Large-Scale Prediction of Collision Cross-Section Values for Metabolites in Ion Mobility-Mass Spectrometry
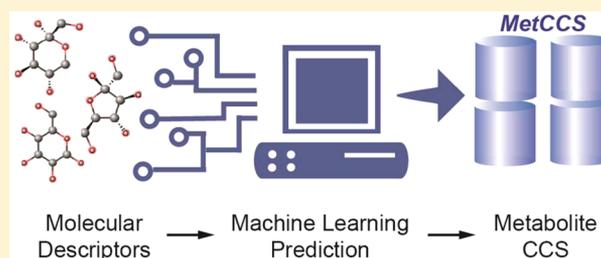
Zhiwei Zhou, Xiaotao Shen, Jia Tu, and Zheng-Jiang Zhu*

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

Ⓢ *Supporting Information*

ABSTRACT: The rapid development of metabolomics has significantly advanced health and disease related research. However, metabolite identification remains a major analytical challenge for untargeted metabolomics. While the use of collision cross-section (CCS) values obtained in ion mobility-mass spectrometry (IM-MS) effectively increases identification confidence of metabolites, it is restricted by the limited number of available CCS values for metabolites. Here, we demonstrated the use of a machine-learning algorithm called support vector regression (SVR) to develop a prediction method that utilized 14 common molecular descriptors to predict CCS values for metabolites. In this work, we first experimentally measured CCS values (Ω_{N2}) of ~400 metabolites in nitrogen buffer gas and used these values as training data to optimize the prediction method. The high prediction precision of this method was externally validated using an independent set of metabolites with a median relative error (MRE) of ~3%, better than conventional theoretical calculation. Using the SVR based prediction method, a large-scale predicted CCS database was generated for 35 203 metabolites in the Human Metabolome Database (HMDB). For each metabolite, five different ion adducts in positive and negative modes were predicted, accounting for 176 015 CCS values in total. Finally, improved metabolite identification accuracy was demonstrated using real biological samples. Conclusively, our results proved that the SVR based prediction method can accurately predict nitrogen CCS values (Ω_{N2}) of metabolites from molecular descriptors and effectively improve identification accuracy and efficiency in untargeted metabolomics. The predicted CCS database, namely, MetCCS, is freely available on the Internet.

I on mobility-mass spectrometry (IM-MS) is a powerful analytical technology which can separate ions rapidly in the gas-phase within a millisecond time frame.[1,2] A number of collisions occurring between ions and inert buffer gas (typically nitrogen or helium) under an electric field result in the differences of drift time (DT).[1] Ion's collision cross-section (CCS) value derived from the drift time is a unique physicochemical property, which is related to the charge, shape, and size of the measured ion and buffer gas.[3–5] Thus, the measurement of an ion's drift time (or CCS value) using IM-MS provides specific structural information on the ion. Recently, IM-MS is becoming a popular analytical tool for many research areas ranging from structural biology,[2] proteomics,[6–8] lipidomics,[9–11] metabolomics,[12–15] to clinical analysis.[16]

The rapid development of metabolomics has largely facilitated biomedical and clinical research.[17,18] However, metabolite identification remains a major analytical challenge for untargeted metabolomics. Current methods for metabolite identification are mainly based on accurate mass, retention time (RT), and MS/MS spectra.[17] Accurate mass can be readily achieved using high-resolution mass spectrometers (such as time-of-flight (TOF), Orbitrap). However, retention time in liquid chromatography (LC) separation is largely affected by the type of columns, mobile phase, gradient, and other factors

and is not easy to be standardized as a physicochemical property. Our previous work has proved that metabolite identification using MS/MS spectra was relatively more accurate and robust.[19,20] Although there are several metabolite MS/MS spectral libraries available (such as METLIN,[21] MassBank[22]), it suffers from the limited number of MS/MS spectra. For example, in METLIN, only 14 034 out of 242 031 metabolites (5.8%) have available MS/MS spectra.[23] Therefore, it becomes more important to use other readily obtained physicochemical properties for metabolite identification.

Recently, measurements of CCS values in metabolomics showed high reproducibility within 2% of precision.[15,24–26] The addition of CCS values to the metabolite library readily increases the identification confidence of metabolites of interest.[15,25,26] Similar to the construction of MS/MS spectral library, it is not feasible for each lab to purchase and measure a large number of metabolite standards to obtain standard CCS values. In addition, the vast majority of metabolite standards are not commercially available. Therefore, there are only a small number of standard CCS values available for metabolites.[15,24,26]
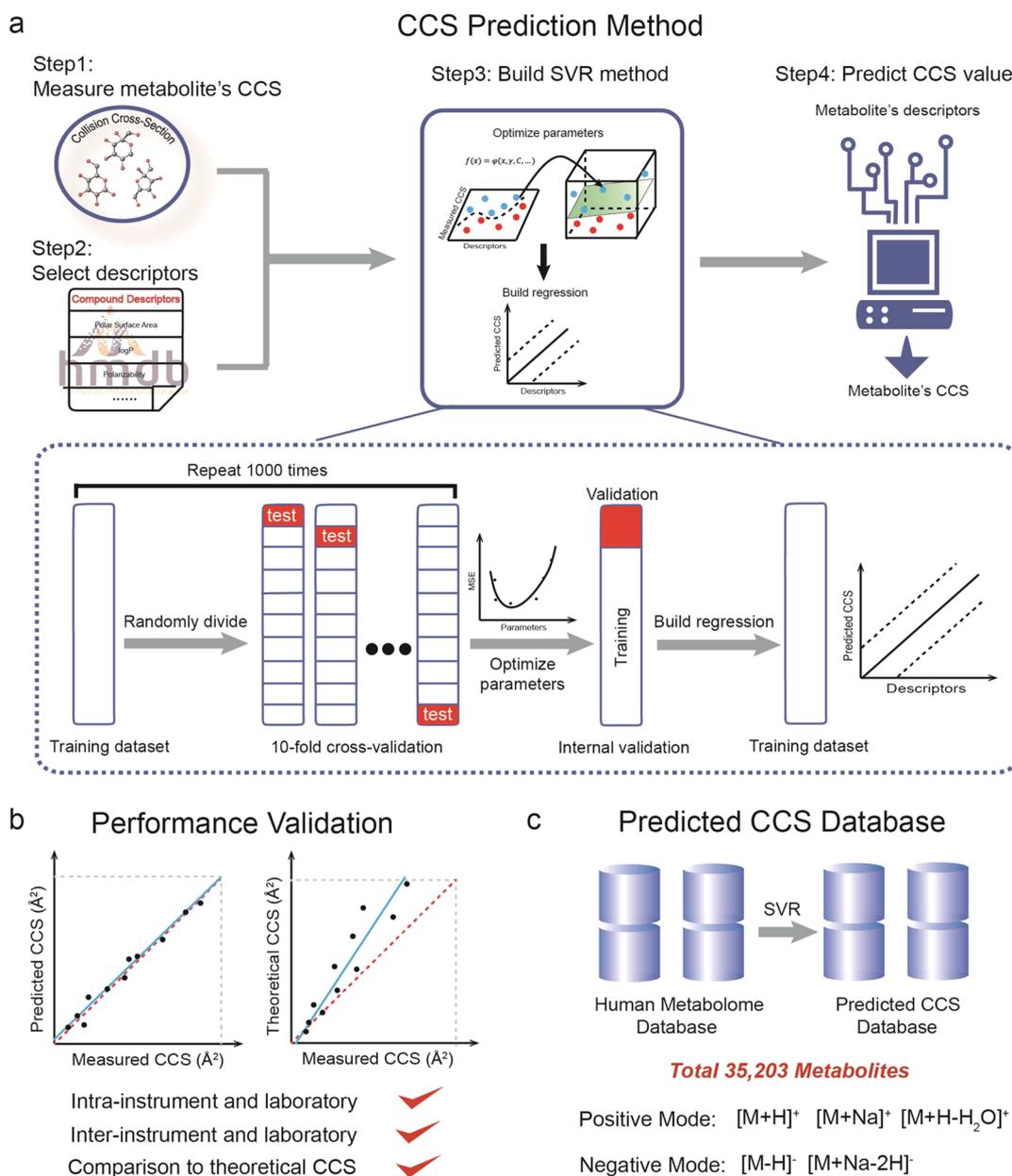
Figure 1. Schematic illustration of CCS value prediction using support vector regression (SVR) algorithm: (a) steps to develop and optimize the SVR based prediction method; (b) validation of prediction performance; (c) generation of a large-scale predicted CCS database containing 35 203 metabolites. Each metabolite has 5 types of adducts for positive and negative polarities.

Another approach to obtain CCS values is using computational chemistry tools like MOBCAL to calculate theoretical CCS values.[27−29] Software simulates the interactions between ions and buffer gas using different modeling methods such as trajectory method (TM),[30] projection approximation (PA),[30] or exact hard sphere scattering (EHSS).[31] In the past decade, theoretically calculated CCS values were widely employed. However, the method required high computational resource and at least several days to complete the calculation of one small molecule. Moreover, compared to the experimentally measured CCS values, estimated relative errors of theoretically calculated CCS values using nitrogen as buffer gas were about 3−30% depending on the modeling method.[15,32,33] Therefore, both precision and efficiency need further improvements for theoretical calculation of CCS values.

Recently, adopting machine-learning based mathematical methods to predict drift times or CCS values was successfully employed for peptides with well-defined amino acid sequences.[34,35] For example, the program imPredict used 134 physicochemical properties of peptides directly from its amino acid sequence, like peptide length and number of nonpolar hydrophobic residues to predict a peptide's drift time (but not CCS values).[34] However, metabolites do not have peptide-like repeating units and own remarkable differences on chemical and structural properties. Meanwhile, time cost and efficiency to obtain tens or hundreds of molecular descriptors for thousands to ten thousands of metabolites is quite challenging for most laboratories. Here, we used a machine-learning algorithm called support vector regression (SVR) to develop a prediction method that utilizes 14 common molecular descriptors for the prediction of metabolite's CCS

values (Figure 1 and Table S1). The molecular descriptors were obtained from the human metabolome database (HMDB, http://www.hmdb.ca/).[36] Specifically, CCS values of 396 and 400 metabolite standards in positive and negative modes, respectively, were first experimentally measured on a drift tube based IM-MS instrument and used as training data to optimize the SVR based prediction method. Then the prediction method was further externally validated using an independent set of metabolites with a median relative error (MRE) of ~3%, significantly better than theoretical calculation. Using this method, we generated a large-scale predicted CCS database containing 35 203 metabolites in HMDB and predicted nitrogen CCS values (Ω_{N2}) were proven to effectively improve the accuracy of metabolite identifications in real biological samples. The predicted CCS database, namely, MetCCS, is freely available on the Internet (http://www.metabolomics-shanghai.org/software.php).

EXPERIMENTAL SECTION

All IM-MS data was measured using a UHPLC system (Agilent 1290 series) coupled to a quadruple time-of-flight mass spectrometer equipped with an ion-mobility drift tube (Agilent DTIM-QTOF-MS 6560, Agilent Technologies). Nitrogen was used for IM separation in this work. The extraction of metabolites from biological samples followed the protocol in our previous publication.[37] Other experimental details about chemicals, preparation of metabolite standards, DTIM-QTOF-MS parameters are provided in the Supporting Information.

Measurements of Experimental CCS Values. All metabolite standards were individually measured in batches, and the design of acquisition batch was provided in Scheme S1 in the Supporting Information. Specifically, every batch has 100 metabolite standards, and Agilent tune mix solution (Tables S2 and S3 in the Supporting Information) was first injected to establish CCS calibration curve for subsequent measurements. During the data acquisition, quality control (QC) samples were repetitively measured every 20 injections to evaluate the instrument performance. The results from QC samples showed that CCS measurements had a relative standard deviation (RSD) less than 0.2% over 6 days during the whole measurement of ~400 metabolite standards, highlighting the good accuracy of the DTIM-QTOF-MS instrument (Figure S1, Tables S4 and S5 in the Supporting Information). The CCS value for each metabolite was calculated using the single-field method.[38,39] A previous report[39] and our results both proved that relative errors between the single-field measured CCS values and the multifield measured CCS values for all metabolites were less than 1% (Figure S2 in Supporting Information). A calibration curve was first established using the Agilent tune mix solution containing 6 compounds with known CCS values. Obtained β and t_{fix} coefficients were applied to samples to calculate CCS values. All calculations were done using IM-MS Browser software (version B.07.01, Agilent Technologies). More details about CCS value calculation using the single-field method are provided in the Supporting Information.

SVR Based Prediction Method. Support vector regression is one of the most widely used machine-learning algorithms. The basic knowledge of SVR algorithm can be found in the literature.[40] Some applications of SVR algorithm in metabolomics[41,42] and analytical chemistry[43] were recently introduced. Specifically in this work, SVR algorithm implicitly maps molecular descriptors of metabolites into a high-dimensional

feature space using a kernel function and constructs a hyperplane in that space to perform the high-dimensional regression between molecular descriptors and CCS values in the training data set (Figure 1a). First, in order to achieve the best regression toward prediction accuracy using SVR algorithm, parameters of kernel function that was used to construct the regression hyperplane needed to be optimized using the training data set. Two common important parameters, cost of constraints violation (C) and gamma (γ), were chosen for optimization (Scheme S2 in the Supporting Information). In total, 85 parameter combinations were evaluated using the training data set via 10-fold cross-validation. The parameter combination with minimal mean squared error (MSE) for prediction was selected for SVR prediction method (Figure S3 in the Supporting Information). Finally, SVR prediction method was built using the whole training data set with the optimized parameter combination. As a result, SVR prediction method can predict the CCS values for metabolites utilizing their molecular descriptors. More details about the SVR prediction method are provided in the Supporting Information. All data processes and calculations were performed using open-source software R (version 3.2.3) and SVR based prediction was performed via the R package e1071 (https://cran.r-project.org/web/packages/e1071).

RESULTS AND DISCUSSION

Develop and Optimize SVR Prediction Method. First, support vector regression algorithm was chosen to develop the CCS prediction method due to its powerful capability of nonlinear regression, and 14 common molecular descriptors were used for prediction. We chose 14 molecular descriptors related to the metabolite's size, shape, charge, polarity, and so on (Table S1 in the Supporting Information). The molecular descriptors of metabolites were obtained from HMDB. Some descriptors like the number of rotatable bonds were recently reported to make a significant contribution to CCS values of small molecules.[28] Descriptor values of each metabolite were first normalized and standardized to z-scores, which benefited numerical stability during the machine-learning process.

Then, we measured CCS values of 396 and 400 metabolites in ESI positive and negative modes, respectively, to serve as a training data set for prediction method development and optimization (Excel files 01 and 02 in the Supporting Information). To the best of our knowledge, this is the largest measurement data set of metabolite chemical standards to obtain experimental CCS values. The chosen metabolites exhibited a wide range of accurate masses (101−1355 Da) and CCS values (117−338 Å^2). In addition, because of the diverse physiochemical properties of metabolites, multiple ionization adducts often appeared, like $[M + H]^+$, $[M + Na]^+$, and $[M + H - H_2O]^+$ for positive ionization. Here, we manually analyzed each metabolite in the training data set and chose one adduct for each metabolite for optimization and development of the prediction method (Figure 2a).

As shown in Figure 1a, utilizing the measured CCS values of metabolites and their molecular descriptors, we optimized parameters of the SVR based prediction method. Then, to validate the performance of the method, 20% of metabolites in the training data set were randomly selected for internal validation, and the rest 80% of metabolites were used to establish the SVR prediction method. For both positive and negative modes, excellent fits with R^2 values of 0.9830 and 0.9813 were obtained, respectively (Figure 2b). Median relative
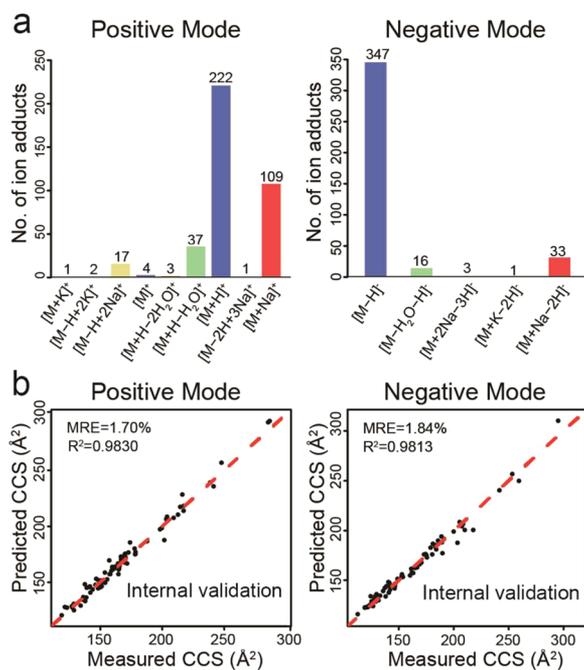
Figure 2. (a) Statistics of chosen metabolite ion adducts in the training data set for positive and negative ionizations. (b) Regression curves between SVR predicted and measured CCS values in internal validation for positive and negative ionizations.

errors (MREs) of the comparison were 1.70% and 1.84%, for data sets of positive and negative modes, respectively. The results proved that these optimized parameters for SVR prediction method performed well for CCS value prediction. Then, SVR prediction method was finally built based on the whole training data set using optimized parameters. For all 396

metabolites in positive mode, median relative error was 1.60%, and R^2 value was 0.9921 (Table S6). The q^2 value of 0.9691 was obtained via 10-fold cross-validation during method development, suggesting that these results were reliable rather than overfitting. Similar parameter optimization and method development were performed on negative mode using molecular descriptors and CCS values from 400 metabolites, where median relative error was 1.62% and R^2 value was 0.9911 (Table S6).

During ESI ionization, one metabolite may form several adducts. The charge site and shape of ions may change among different adducts and cause subtle differences to CCS values. In order to confirm the effective performance of our SVR method for the prediction for CCS values derived from different adducts, we randomly chose 43 metabolites with 111 different adducts in positive polarity (Data was provided in the Supporting Information) and compared the measured CCS values of 111 adducts from SVR prediction (Figure S4). A good linear regression was observed with an R^2 value of 0.9655 between predicted and measured CCS values. The median relative error was 1.73%, indicating the excellent prediction performance of our SVR method for different adducts. Similar results were obtained in negative mode, which included 57 metabolites with 136 adducts. Here, the inclusion of uncommon adducts for comparison, like $[M - H_2O - H]^-$, $[M + 2Na - 3H]^-$, $[M + K - 2H]^-$, aims to demonstrate the capability of SVR method for the prediction of different ionization adducts. Overall, these results showed the validity of SVR based CCS prediction using the complex training data set and 14 common molecular descriptors.

External Validation and Prediction Performance. We further evaluated the performance of the SVR prediction method using an independent set of metabolites (78 and 79 metabolites in positive and negative modes, respectively) other
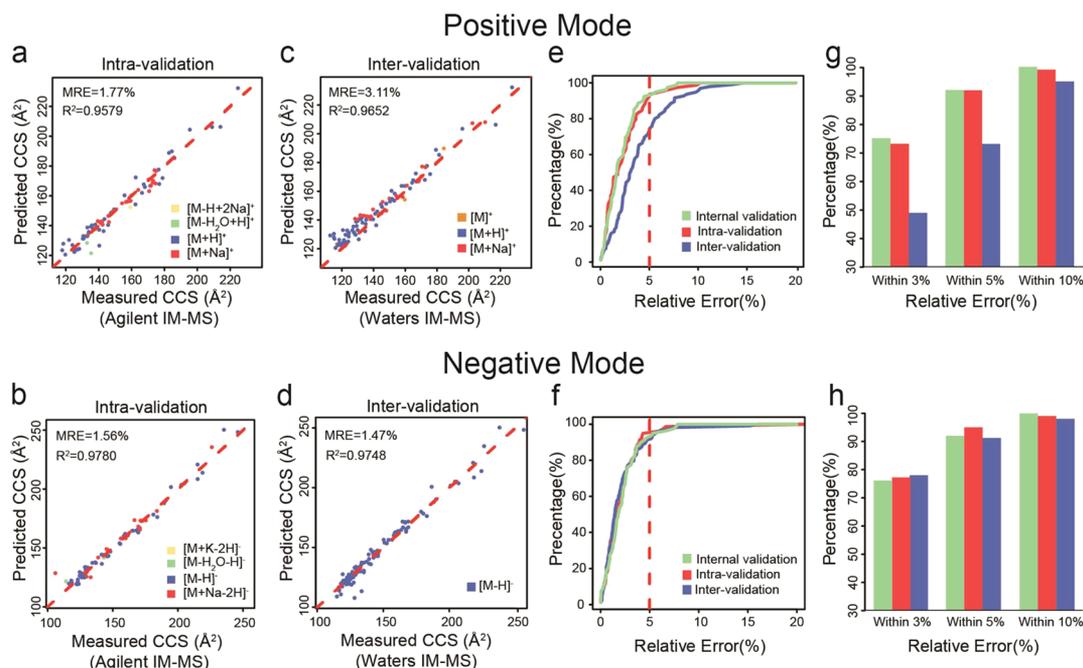

Figure 3. External validations of CCS prediction across instruments and laboratories: (a,b) regression curves between predicted and measured CCS values in intrainstrument and intralab validations in positive (a) and negative modes (b); (c,d) regression curves between predicted and measured CCS values in interinstrument and interlab validations in positive (c) and negative modes (d); (e,f) cumulative curves of relative errors across all validations in positive (e) and negative modes (f); (g,h) bar graphs of relative errors across all validations in positive (g) and negative modes (h).
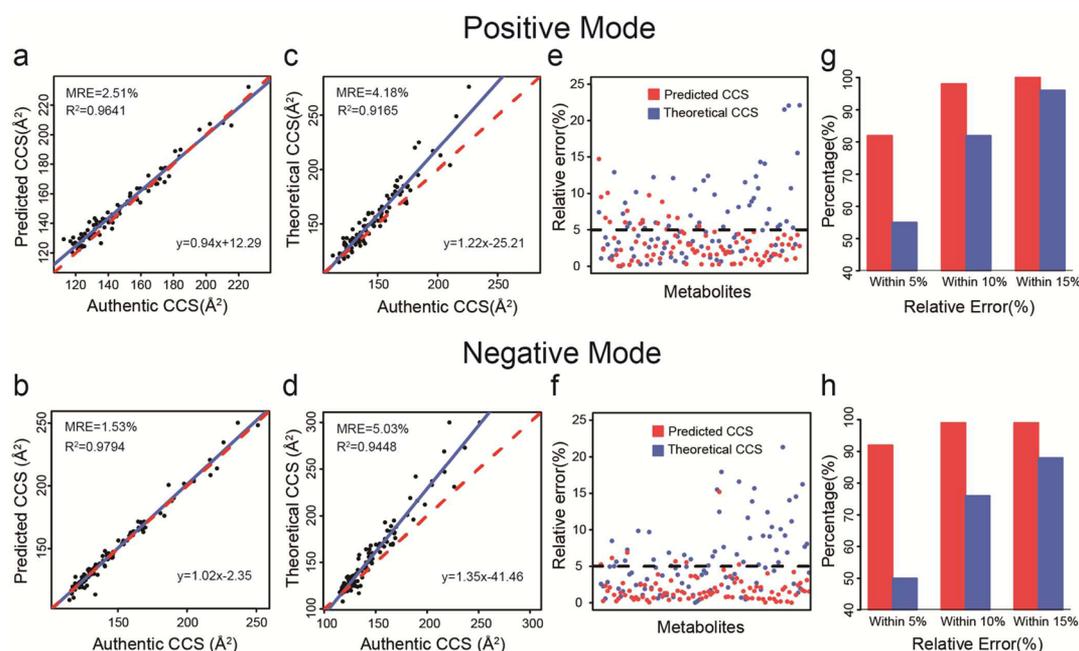
Figure 4. Precision comparison between SVR prediction method and theoretical calculation method: (a,b) regression curves between SVR predicted and authentic CCS values in positive (a) and negative (b) modes; (c,d) regression curves between theoretical and authentic CCS values in positive (c) and negative (d) modes; (e,f) dot plots of the distributions of relative errors in positive (e) and negative (f) modes; (g,h) bar graphs of cumulative relative errors in positive (g) and negative (h) modes. Authentic CCS value for each metabolite was calculated using mean measurement value from both Agilent and Waters IM-MS systems.

than from the training data set. The CCS values of these metabolites were measured in both the Agilent IM-MS system in our lab and Waters Synapt IM-MS system previously reported by Astarita et al.[15] Please note that these metabolites were not used for prediction method optimization and building, therefore, referred to as an external validation data set. As shown in Figure 3, predicted CCS values from the SVR method matched very well with the experimentally measured CCS values from the Agilent IM-MS system. The R^2 values of regression curves were 0.9579 and 0.9780 in positive and negative modes, respectively. Median relative errors were 1.77% and 1.56% in positive and negative modes, respectively (Figure 3a,b). Please note all the measurements were done in the same instrument and the same lab, therefore, defined as intrainstrument and intralab validation in Figure 3.

Then, we further compared the SVR predicted CCS values with measured CCS values from the Waters Synapt IM-MS system.[15] Since these CCS measurements were carried out in different instruments and laboratories, the comparison was defined as interinstrument and interlab comparison. Again, the results showed good prediction precision that the R² values of regression curves were 0.9652 and 0.9748 in positive and negative modes, respectively. Also median relative errors were 3.11% and 1.47% in positive and negative modes, respectively (Figure 3c,d).

Moreover, we further compared SVR predicted CCS values of 43 compounds (derived from single-field measurements) to the experimental CCS values measured by the multifield method available from the McLean,[24] Bush,[27] and Schmitz[26] groups (Data was provided in the Supporting Information). The median relative error was 2.07%, and the R^2 value of the regression curve was 0.9338 (Figure S6 in the Supporting Information). Taken together, it suggested that the SVR predicted CCS values matched well with experimental

measured CCS values from different instruments and laboratories even with different measurement methods.

To summarize the performance of the SVR prediction method, for intrainstrument/lab comparisons, we discovered that more than 90% of metabolites had predicted CCS values within 5% of relative errors compared to experimentally measured values in both positive and negative ionization modes, while more than 70% of metabolites had predicted CCS values within 3% of relative errors. The performance was similar to internal validation using metabolites in training data sets (Figure 3e−h). We manually analyzed the metabolites with relative errors larger than 5% and found that the presence of multiple isomers for one metabolite may be the reason for inaccurate prediction (Figure S5 in Supporting Information). However, there is no clear explanation why predicted CCS values match well with more extended isomers instead of compact ones. Through comparing against the Waters IM-MS CCS data set, we discovered that more than 90% of metabolites had predicted CCS values within 10% relative error in positive mode, while more than 90% of metabolites had predicted CCS values within 5% relative error in the negative mode (Figure 3e−h). Therefore, these results demonstrated that our SVR prediction method had very good predictive capability and performance and fitted with different instruments and laboratories.

Precision Improvement Compared to Theoretical Calculation. Currently, theoretical calculation is widely employed to obtain CCS values in IM-MS studies. We further compared the performance of SVR prediction method with the theoretical calculation method. In the previous work, Astarita et al. reported 94 and 111 theoretically calculated CCS values for metabolites in positive and negative modes, respectively.[15] These CCS values were obtained from MOBCAL software, denoted as "Theoretical CCS" in Figure 4. Meanwhile, we also

obtained the predicted CCS values for these metabolites using SVR prediction method, denoted as "Predicted CCS" in Figure 4. Then we compared the precision of these CCS values to the authentic ones. In order to avoid differences between different instruments and laboratories, authentic CCS value for each metabolite was calculated using the mean value from both Agilent and Waters IM-MS systems.

The improved precision was observed for "Predicted CCS" values in both positive and negative modes. Specifically, predicted CCS values from SVR method had excellent precision when comparing with authentic CCS values. The results showed that the R^2 values for the regression curves were 0.9641 and 0.9794, and median relative errors (MREs) were 2.51% and 1.53% in positive and negative modes, respectively (Figure 4a,b). By contrast, theoretical CCS values had relatively poor precision when comparing with authentic CCS values. The results showed that the R^2 values for the regression curves were only 0.9165 and 0.9448, and MREs were as large as 4.18% and 5.03% in positive and negative modes, respectively (Figure 4c,d). Interestingly, it was observed that most ions have theoretical CCS values significantly larger than authentic CCS values, especially for relative larger ions with CCS values larger than 150 Å^2. For instance, adenosine monophosphate (AMP) had an authentic CCS value of 169.2 Å^2, which was very similar to the predicted CCS value of 166.8 Å^2 (1.4% of relative error) using SVR prediction method. As a comparison, the theoretically calculated CCS value was as large as 193 Å^2 (14% of relative error). The unreasonably large errors in theoretical calculation may be caused by a larger distribution of conformational complexity, which leaded to inaccuracy of averaged molecular conformations. To summarize the results, we discovered that 82% and 92% of the predicted CCS values had relative errors less than 5% in positive and negative modes, respectively. As a comparison, only 55% and 50% of theoretical CCS values had relative errors within 5% in positive and negative modes, respectively (Figure 4e−h). These results supported precision improvement of SVR prediction method in comparison to theoretical calculation.

Generation of Predicted CCS Database for Metabolomics. Recently, the use of CCS values for metabolite identification in metabolomics has been proven to effectively increase identification confidence.[15,25,26] However, this metabolomics workflow suffers from the limited number of available metabolite CCS values. Here, we demonstrated that using our SVR prediction method, CCS values for thousands of metabolites can be readily predicted within 10 min. Taken HMDB as an example, we successfully predicted 35 203 metabolites with accurate mass between 60 and 1000 Da. In order to reduce false positives caused by the uncommon ionization adducts, we only predicted CCS values for five common adducts such as $[M + H]^+$, $[M + Na]^+$, and $[M − H_2O + H]^+$ in positive mode, and $[M − H]^−$ and $[M + Na − 2H]^−$ in negative mode for each metabolite, accounting for 176 015 CCS values in total. To the best of our knowledge, this is the first available large-scale CCS database for metabolomics study. The predicted metabolite CCS database, namely, MetCCS, can be freely downloaded from our group Web site (http://www.metabolomics-shanghai.org/software.php).

In order to demonstrate the applicability of the predicted CCS database to metabolomics, we analyzed human urine samples using LC-DTIM-QTOF-MS. Using the IMFE function from IM-MS Browser software for feature extraction, a total of 3 618 features were detected. A feature is a collection of peaks

from one metabolite including isotopic peaks. Then these features were searched against to HMDB database (with 35 203 metabolites in total) for metabolite identification. For the purpose of comparison, two match methods, i.e., m/z match only and m/z and CCS match, were used for metabolite identification. Here, we set the $\Delta m/z$ match tolerances as 15 ppm and ΔCCS match tolerances as 3% to balance the filter efficiency and to avoid overfiltering (Figure S7 in the Supporting Information). With the increased ΔCCS match tolerance, filtering efficiency decreased while the overfiltering reduced accordingly. As shown in Figure 5a, for m/z match
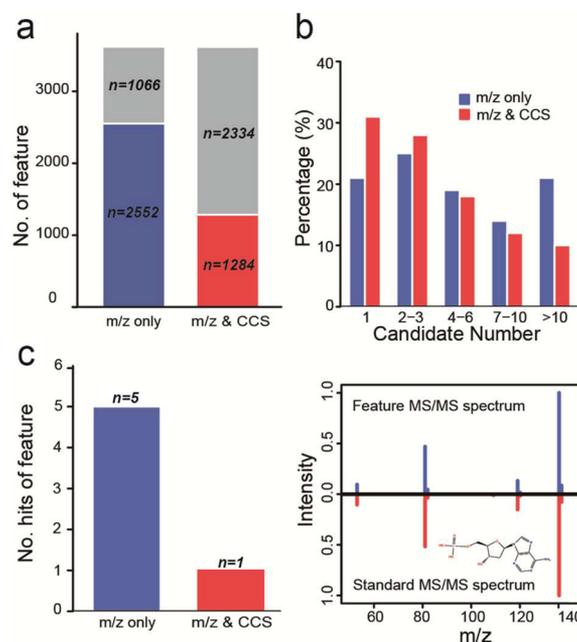

Figure 5. Application of predicted CCS database to metabolite identification in untargeted metabolomics: (a) number of metabolic features with metabolite hits using m/z only or both m/z and CCS match searches; (b) percentage distribution of features with different metabolite hits from two match methods; (c) number of metabolite hits for metabolic feature (M332T13CCS168, left panel) and its MS/MS spectral match with standard deoxyadenosine monophosphate (dAMP, right panel).

only search, 2 552 out of 3 618 features (70%) had at least one metabolite hit. However, for m/z and CCS match, only 1 284 out of 3 618 features (35%) had at least one metabolite hit. Therefore, 50% of features were filtered using the additional ΔCCS match. The results demonstrated that the introduction of CCS values for metabolite identification could significantly reduce false positive identifications.

As shown in Figure 5b, the percentage of features with less than three metabolite hits significantly increased with the additional ΔCCS match. Concurrently, the percentage of features with more than four metabolite hits decreased with the additional ΔCCS match. For other biological samples such as human serum, mammalian cell (Jurkat cells), rat liver tissue, and *Escherichia coli* bacteria sample, similar results can be readily obtained (Figure S8). For example, for the feature M332T13CCS168 (m/z 332.0746 Da; RT 13.08 min; CCS 168 Å^2, Figure 5c), 5 metabolite candidates were obtained in HMDB using m/z only match method. However, 4 candidates were further filtered by the additional CCS match. Then the feature was identified as deoxyadenosine monophosphate

(dAMP, m/z 332.0760 Da; CCS 166 Å^2 in the database). To confirm the identification of this metabolite, we further purchased a chemical standard to acquire the standard MS/MS spectrum and retention time (RT) and validated the identification (Figure 5c and Figure S9). Overall, these results demonstrated that the addition of predicted CCS values for metabolite identification could significantly narrow the search scope and improve the identification accuracy.

CONCLUSION

In conclusion, we developed a SVR algorithm based prediction method that can utilize the molecular descriptors of a metabolite to accurately predict its CCS value. The high precision of the prediction method was externally validated using an independent set of metabolites with measured CCS values from different IM-MS instruments and different laboratories. The results proved that the SVR based prediction method had a high prediction precision for CCS values with median relative errors of ~3%. Then, this prediction method was used to generate a large-scale predicted metabolite CCS database, namely, MetCCS, which contains 35 203 metabolites in total. The whole MetCCS database can be freely download on the Internet. In this work, since CCS values of metabolites in the training data set were all acquired in nitrogen buffer gas, predicted CCS values are all nitrogen CCS values, denoted as Ω_{N2}. Presumably, this method can also be applied to predict helium CCS values (Ω_{He}) if we use helium CCS values of metabolites as the training data set. Finally, applying this predicted CCS database to untargeted metabolomics significantly reduced false positive identifications of metabolites by improving identification accuracy. Therefore, we believe that the MetCCS database will have a broad application in untargeted metabolomics.

ASSOCIATED CONTENT

© Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.6b03091.

> Tables S1−S6, Schemes S1 and S2, Figures S1−S9, and supplemental experimental details (PDF)
> Excel files for experimental and theoretical CCS values (ZIP)

AUTHOR INFORMATION

Corresponding Author

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

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

REFERENCES

(1) May, J. C.; McLean, J. A. *Anal. Chem.* **2015**, *87*, 1422−1436.
(2) Lanucara, F.; Holman, S. W.; Gray, C. J.; Eyers, C. E. *Nat. Chem.* **2014**, *6*, 281−294.
(3) Bush, M. F.; Hall, Z.; Giles, K.; Hoyes, J.; Robinson, C. V.; Ruotolo, B. T. *Anal. Chem.* **2010**, *82*, 9557−9565.
(4) Lapthorn, C.; Pullen, F.; Chowdhry, B. Z. *Mass Spectrom. Rev.* **2013**, *32*, 43−71.
(5) Cajka, T.; Fiehn, O. *Anal. Chem.* **2016**, *88*, 524−545.
(6) Jia, C.; Wu, Z.; Lietz, C. B.; Liang, Z.; Cui, Q.; Li, L. *Anal. Chem.* **2014**, *86*, 2917−2924.
(7) Lietz, C. B.; Yu, Q.; Li, L. *J. Am. Soc. Mass Spectrom.* **2014**, *25*, 2009−2019.
(8) Baker, E. S.; Burnum-Johnson, K. E.; Ibrahim, Y. M.; Orton, D. J.; Monroe, M. E.; Kelly, R. T.; Moore, R. J.; Zhang, X.; Theberge, R.; Costello, C. E.; Smith, R. D. *Proteomics* **2015**, *15*, 2766−2776.
(9) 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.
(10) Paglia, G.; Kliman, M.; Claude, E.; Geromanos, S.; Astarita, G. *Anal. Bioanal. Chem.* **2015**, *407*, 4995−5007.
(11) 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.
(12) Dwivedi, P.; Wu, P.; Klopsch, S. J.; Puzon, G. J.; Xun, L.; Hill, H. H., Jr *Metabolomics* **2008**, *4*, 63−80.
(13) Dwivedi, P.; Schultz, A. J.; Hill, H. H. *Int. J. Mass Spectrom.* **2010**, *298*, 78−90.
(14) Hines, K. M.; Ashfaq, S.; Davidson, J. M.; Opalenik, S. R.; Wikswo, J. P.; McLean, J. A. *Anal. Chem.* **2013**, *85*, 3651−3659.
(15) Paglia, G.; Williams, J. P.; Menikarachchi, L.; Thompson, J. W.; Tyldesley-Worster, R.; Halldorsson, S.; Rolfsson, O.; Moseley, A.; Grant, D.; Langridge, J.; Palsson, B. O.; Astarita, G. *Anal. Chem.* **2014**, *86*, 3985−3993.
(16) Chouinard, C. D.; Wei, M. S.; Beekman, C. R.; Kemperman, R. H.; Yost, R. A. *Clin. Chem.* **2016**, *62*, 124−133.
(17) Patti, G. J.; Yanes, O.; Siuzdak, G. *Nat. Rev. Mol. Cell Biol.* **2012**, *13*, 263−269.
(18) Wishart, D. S. *Nat. Rev. Drug Discovery* **2016**, *15*, 473−484.
(19) Zhu, Z. J.; Schultz, A. W.; Wang, J.; Johnson, C. H.; Yannone, S. M.; Patti, G. J.; Siuzdak, G. *Nat. Protoc.* **2013**, *8*, 451−460.
(20) Tautenhahn, R.; Cho, K.; Uritboonthai, W.; Zhu, Z.; Patti, G. J.; Siuzdak, G. *Nat. Biotechnol.* **2012**, *30*, 826−828.
(21) Smith, C. A.; O'Maille, G.; Want, E. J.; Qin, C.; Trauger, S. A.; Brandon, T. R.; Custodio, D. E.; Abagyan, R.; Siuzdak, G. *Ther. Drug Monit.* **2005**, *27*, 747−751.
(22) Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; Oda, Y.; Kakazu, Y.; Kusano, M.; Tohge, T.; Matsuda, F.; Sawada, Y.; Hirai, M. Y.; Nakanishi, H.; Ikeda, K.; Akimoto, N.; Maoka, T.; Takahashi, H.; Ara, T.; Sakurai, N.; Suzuki, H.; Shibata, D.; Neumann, S.; Iida, T.; Tanaka, K.; Funatsu, K.; Matsuura, F.; Soga, T.; Taguchi, R.; Saito, K.; Nishioka, T. *J. Mass Spectrom.* **2010**, *45*, 703−714.
(23) *METLIN*, https://metlin.scripps.edu/index, accessed on August 8, 2016.
(24) 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.
(25) Stephan, S.; Jakob, C.; Hippler, J.; Schmitz, O. J. *Anal. Bioanal. Chem.* **2016**, *408*, 3751−3759.
(26) Stephan, S.; Hippler, J.; Kohler, T.; Deeb, A. A.; Schmidt, T. C.; Schmitz, O. J. *Anal. Bioanal. Chem.* **2016**, *408*, 6545−6555.
(27) Campuzano, I.; Bush, M. F.; Robinson, C. V.; Beaumont, C.; Richardson, K.; Kim, H.; Kim, H. I. *Anal. Chem.* **2012**, *84*, 1026−1033.
(28) Lapthorn, C.; Pullen, F. S.; Chowdhry, B. Z.; Wright, P.; Perkins, G. L.; Heredia, Y. *Analyst* **2015**, *140*, 6814−6823.

(29) Boschmans, J.; Jacobs, S.; Williams, J. P.; Palmer, M.; Richardson, K.; Giles, K.; Lapthorn, C.; Herrebout, W. A.; Lemiere, F.; Sobott, F. *Analyst* **2016**, *141*, 4044−4054.

(30) Mesleh, M.; Hunter, J.; Shvartsburg, A.; Schatz, G. C.; Jarrold, M. *J. Phys. Chem.* **1996**, *100*, 16082−16086.

(31) Shvartsburg, A. A.; Jarrold, M. F. *Chem. Phys. Lett.* **1996**, *261*, 86−91.

(32) Lapthorn, C.; Dines, T. J.; Chowdhry, B. Z.; Perkins, G. L.; Pullen, F. S. *Rapid Commun. Mass Spectrom.* **2013**, *27*, 2399−2410.

(33) Poyer, S.; Loutelier-Bourhis, C.; Coadou, G.; Mondeguer, F.; Enche, J.; Bossee, A.; Hess, P.; Afonso, C. *J. Mass Spectrom.* **2015**, *50*, 175−181.

(34) Shah, A. R.; Agarwal, K.; Baker, E. S.; Singhal, M.; Mayampurath, A. M.; Ibrahim, Y. M.; Kangas, L. J.; Monroe, M. E.; Zhao, R.; Belov, M. E.; Anderson, G. A.; Smith, R. D. *Bioinformatics* **2010**, *26*, 1601−1607.

(35) Wang, B.; Valentine, S.; Plasencia, M.; Raghuraman, S.; Zhang, X. *BMC Bioinf.* **2010**, *11*, 182.

(36) Wishart, D. S.; Jewison, T.; Guo, A. C.; Wilson, M.; Knox, C.; Liu, Y.; Djoumbou, Y.; Mandal, R.; Aziat, F.; Dong, E.; Bouatra, S.; Sinelnikov, I.; Arndt, D.; Xia, J.; Liu, P.; Yallou, F.; Bjorndahl, T.; Perez-Pineiro, R.; Eisner, R.; Allen, F.; Neveu, V.; Greiner, R.; Scalbert, A. *Nucleic Acids Res.* **2013**, *41*, D801−807.

(37) Cai, Y.; Weng, K.; Guo, Y.; Peng, J.; Zhu, Z.-J. *Metabolomics* **2015**, *11*, 1575−1586.

(38) Kurulugama, R. T.; Darland, E.; Kuhlmann, F.; Stafford, G.; Fjeldsted, J. *Analyst* **2015**, *140*, 6834−6844.

(39) Mordehai, A.; Kurulugama, R.; Darland, E.; Stafford, G.; Fjeldsted, J. *Proceedings of the 63rd ASMS Conference on Mass Spectrometry and Allied Topics*, St. Louis, MO, May 31−June 4, 2015; MP 126.

(40) Steinwart, I.; Christmann, A. *Support Vector Machines*; Springer Science & Business Media: New York, 2008.

(41) Ren, S.; Hinzman, A. A.; Kang, E. L.; Szczesniak, R. D.; Lu, L. J. *Metabolomics* **2015**, *11*, 1492−1513.

(42) Shen, X.; Gong, X.; Cai, Y.; Guo, Y.; Tu, J.; Li, H.; Zhang, T.; Wang, J.; Xue, F.; Zhu, Z.-J. *Metabolomics* **2016**, *12*, 89.

(43) Brereton, R. G.; Lloyd, G. R. *Analyst* **2010**, *135*, 230−267.