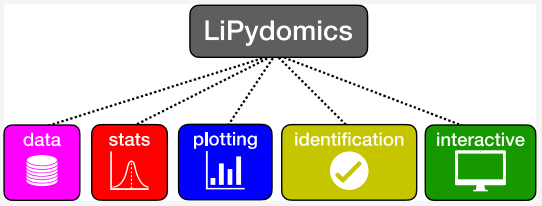
Dylan H. Ross, Jang Ho Cho, Rutan Zhang, Kelly M. Hines, and Libin Xu*

Cite This: *Anal. Chem.* 2020, 92, 14967−14975

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Comprehensive profiling of lipid species in a biological sample, or lipidomics, is a valuable approach to elucidating disease pathogenesis and identifying biomarkers. Currently, a typical lipidomics experiment may track hundreds to thousands of individual lipid species. However, drawing biological conclusions requires multiple steps of data processing to enrich significantly altered features and confident identification of these features. Existing solutions for these data analysis challenges (i.e., multivariate statistics and lipid identification) involve performing various steps using different software applications, which imposes a practical limitation and potentially a negative impact on reproducibility. Hydrophilic interaction liquid chromatography-ion mobility-mass spectrometry (HILIC-IM-MS) has shown advantages in separating lipids through orthogonal dimensions. However, there are still gaps in the coverage of lipid classes in the literature. To enable reproducible and efficient analysis of HILIC-IM-MS lipidomics data, we developed an open-source Python package, LiPydomics, which enables performing statistical and multivariate analyses ("stats" module), generating informative plots ("plotting" module), identifying lipid species at different confidence levels ("identification" module), and carrying out all functions using a user-friendly text-based interface ("interactive" module). To support lipid identification, we assembled a comprehensive experimental database of *m/z* and CCS of 45 lipid classes with 23 classes containing HILIC retention times. Prediction models for CCS and HILIC retention time for 22 and 23 lipid classes, respectively, were trained using the large experimental data set, which enabled the generation of a large predicted lipid database with 145,388 entries. Finally, we demonstrated the utility of the Python package using *Staphylococcus aureus* strains that are resistant to various antimicrobials.



INTRODUCTION

Lipids are a class of biomolecules with broad biological importance, from being structural components of the cell membrane and microdomains to serving as signaling molecules, and dysregulation of lipid metabolism is a common feature of many disease states.[1,2] Lipidomics, the comprehensive analysis of lipids within a biological system, continues to gain popularity as it offers insight into metabolic phenotype and underlying mechanisms of these disease states.[3−5]

Lipid species can be broken into classes and subclasses on the basis of their headgroup chemistry, in addition to the composition of their fatty acyl tails (chain length, number, arrangement, and stereochemistry of double bonds).[6−8] Identification of lipid species may be performed at a variety of levels of structural detail, ranging from basic lipid class (Level 1) to complete molecular species (lipid class, subclass, and fatty acid isomeric composition, Level 5),[8,9] according to the Lipidomics Standard Initiative (LSI). In lipidomics experiments, it is desirable to identify lipid species at the highest level possible in order to gain the most complete understanding of the biological processes being studied. The use of liquid chromatography coupled to ion mobility-mass spectrometry (LC-IM-MS) for lipidomics experiments has been demonstrated to provide a good balance between analytical throughput, resolution, and confidence in lipid identifications.[5,10−12] Hydrophilic interaction liquid chromatography (HILIC) is particularly advantageous as it provides resolution on the basis of lipid headgroups in the retention time dimension, while the orthogonal IM and MS separations allow for further delineation of overlapping subclass and fatty acid sum composition.[11−13] Therefore, this method generally allows Level 3 lipid identifications (lipid class/subclass and fatty acid sum composition).[8,9]

Lipid identifications by IM-MS rely on reference CCS values to compare against, and although there are several large collections of experimental lipid CCS values in the literature,[11−19] these collections do not yet comprehensively cover the vast lipid chemical space (both in terms of class and composition). CCS prediction using machine learning (ML) is

Received: June 16, 2020
Accepted: October 15, 2020
Published: October 29, 2020

https://dx.doi.org/10.1021/acs.analchem.0c02560
Anal. Chem. 2020, 92, 14967−14975

one solution that has gained traction in recent years,[14,19−24] and variants of this general technique have been used by multiple groups to generate predicted CCS databases for lipids.[14−16] Zhou et al. were the first to construct regression models for predicting lipid CCS from a large set of molecular descriptors (45 and 66 for positive and negative modes, respectively) using support vector regression.[14,25] Blaženović et al. trained several classification models (primarily K-nearest neighbor algorithm) using combinations of *m/z*, retention time, and CCS for the prediction of lipid class and carbon number,[15] but their approach did not result in a predicted database covering theoretical lipids. We recently reported a clustering-to-prediction approach for a comprehensive prediction of CCS of diverse chemical structures, including lipids and other types of molecules, but a comprehensive predicted lipid CCS database is still needed.[19] More recently, a large predicted CCS database was constructed using a regression model (XGBoost algorithm) that predicts lipid CCS from 328 molecular descriptors.[16] However, although the previous approaches perform well in lipid identification or classification,[14−16,25] previous databases mostly cover mammalian lipid species, have limited coverage of bacterial lipids, and have no built-in statistical functions, which are needed for a complete lipidomics workflow.

A typical lipidomics experiment may track hundreds to thousands of individual lipid species (features) across a large number of biological samples. The dimensionality of these data sets (many features, fewer samples) can make the interpretation of results difficult since macroscopic differences between samples often correspond to nuanced patterns of change across many features. To address this challenge, multivariate statistical analyses are often applied to lipidomics data in order to draw out the features that are most important or explanatory with regard to the specific biological question being probed. Commonly employed analyses range from simple statistical tests like per-feature ANOVA or Pearson correlation analysis to multivariate dimensionality reduction analyses like principal components analysis (PCA) and partial least-squares discriminant analysis (PLS-DA). At a high level, the use of such analyses allows large lipidomic data sets to be pared down to the set of lipid features that are altered by the specific biological conditions. Owing to the complexity of the entire process and the fact that they are often implemented in different pieces of software, thus requiring moving the data between different programs and converting them between different formats, these analyses can be laborious to perform and difficult to apply consistently across multiple data sets.

To address the primary challenges faced in the analysis of lipidomics data (lipid identification and data complexity), we have prepared a Python package, LiPydomics, which contains a suite of tools for performing data analysis and lipid identification on HILIC-IM-MS lipidomics data in an efficient and reproducible fashion. To support lipid identification, we assembled a comprehensive experimental CCS database from the literature, trained ML models for the prediction of CCS and HILIC retention times using simple but specialized feature sets, and built a predicted lipid database with a broad coverage of lipid classes.

EXPERIMENTAL SECTION

Reference Lipids Database Assembly. A comprehensive collection of lipid CCS values was assembled from individual CCS collections available in the literature[11−18] into a single database of reference lipids for use in lipid identification. Briefly, the source data sets were manually examined for errors and the relevant data (i.e., lipid name, MS adduct, *m/z*, and CCS) from each was converted into the JSON format, yielding clean and consistently formatted data with separate files for each data set. The SQLite3 relational database was initialized with a table to hold the reference CCS values. A series of build scripts developed in-house was used to assemble the combined database from individual cleaned data files in a reproducible fashion. During database assembly, the lipid names were parsed for relevant information (i.e., lipid class, sum composition of fatty acids [number of carbons and unsaturation degrees], and presence of ether lipids), and this information along with metadata reflecting measurement conditions was associated with each entry. CCS values measured on drift tube (DT), traveling wave (TW), and trapped ion mobility spectrometry (TIMS) instruments were included, and those measured on TW were calibrated using lipid standards. For the individual data sets that were measured using the same HILIC-IM-MS protocol as reported previously (referred to hereafter as the established HILIC method),[11−13] the retention time was also stored with each lipid measurement. Additional tables containing predicted *m/z*, CCS, and retention times were also added to the database and populated as described below.

Generation of Exact Lipid *m/z* Values. Theoretical *m/z* values were systematically produced for lipid classes using a subpackage within LiPydomics (*lipydomics/identification/LipidMass*). Monoisotopic masses were computed from the lipid classes and subclasses, fatty acid compositions (ranging from 10 to 30 carbons, including both even and odd numbers, and 0−6 unsaturations per fatty acid), and MS adducts using a method similar to that used in LipidPioneer.[26] Separate functions were used for each lipid class, and lipid classes are further grouped into sphingolipids (Cer, GlcCer, SM), glycerolipids (DG, TG), glycolipids (MGDG, DGDG, GlcADG), glycerophospholipids (AcylPG, AcylPE, AlanylPG, CL, LysylPG, PA, PC, PE, PG, PI, PIP, PIP2, PIP3, PS), lysoglycerophospholipids (LPA, LPC, LPE, LPG, LPI, LPS, LCL), and free fatty acids (FA). Lipid abbreviations follow the standards established by LIPID MAPS (see Table S1 in the Supporting Information for lipid class abbreviations).[6,7] Exact *m/z* values were computed for lipids using a number of commonly observed ESI adducts in positive ($[M]^+$, $[M + H]^+$, $[M + Na]^+$, $[M + K]^+$, $[M + NH_4]^+$, $[M + H\text{-}H_2O]^+$, $[M + 2Na\text{-}H]^+$, $[M + 2\ K]^{2+}$) and negative ($[M\text{-}H]^-$, $[M + HCOO]^-$, $[M + CH_3COO]^-$, $[M + Cl]^-$, $[M\text{-}2H]^{2-}$) modes.

Prediction of CCS Using Machine Learning. Predicted CCS values for lipids were produced using a predictive model trained on the reference lipid database. For all reference lipids, lipid classes, fatty acid modifiers (e.g., "p" indicating a plasmenyl lipid), and MS adducts were each encoded into one-hot binary vectors (22, 3, and 11 features, respectively; see the Supporting Information for specific encodings). Only the lipid classes, fatty acid modifiers, and MS adducts with sufficient representation (at least 20 measurements) in the database were explicitly encoded. The final feature vector was prepared by appending fatty acid sum composition (number of carbons and unsaturations) and observed *m/z* to the binary encoded vectors for each lipid (a total of 39 features; Tables S2−S4). A subset of the reference lipid database (6394 measurements; Table S2) consisting of only the explicitly encoded lipid classes, fatty acid modifiers, and MS adducts was

https://dx.doi.org/10.1021/acs.analchem.0c02560
Anal. Chem. 2020, 92, 14967−14975

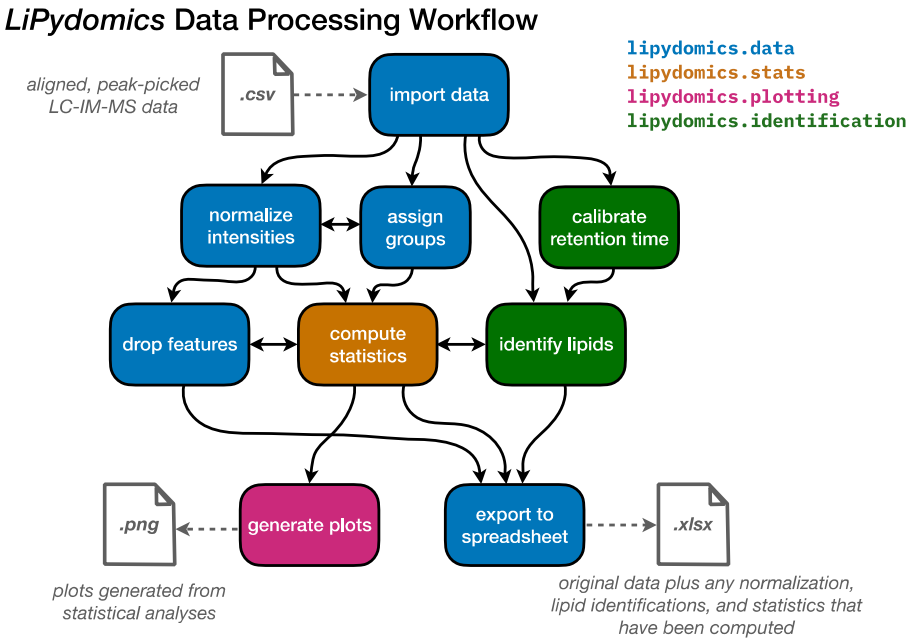


Figure 1. Schematic representation of the LiPydomics data processing workflow. Input/output files (with corresponding file formats) are depicted in gray. Each cell represents an individual data processing step, and arrows reflect possible workflow sequences. Each cell is color-coded according to the specific module used to perform each step. The consistent and modular API of LiPydomics allows for data processing workflows to be customized to the needs of a particular experiment.

selected for use in CCS prediction. This subset was randomly split into training and test data sets in proportions of 80 and 20%, respectively, and the test data set was set aside until model training was complete. The training data were scaled such that all features had a variance of 1 to avoid arbitrary overweighting of individual features based on their scale. A support vector machine with radial basis function kernel (svr) was selected for CCS prediction based on preliminary testing, and hyperparameters were optimized using a grid search with fivefold cross-validation on the training data. Using the optimized hyperparameters, the model was trained on the full set of training data, and performance metrics [mean absolute error (MAE), median absolute error (MDAE), median relative error (MDRE), and root mean squared error (RMSE)] were computed on the training data. Finally, the same performance metrics were computed with the trained model on the test set data to validate model performance on unseen data.

Prediction of HILIC Retention Time Using Machine Learning. Predicted HILIC retention times were produced using a predictive model trained on all entries in the reference lipid database that contain HILIC retention times measured using the HILIC method mentioned above (596 lipids in total; Table S5).[11−13] A smaller feature set (26 features) was used for retention time prediction compared with CCS prediction: binary encoded lipid class (22 features), fatty acid modifier (2 features), and sum composition (2 features). The smaller number of lipid classes and fatty acid modifiers present in the feature set are reflective of the fact that this subset represents less than 10% of the complete reference lipids database (596 of 7907 lipids, see Table S5 for specific encodings). In addition, m/z and encoded MS adduct were not included since these do not relate directly to chromatographic retention time. This subset was split into training and test data sets as described above for CCS prediction. A multivariate linear regression model was used for retention time prediction. The model was fit, and performance metrics (MAE, MDAE, and RMSE) were computed using the training data. Finally, performance metrics were computed with the trained model on the test set data to validate model performance on unseen data.

Calibration of HILIC Retention Time. HILIC retention times present in the reference lipid database were measured using an established HILIC method mentioned above,[11−13] and the ML model for predicting retention times was trained on these retention times. In order to be able to compare retention times acquired using other HILIC conditions, a retention time calibration utility was developed and included in the library. This utility uses linear interpolation of known standards to calibrate retention times of a given HILIC gradient to the retention times in the database. Multiple calibration points can be used in order to approximate nonlinear relationships between reference and measured HILIC retention times. This approach offers excellent calibration accuracy and flexibility, without the complications of choosing a fitting function when the relationship is nonlinear. Once a retention time calibration has been set, the calibrated retention time is automatically used for compound identification. To evaluate this calibration strategy, we first examined three different gradients on the same Phenomenex Kinetex HILIC column (100 × 2.1 mm, 1.7 μm) with Solvent A being acetonitrile/water (50/50) with 5 mM ammonium acetate and Solvent B being acetonitrile/water (95/5) with 5 mM ammonium acetate: (1) 0−1 min, 100% B; 1−4 min, 100−90% B; 4−7 min, 90−70% B; 7−8 min, 70% B; 8−9 min, 70−100% B, 9−12 min, 100% B; (2) 0−0.8 min, 100% B; 0.8−1.8 min, 100−90% B; 1.8−2.8 min, 90−70% B; 2.8−3.8 min, 70% B; 3.8−4.8 min, 70−100% B, 4.8−8 min, 100% B; (3) 0−2 min, 100% B; 2−8 min, 100−90% B; 8−14 min, 90−70% B; 14−15 min, 70% B; 15−16 min, 70−100% B, 16−19 min, 100% B. We then examined three different columns from the Phenomenex Kinetex HILIC series (100 × 2.1, 50 × 2.1, or 30 × 2.1 mm; 1.7 μm). The gradients for

https://dx.doi.org/10.1021/acs.analchem.0c02560
Anal. Chem. 2020, 92, 14967−14975

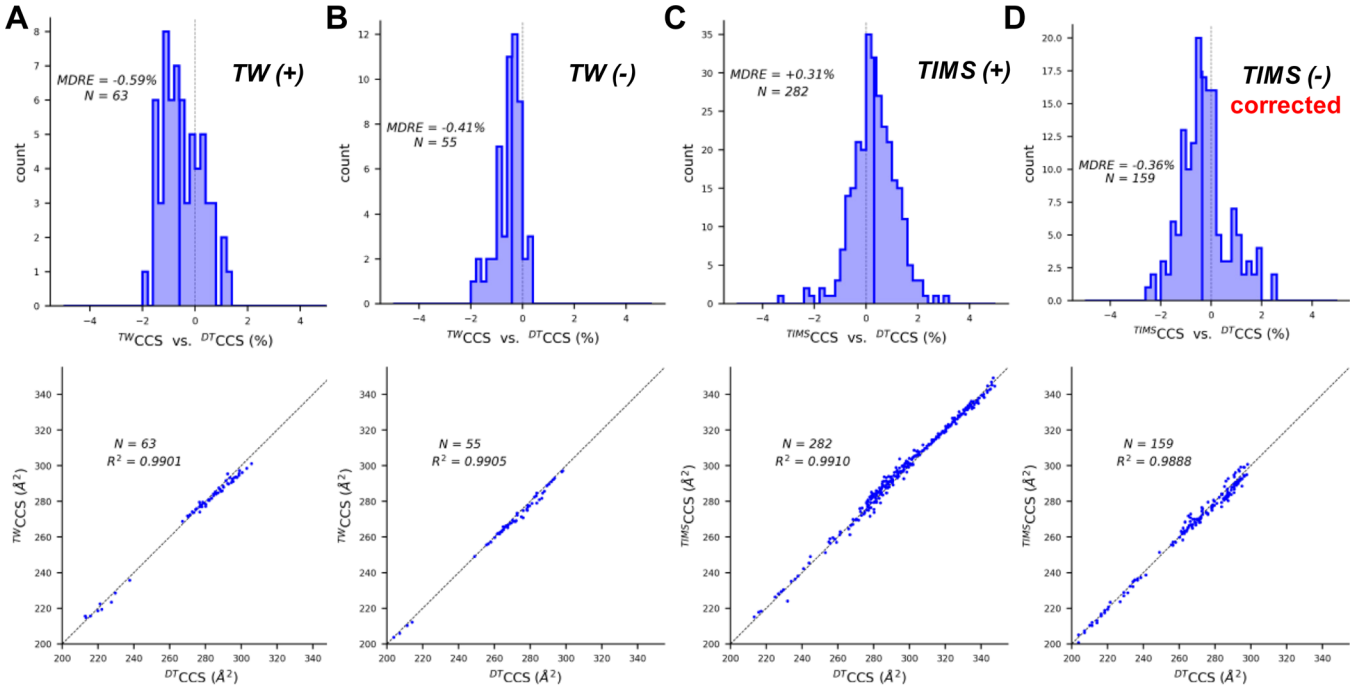


Figure 2. Comparisons of (A, B) ^{TW}CCS and (C, D) ^{TIMS}CCS vs ^{DT}CCS values for lipids in the experimental database. Histograms and CCS-CCS plots provided for the comparisons of the following groups to corresponding overlapping DT values: (A) TW positive mode, (B) TW negative mode, (C) TIMS positive mode, and (D) TIMS negative mode with linear corrections applied. Dotted lines show the linear equation $y = x$.

different columns were changed in linear relation to their lengths. Specifically, the gradients for 50 and 30 mm columns were as follows: (1) 0−0.5 min, 100% B; 0.5−2 min, 100−90% B; 2−3.5 min, 90−70% B; 3.5−4 min, 70% B; 4−4.5 min, 70−100% B, 4.5−6 min, 100% B; (2) 0−0.3 min, 100% B; 0.3−1.2 min, 100−90% B; 1.2−2.1 min, 90−70% B; 2.1−2.4 min, 70% B; 2.4−2.7 min, 70−100% B, 2.7−3.6 min, 100% B.

Statistical and Multivariate Analyses for Lipidomics Data. All statistical and multivariate analyses implemented in this library are available from the SciPy[27] and Scikit-Learn[28] Python libraries, respectively. These analyses use either the raw or normalized intensities from samples belonging to user-specified groups, and the computed statistics are automatically stored along with the data set. The analyses generally fall into two categories: untargeted and targeted. The untargeted analyses (ANOVA and PCA) can be computed on two or more groups in an unsupervised fashion, that is, they report on intrinsic characteristics of the data used in their calculation. The targeted analyses [Pearson correlation analysis, PLS-DA, Log_2(fold-change)] are performed between two specified groups in a supervised fashion, where features that differ between the specified groups are highlighted. In addition, partial least-squares regression analysis may be performed in order to find correlations between lipidomic data and an external continuous variable.

RESULTS AND DISCUSSION

Development of an All-in-One Python Package for Comprehensive Lipidomics: LiPydomics. To enable efficient and reproducible analysis of HILIC-IM-MS data, we developed a free and open-source (MIT license) Python package, LiPydomics. The library contains several modules, each responsible for handling different aspects of lipidomics data analysis (Figure 1). The *data* module is responsible for the organization and storage of the lipidomics data set itself, along with relevant metadata and any statistics calculated on the data set using the *stats* module. It also contains utilities for saving/loading a data set to file, exporting to a spreadsheet, and normalizing intensities. The *stats* module contains functions for applying statistical and multivariate analyses [ANOVA *p*-value, Pearson correlation, PCA, PLS-DA, partial least-squares regression analysis, two-group Log2(fold-change)] on the data set, and the *plotting* module contains functions for extracting data and generating standard plots, such as bar graph and heatmap, for the visualization of the data set and statistical analyses. The *identification* module is used for calibrating HILIC retention times and identifying lipid features at various confidence levels using *m/z*, HILIC retention times, and CCS, and contains utilities for accessing and retraining the CCS and HILIC retention time predictive models as discussed below. The *identification* module additionally contains a subpackage, *LipidMass*, which allows for easy generation of exact masses for a large selection of lipid classes. The *interactive* module contains a user-friendly text-based interface for performing lipidomics data analysis (see the Supporting Information, Figure S1). This entire package, including the interface, can be easily installed on any computer with a compatible Python interpreter (version 3.5 or greater). The assembly of the experimental database, development of CCS and retention time prediction models, the assembly of the predicted database, and demonstration of various modules are discussed in the following sections. A more in-depth overview of the library structure and function is available in the package documentation on GitHub (https://github.com/dylanhross/lipydomics).

Assembly of an Experimental Reference Lipid Database. A database of experimental reference lipid CCS values was assembled from data sets available in the literature.[11−18] In total, 7907 experimental CCS values were included in the database, representing 45 lipid classes (Table S6) and covering

https://dx.doi.org/10.1021/acs.analchem.0c02560
Anal. Chem. 2020, 92, 14967−14975

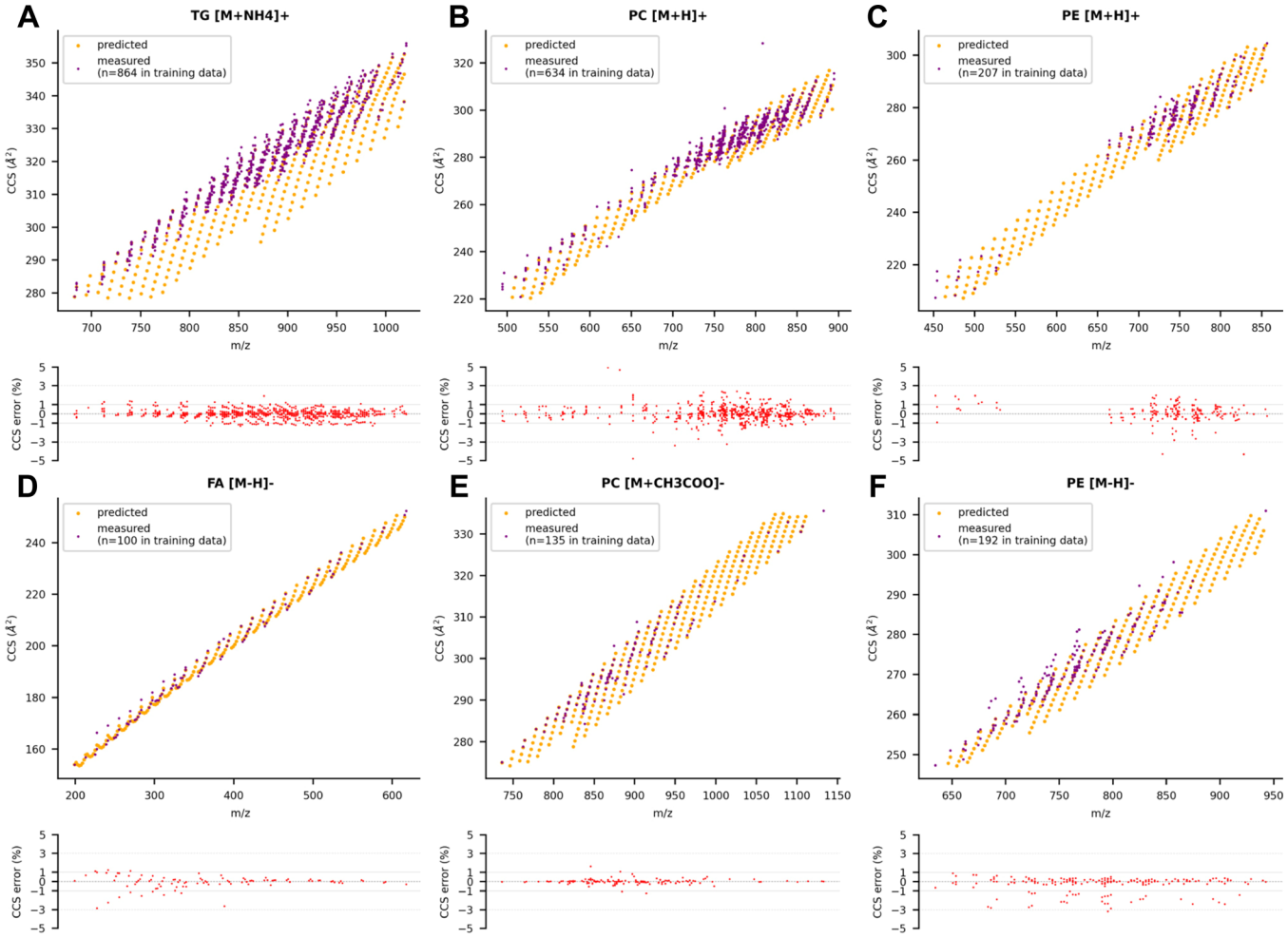


Figure 3. Predicted (gold) and measured (purple) lipid CCS values and relative prediction errors for abundant lipid species in the lipid CCS database in (A–C) positive and (D–F) negative ESI modes.

major lipid species present in both mammalian and bacterial systems. The database covers a variety of MS adducts with 5110 positive mode measurements and 2797 negative mode measurements. CCS measurements made on DTIM, TWIM, and TIMS instruments were included in the database (1285, 596, and 6026 values, respectively). Excellent agreement has already been demonstrated between measurements made on DT and TW platforms when lipid calibrants are used to calibrate CCS values in TW measurements.[29] However, a systematic comparison of TIMS[16,18] CCS values against the established DT method has not yet been performed. To this end, we assessed the agreement between CCS values of overlapping lipids present in TW and TIMS data sets relative to DT values (Figure 2). Both positive and negative mode TW CCS values (Figure 2A,B) show excellent agreement with DT values as evidenced by median relative errors (MDRE) much less than 1% and high degrees of correlation in CCS-CCS plots. Positive mode TIMS CCS values also showed excellent agreement with DT values (Figure 2C); however, negative mode TIMS values (Figure S2A) displayed an MDRE of ∼1% with two apparent populations in the histogram. Negative mode TIMS CCS values from the two constituent data sets[16,18] were examined separately (Figure S2B,C), and it was found that both data sets displayed MDREs >1%, but in opposite directions. The CCS-CCS plots indicated distinct linear relationships between these TIMS CCS values and DT values for the two data sets. Therefore, in order to utilize both data sets for building the CCS prediction model, we applied a linear correction to each data set toward DT values using equations shown in Figure S2 prior to ML model training. After this correction, the MDRE for negative mode TIMS CCS values is −0.36% (Figure 2D). Overall, this database represents comprehensive coverage of currently available experimental lipid CCS values, with a broad representation of lipid classes and IM-MS platforms. A particular strength of this comprehensive lipid database is the extended coverage of bacterial lipids, such as LysylPGs, AlanylPGs, AcylPGs, AcylPEs, GlcADG, and doubly charged lipids, such as CLs and LCLs, which were not covered in previous large-scale lipidomics data sets that contain mostly mammalian lipids.[14,16]

Performance of CCS Prediction Using Machine Learning. An ML model was trained on data from the experimental lipid database to predict CCS values using only a minimal feature set consisting of encoded lipid class, fatty acid composition, encoded MS adduct, and *m/z*. These features do not require computation, which make them easy to assemble for a wide range of lipids and avoids reproducibility issues regarding structural assignment and descriptor generation. It has also been demonstrated that lipids display distinct trends in CCS with respect to *m/z*, lipid class, MS adduct, and acyl chain composition (visit CCSbase.net for interactive visualization of such trends),[11,14,17,19] supporting their inclusion in our

https://dx.doi.org/10.1021/acs.analchem.0c02560
Anal. Chem. 2020, 92, 14967−14975

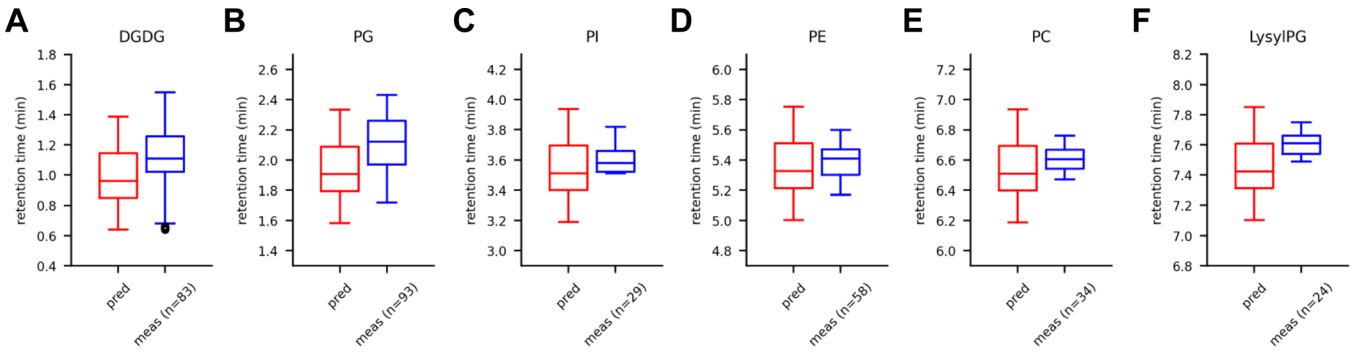


Figure 4. Distributions of predicted (red) and measured (blue) HILIC retention times for major lipid classes (A: DGDG; B: PG; C: PI; D: PE; E: PC; F: LysylPG) spanning the retention time range of the established HILIC method described in the Experimental Section.

minimal feature set. Lipid classes of the same MS adducts with at least 20 measurements, resulting in 6394 CCS values in 22 lipid classes, were included for building the prediction model. This selected subset of measurements was split in an 80/20 proportion for training and test data sets, respectively. The predictive model was trained using support vector regression with a radial basis function kernel as described in the Experimental Section. This model was able to predict CCS values for lipids with high accuracy, achieving MAE, MDAE, and RMSE scores of 1.05, 0.55, and 1.79 Å^2, respectively, on the training data set and 1.34, 0.78, and 3.03 Å^2, respectively, on the test data set. Our model slightly outperformed a recently reported lipid-specific CCS prediction model trained on TIMS CCS values,[16] which achieved RMSE scores of 1.4 and 2.8 Å^2 on their training and test set data, respectively. With MDRE scores of 0.20 and 0.27% on the training and testing data, respectively, our model also modestly outperforms the established Lipid CCS predictor, which achieved MDRE scores of 0.50 and 0.42%, respectively, on positive and negative mode intralab external validation sets (i.e., data not seen during model training).[14] Relative standard deviation (RSD) was computed for 1667 lipid species having multiple reported CCS measurements in the combined CCS database (CCS was corrected as described above for negative mode data from Vasilopoulou et al. and Tsugawa et al.[16,18]), and the mean and median RSD for this group were 0.60 and 0.50%, respectively. Thus, the performance of our predictive model (specifically by MDRE) also compares favorably with variance in experimentally measured CCS values. Figure 3 shows CCS versus *m/z* plots for MS adducts of several major lipid classes in both positive and negative modes along with corresponding relative errors of predicted CCS values relative to available measured values, where predicted values were produced using the ML model and measured values are taken from the experimental lipid CCS database. The predicted CCS and theoretical *m/z* values for all lipids span a comprehensive range of fatty acyl chain lengths and unsaturation degrees, with clear structural trends visible in this space as a function of both characteristics. The predicted CCS values for these lipid classes generally show excellent agreement with the measured values, with residual CCS of predicted values falling mostly within 1% of measured values for most lipid species. We note that although there are some outliers in the measured values (possibly attributable to misidentified lipids), the contribution of these outliers to the training of the overall prediction model appears to be minimum as the majority of the consistent data outweigh the small number of outliers during model training. Plots for additional abundant lipid classes are available in the Supporting Information, Figure S3. These results demonstrate that high-quality lipid CCS predictions can be obtained using a relatively small but specialized feature set, which includes lipid-specific information, such as lipid class, sum fatty acid composition, and fatty acid modifiers (Tables S2−S4), with sufficient training data. Using these specialized features also allows easy expansion of the prediction model as experimental data for additional lipid classes becomes available since these features are easy to generate without computational effort.

Performance of HILIC Retention Time Prediction Using Machine Learning. A separate ML model was trained on data from the reference lipid database using a smaller feature set (minus the adduct types; see the Experimental Section) to predict HILIC retention times based on the HILIC-IM-MS method established previously.[11−13] The trained predictive model achieved MAE, MDAE, and RMSE scores of 0.11, 0.08, and 0.15 min on the test set data, respectively. Figure 4 shows the distributions of predicted and measured retention times for the lipid classes that are well represented in the database spanning the retention time range of the established HILIC method. The predicted HILIC retention times show excellent agreement with measured values for all of these abundant lipid classes, and good agreement with values for less represented lipid classes (Figure S4). To allow the retention time database broadly applicable for HILIC methods run with different gradients and on different columns, we implemented a calibration method using multiple segments of linear interpolation between calibrants. To demonstrate this utility, retention times of lipids extracted from a *Staphylococcus aureus* strain were measured using the established HILIC method,[12] as well as modified methods (see the Experimental Section) using columns of different lengths (Figure 5A) and/or different gradients (Figure 5B). For each set of conditions, two to four individual lipids were used as calibrants to convert measured retention times to reference retention times. The lines in these plots represent the linear interpolation that occurs between the calibrants, and their overlap with the rest of the lipids not used for calibration demonstrates the utility and accuracy of this flexible retention time calibration scheme.

Assembly of a Predicted Lipid Database. Separate data tables were added to the reference lipid database containing predicted *m/z*, CCS, and HILIC retention time for a large collection of lipid species (145,388) comprising a broad representation of lipid classes found in both mammalian and bacterial systems. These predicted data were produced by systematic enumeration of fatty acyl chain length (from 10 to 30 carbons per fatty acid, including both even and odd

https://dx.doi.org/10.1021/acs.analchem.0c02560
Anal. Chem. 2020, 92, 14967−14975

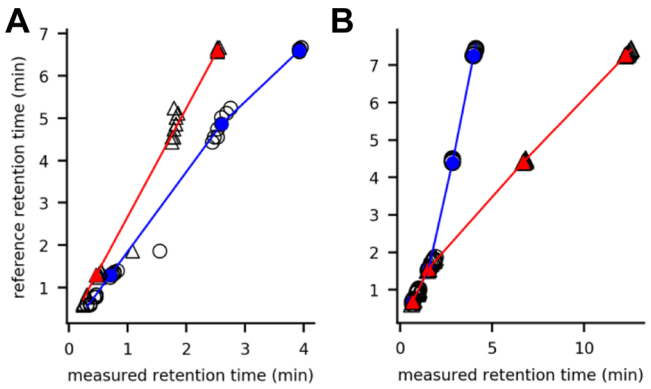


Figure 5. Demonstration of linear interpolation retention time calibration using data collected with columns of (A) different lengths or (B) different gradients. Open circles and triangles in (A) represent measured retention times from experiments using 50 and 30 mm columns, respectively, plotted against retention time from the established HILIC method (100 mm column). Open circles and triangles in (B) represent measured retention times from experiments using a faster and slower gradient, respectively, plotted against retention time from the established HILIC method using the same 100 mm column. Solid colored points represent the individual lipids chosen as calibrants, with colors distinguishing between the two experiments. The colored lines reflect the linear interpolation between calibrants that used for converting measured retention times to their reference equivalent.

numbers) and unsaturations (from 0 to 6 per fatty acid) for 31 lipid classes (see Table S6) defined in the *LipidMass* module in *LiPydomics* (*LipidMass,* see below) and using ML models trained to predict HILIC retention time and CCS; 94,451 and 106,020 predicted CCS and HILIC retention time values were generated, respectively, covering 22 and 23 lipid classes, respectively. Together, this predicted database vastly expands the coverage and depth of the reference lipid database and enables identifications of more lipid species than using the experimental reference data alone.

Automated Identification of Lipid Species at Different Confidence Levels. Identification of lipid species is performed by matching *m/z*, retention time, and CCS against values from the reference lipid database. Lipid identifications can be made at several levels of confidence based on the number of components used for the identification and whether these were compared against experimental or predicted values. The available identification levels in this package are (in descending order of confidence): measured *m/z*, retention time, and CCS; predicted *m/z*, retention time, CCS; measured *m/z* and retention time; predicted *m/z* and retention time; measured *m/z* and CCS; predicted *m/z* and CCS; measured *m/z*; and predicted *m/z*. The user may specify one of these confidence levels when undertaking lipid identification or use a tiered approach, where the highest confidence level is tried first for each lipid species and successive levels are attempted until an identification is made. If retention time calibration has been set up, the calibrated retention time is automatically used for lipid identification. Whenever lipid identifications are made, both the putative identification(s) and the level of confidence are stored for each lipid feature. When multiple annotations are made for a single feature, the putative identifications are ranked by a score reflecting the agreement between query and reference values, computed as the dot product of residuals from the matched values normalized by their respective search tolerances. All lipid identifications made by this method are of LSI Level 3,[9] i.e., lipid class, subclass, and fatty acid sum composition. Overall, this utility allows users to identify lipids in an efficient, automated fashion. In addition, the predicted lipid database was added to our existing web interface (https://CCSbase.net)[19] so that users can query these data without using the complete LiPydomics package.

Demonstration of LiPydomics Functionality. In order to demonstrate the functionality of LiPydomics, we reanalyzed data from our recently published study examining lipidomic changes associated with antibiotic resistance in methicillin-resistant *Staphylococcus aureus* (MRSA) strains.[30] Aligned and peak-picked HILIC-IM-MS data acquired in negative ESI mode were used for this analysis. The data contained normalized intensities for 3647 features from four different MRSA strains (JE2 parent strain, "Par"; JE2-derived strain with reduced susceptibility to daptomycin, "Dap2"; reduced susceptibility to dalbavancin, "Dal2"; and reduced susceptibility to vancomycin, "Van4"), each with four biological replicates. Lipids were identified by matching on predicted *m/z*, retention time, and CCS (using search tolerances of 0.02 Da, 0.2 min, and 3.0%, respectively), or measured *m/z* and CCS to cover lipid classes without retention time information. Using the *stats* module, we computed a three-component PCA to see how the groups separated according to their overall variance. Figure 6A shows the PCA projections for each sample along the first two principal components, colored by strain. These components capture around 90% of the total variance in the data set, and samples from each group cluster together and separate from other groups in this space, indicating that there are distinct characteristics that are associated with each strain. We next looked specifically at the comparison between the lipid profiles of the daptomycin-resistant Dap2 and the parent strains that have been examined previously. First, we performed PLS-DA and Pearson correlation between Dap2 and Par. The PLS-DA projections (Figure 6B) showed excellent separation between the strains, and similar levels of intragroup variance. The S-plot (PLS-DA x-loadings vs. Pearson correlation) highlights multiple features that are highly abundant in either strain and different between strains (Figure 6C). Examination of these discriminating features reveals systematic changes in the DGDG, LysylPG, and FA lipid classes between these strains. To explore these effects at a higher level, we computed the Log2(fold-change) between Dap2 and Par and produced heat maps of all annotated lipids from each of these classes using the *plotting* module (Figure 6D−F). From these heat maps, we observed a general decrease in DGDGs, an increase in LysylPGs, and an increase in FAs between 15 and 21 carbons in length in Dap2 strains relative to Par. It should be noted that these heat maps include lipid features annotated as unsaturated lipids; however, these are unlikely to be found in the bacterial system studied. Indeed, a close examination of those features suggests that most have low signal intensities likely corresponding to background signals. We also produced bar plots using the *plotting* module, showing the mean intensities with standard deviation in Dap2 and Par strains for the most significantly altered lipids in each of the previously discussed lipid classes (Figure 6G−I). Overall, this analysis using LiPydomics reproduced the key findings of the previous report[30] and was performed with only 19 lines of Python code on minimally processed data.

Separately, we used both positive and negative ESI mode data from the same study to perform lipid identification using

https://dx.doi.org/10.1021/acs.analchem.0c02560
Anal. Chem. 2020, 92, 14967−14975

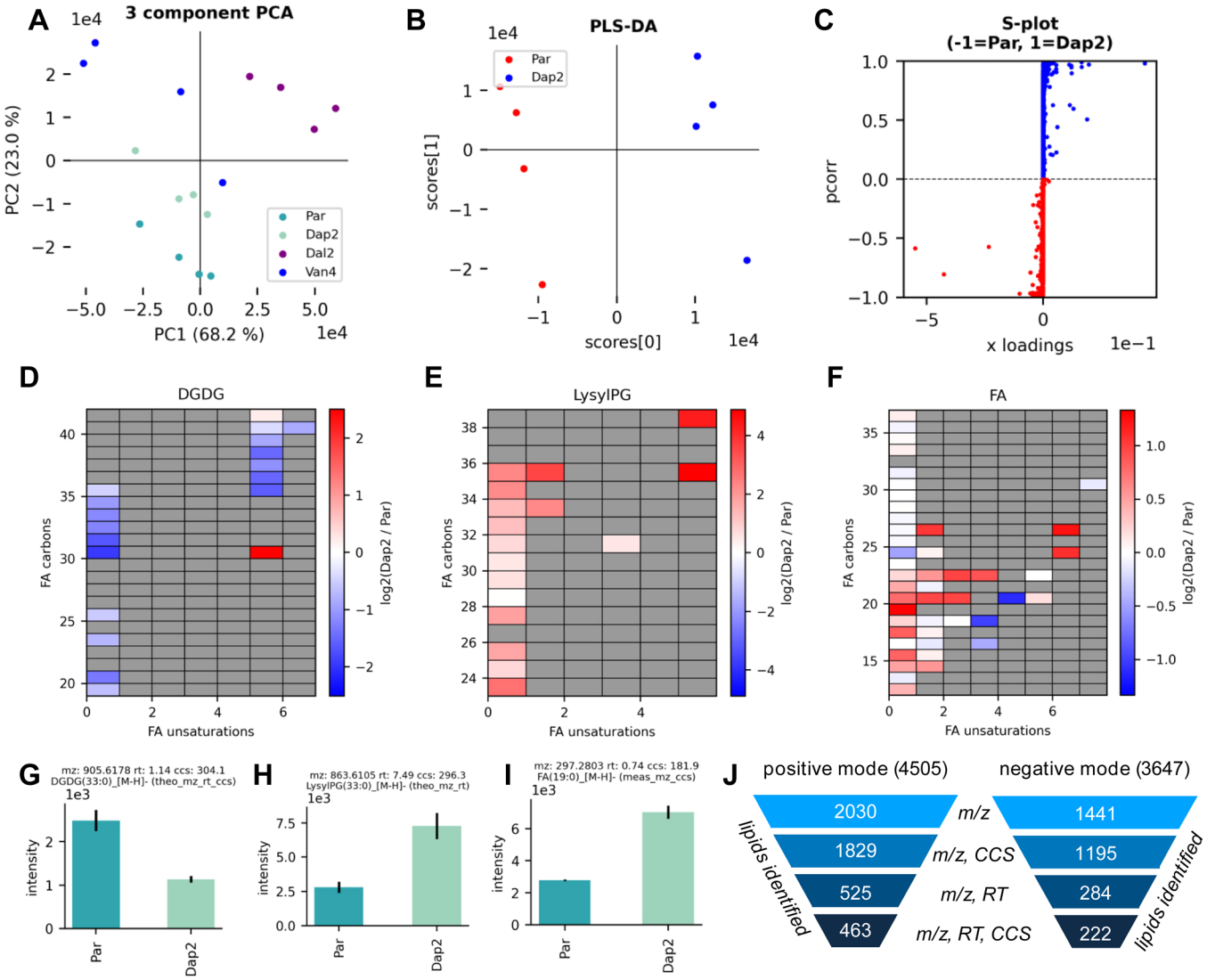


Figure 6. Illustration of LiPydomics functions by analyzing antibiotic-resistant MRSA strains. (A) PCA projections for parent strain (Par) and strains with resistance to daptomycin, dalbavancin, or vancomycin (Dap2, Dal2, Van4, respectively). (B) PLS-DA projections computed between Par (red) and Dap2 (blue) strains. (C) S-plot showing individual features driving separation between Par (red) and Dap2 (blue) strains. (D-F) Heatmaps of Log_2(fold-change) between Par and Dap2 strains for major bacterial lipid classes. (G-I) Bar plots of individual lipids displaying the most significant differences between Par and Dap2 strains. (J) Number of lipids identified from positive and negative mode data using various combinations of predicted identifiers (m/z, CCS, and/or HILIC retention time).

the predicted lipid database at varying levels of confidence. Figure 6J shows the number of lipids identified at each level of confidence for both ESI modes. The number of lipids identified decreases steadily as we progress from matching solely based on *m/z* (lowest confidence) to matching based on *m/z*, retention time, and CCS (highest confidence), using search tolerances of 0.02 Da, 0.2 min, and 3.0%, respectively, across all tests. This example demonstrates the flexibility of the lipid identification utility in LiPydomics, which allows a user to prioritize annotation coverage or confidence as it suits the biological problem being studied.

CONCLUSIONS

The key strengths of LiPydomics as a resource for lipidomics data analysis lie in its large coverage of lipid classes (both experimental and predicted), versatility (from statistical analysis to identification), reproducibility, extensibility, and ease of use. In addition, data analyses can be partially or fully automated through scripting, further enhancing the reproducibility and efficiency of such analyses. The unique reference lipid database contains measured and predicted *m/z*, retention time, and CCS values, with broad coverage of common and rare lipid species from both mammalian and bacterial systems, the latter being underrepresented in other lipid databases to date. The predicted *m/z*, retention times, and CCS values display good agreement with measured values and cover a comprehensive range of lipid classes and fatty acid compositions, enabling identification of more lipids than would be possible using measured values alone. Thus, this comprehensive lipid database enables the identification of lipid species at the level of class, subclass, and sum fatty acid composition (LSI Level 3) from diverse biological systems. The package (including the lipid database and prediction models) is also built to be highly extensible and customizable, allowing easy expansion as more data becomes available and optimization for specific analysis workflows via its flexible and

https://dx.doi.org/10.1021/acs.analchem.0c02560
Anal. Chem. 2020, 92, 14967−14975

well-documented API. The text-based user interface makes the library more broadly accessible to those who are not familiar with Python programming. Together, these attributes make LiPydomics a unique and comprehensive tool for performing analysis of HILIC-IM-MS lipidomic data.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.0c02560.

Lipid class full names and abbreviations; binary encodings for CCS and HILIC retention time prediction; overview of the *Lipydomics* architecture; screenshots of the interactive interface; additional data on performance of CCS and retention prediction; and CCS and retention time training and test data sets (PDF)

Lipydomics_SI_CCS_training_and_test_datasets (XLSX)

Lipydomics_SI_RT_training_and_test_datasets (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Libin Xu − *Department of Medicinal Chemistry, University of Washington, Seattle, Washington 98195, United States;* orcid.org/0000-0003-1021-5200; Phone: (206) 543-1080; Email: libinxu@uw.edu; Fax: (206) 685-3252

Authors

Dylan H. Ross − *Department of Medicinal Chemistry, University of Washington, Seattle, Washington 98195, United States*

Jang Ho Cho − *Department of Medicinal Chemistry, University of Washington, Seattle, Washington 98195, United States*

Rutan Zhang − *Department of Medicinal Chemistry, University of Washington, Seattle, Washington 98195, United States*

Kelly M. Hines − *Department of Medicinal Chemistry, University of Washington, Seattle, Washington 98195, United States;* orcid.org/0000-0001-9125-5268

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.0c02560

Author Contributions

D.H.R. and J.H.C. developed the Python package. R.Z. measured HILIC retention times for lipids under different conditions. K.M.H. performed lipidomic characterization of antibiotic-resistant MRSA strains. The manuscript was written through the contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health's grants (R01HD092659, R01AI136979, and R01GM12757), UW CoMotion Innovation Gap Fund, and startup funds from the Department of Medicinal Chemistry to L.X. The authors would like to thank Emily Pruitt and Amy Li for their constructive feedback during the development of *LiPydomics*.

■ REFERENCES

(1) Harayama, T.; Riezman, H. *Nat. Rev. Mol. Cell Biol.* **2018**, *19*, 281−296.
(2) Wymann, M. P.; Schneiter, R. *Nat. Rev. Mol. Cell Biol.* **2008**, *9*, 162−176.
(3) Tumanov, S.; Kamphorst, J. J. *Curr. Opin. Biotechnol.* **2017**, *43*, 127−133.
(4) Rustam, Y. H.; Reid, G. E. *Anal. Chem.* **2018**, *90*, 374−397.
(5) Tu, J.; Zhou, Z.; Li, T.; Zhu, Z.-J. *TrAC Trends Anal. Chem.* **2019**, *116*, 332−339.
(6) Fahy, E.; et al. *J. Lipid Res.* **2005**, *46*, 839−862.
(7) Fahy, E.; et al. *J. Lipid Res.* **2009**, *50 Suppl*, S9−S14.
(8) Liebisch, G.; Vizcaíno, J. A.; Kofeler, H.; Trotzmüller, M.; Griffiths, W. J.; Schmitz, G.; Spener, F.; Wakelam, M. J. *J. Lipid Res.* **2013**, *54*, 1523−1530.
(9) Ryan, E.; Reid, G. E. *Acc. Chem. Res.* **2016**, *49*, 1596−1604.
(10) Kyle, J. E.; et al. *Analyst* **2016**, *141*, 1649−1659.
(11) Hines, K. M.; Herron, J.; Xu, L. *J. Lipid Res.* **2017**, *58*, 809−819.
(12) Hines, K. M.; Waalkes, A.; Penewit, K.; Holmes, E. A.; Salipante, S. J.; Werth, B. J.; Xu, L. *mSphere* **2017**, *2*, e00492−e00417.
(13) Hines, K. M.; Xu, L. *Chem. Phys. Lipids* **2019**, *219*, 15−22.
(14) Zhou, Z.; Tu, J.; Xiong, X.; Shen, X.; Zhu, Z. J. *Anal. Chem.* **2017**, *89*, 9559−9566.
(15) Blaženović, I.; Shen, T.; Mehta, S. S.; Kind, T.; Ji, J.; Piparo, M.; Cacciola, F.; Mondello, L.; Fiehn, O. *Anal. Chem.* **2018**, *90*, 10758−10764.
(16) Tsugawa, H.; et al. *Nat. Biotechnol.* **2020**, *38*, 1159−1163.
(17) Leaptrot, K. L.; May, J. C.; Dodds, J. N.; McLean, J. A. *Nat. Commun.* **2019**, *10*, 985.
(18) Vasilopoulou, C. G.; Sulek, K.; Brunner, A. D.; Meitei, N. S.; Schweiger-Hufnagel, U.; Meyer, S. W.; Barsch, A.; Mann, M.; Meier, F. *Nat. Commun.* **2020**, *11*, 331.
(19) Ross, D. H.; Cho, J. H.; Xu, L. *Anal. Chem.* **2020**, *92*, 4548−4557.
(20) Zhou, Z.; Shen, X.; Tu, J.; Zhu, Z. J. *Anal. Chem.* **2016**, *88*, 11084−11091.
(21) Bijlsma, L.; Bade, R.; Celma, A.; Mullin, L.; Cleland, G.; Stead, S.; Hernandez, F.; Sancho, J. V. *Anal. Chem.* **2017**, *89*, 6583−6589.
(22) Soper-Hopper, M. T.; Petrov, A. S.; Howard, J. N.; Yu, S.-S.; Forsythe, J. G.; Grover, M. A.; Fernández, F. M. *Chem. Commun.* **2017**, *53*, 7624−7627.
(23) Mollerup, C. B.; Mardal, M.; Dalsgaard, P. W.; Linnet, K.; Barron, L. P. *J. Chromatogr., A* **2018**, *1542*, 82−88.
(24) Plante, P.-L.; Francovic-Fontaine, É.; May, J. C.; McLean, J. A.; Baker, E. S.; Laviolette, F.; Marchand, M.; Corbeil, J. *Anal. Chem.* **2019**, *91*, 5191−5199.
(25) Zhou, Z.; Shen, X.; Chen, X.; Tu, J.; Xiong, X.; Zhu, Z. J. *Bioinformatics* **2019**, *35*, 698−700.
(26) Ulmer, C. Z.; Koelmel, J. P.; Ragland, J. M.; Garrett, T. J.; Bowden, J. A. *J. Am. Soc. Mass Spectrom.* **2017**, *28*, 562−565.
(27) Virtanen, P.; et al. *Nat. Methods* **2020**, *17*, 261−272.
(28) Pedregosa, F.; et al. *J. Mach Learn. Res.* **2011**, *12*, 2825−2830.
(29) Hines, K. M.; May, J. C.; McLean, J. A.; Xu, L. *Anal. Chem.* **2016**, *88*, 7329−7336.
(30) Hines, K. M.; et al. *J. Antimicrob. Chemother.* **2020**, *75*, 1182−1186.