Using differential mobility spectrometry and machine learning-based modeling to predict the physicochemical properties of molecules

J. Larry Campbell[1] and W. Scott Hopkins[2]

[1]SCIEX, Concord, ON, Canada; [2] Department of Chemistry, University of Waterloo, Waterloo, ON, Canada

INTRODUCTION

Over the past several years, a number of studies have shown how various structural features of compounds influence their observed behavior during differential mobility spectrometry (DMS) experiments. For example, DMS data have revealed the specific influences of a molecule's site of charging (Campbell et al., 2012; Kovačević et al., 2014; Walker et al., 2018), the steric hinderance proximal to the charge site (Campbell et al., 2014; Liu et al., 2015), or the resonance stabilization/delocalization of an ion's charge (Liu et al., 2017). As an overall result, the applicability and understanding of processes within the DMS have increased substantially.

Beyond, these structural relationships, we have also identified that the observed DMS behaviors of ions (i.e., the relationship between the optimal separation voltage (SV) and compensation voltage (CV) needed for transmission) also encode for the physicochemical properties of those molecules. For example, we observed how DMS behaviors correlated strongly with the measured pKa and pKb values, as well as the passive cell permeabilities for a set of structurally related drug molecules, even discriminating between isomeric forms of these drugs.(Liu et al., 2017) However, these initial correlations made use of a select key DMS-related metric – the $SV@CV_{min}$ – where the ions' dispersion plots (SV vs. CV) displayed a minimum value. Undoubtedly, more knowledge could be garnered from these rich data sets if a more global view of the DMS data were available.

To provide a more comprehensive, precise, and accurate method that links DMS behavior to physicochemical properties, we have adopted supervised machine learning to treat the DMS data from over 250 molecules having varied chemical structures. Here, we demonstrate that indeed the gas-phase clustering behavior of an ion in a DMS cell can be used to predict a number of physicochemical properties, including collision cross sections (CCSs), as well as condensed phase molecular properties like cell permeability, chemical reactivity, solubility, polar surface area, and water/octanol distribution coefficient.

MATERIALS AND METHODS

Sample Preparation. More than 250 compounds were analyzed during the course of this study. Each compound was present at 100 ng/mL and subjected to ESI(+) prior to DMS analysis. These compounds included species that had been the subject of previous studies, including quinoline-based drugs (Liu et al., 2015), quinoline-8-ol-based drugs (Liu et al., 2017), tetraalkylammonium ions (Campbell et al., 2014), drugs designed to incorporate intramolecular hydrogen bonds (IMHBs) (Goetz et al., 2014), electrophilic chemically reactive groups (CRGs) used in drug design (Flanagan et al., 2014), and a test mixture of ~180 compounds (Schneider et al., 2015) used to evaluate DMS and MS performance.

DMS-MS Conditions. A differential mobility spectrometer (Figure 1) was mounted in the atmospheric region between the sampling orifice and ESI source (5500V) of a hybrid triple quadrupole – linear ion trap mass spectrometer (Figure 1). The fundamentals of the DMS device have been described elsewhere. (Schneider et al., 2010) The temperature of the DMS cell was maintained at a selected temperature (150, 225, or 300 ºC) during the course of an experiment, and the nitrogen curtain gas was operated at 10 psi. In this study, the separation voltage (SV) was held at a constant value (3250, 3500, 3750, or 4000 V) while the compensation voltage (CV) was scanned from -40 V to +20 V in 0.10-V increments.

Data analysis and Machine Learning (ML) Modeling. All data were analyzed using a research version of PeakView® Software (SCIEX) and the DMS ionogram data (SV versus CV versus Intensity) were output to Orange Canvas (v. 3.4.2) – a Python-based machine learning interface. These data were treated with five different ML algorithms: [1] k Nearest Neighbors (kNN), [2] Random Forest, [3] Decision Tree, [4] Linear Regression, and [5] Adaptive Boosting (AdaBoost) to evaluate the quality of the predictive models; ultimately, random forest regression was selected based upon this appraisal. This supervised ML uses multiple decision trees and statistically analyzes outcomes to generate a predictive model. The data (DMS and associated meta data, including m/z, MOBCAL-modeled (Mesleh et al., 1996; Shvartsburg and Jarrold, 1996) CCS values, experimentally determined cell permeability, pka, ion/solvent binding energies, etc.) were randomly split differently for each tree, of which there were 10 trees. The data were randomly binned into 5 folds and the algorithm was run a total of 5 times (matching the number of folds), each time leaving 1 fold out of the training set for cross validation. This allows us to infer relationships in the labeled data sets.

RESULTS

Building upon earlier studies relating DMS behavior with ion structure

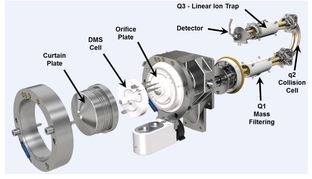


Figure 1. Exploded view of the DMS system employed in this study, including the hybrid quadrupole linear ion trap mass spectrometer.

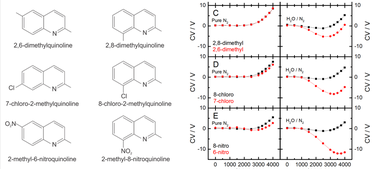


Figure 2. The effect of steric hinderance of a modifier molecules (e.g., water) to a charge site (quinoline ring N) is demonstrated here. When a substituent is located in the 8-position (right-hand column, black traces), the CV shifts are less negative, indicating weaker ion/solvent binding (Liu et al., 2015).

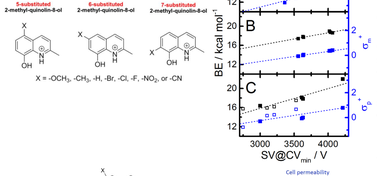


Figure 3. For the isomeric compound sets (2-methyl-quinoline-8-ols, left-hand side) analyzed by DMS, the positive charge originating from the protonated ring nitrogen can be delocalized efficiently by electron donating groups in the 5- and 7-postions, but not the 6-position. As a consequence, the 5- and 7-substituted species were found to bind more weakly than their 6-substituted isomers. (Liu et al., 2017). The strong correlation between the substituents' respective SV@Cvmin values with σ^+ parameters (Brown and Okamoto, 1958) and the calculated ion/modifier binding energies (right hand plots) support these findings.

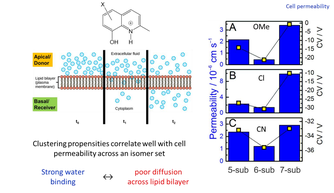


Figure 4. Relating the DMS behaviors of isomeric drug molecules to their relative passive cell permeabilities. With more charge delocalization from the 5- and 7-substituted isomers, greater passive cell permeabilities were observed for these species compared to the 6-substituted isomers for a given substituent (Liu et al., 2017). It is postulated that the greater permeability derives from the weaker binding of those isomers to water, which must be shed prior to passage through a lipid bilayer.

RESULTS

Developing a machine learning-based model using DMS and other meta data

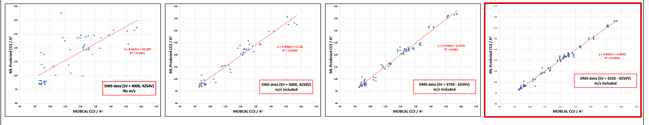


Figure 5. The effect of including various amounts of DMS-based data into the ML model is highlighted in the plots above. If only a limited number of DMS (SV, CV) data points are used to build and train the ML model (far left plot), a model of weak predictive power results (low R^2 value). However, as more and more of the DMS dispersion plot data are included in the ML modeling (plots advancing to the right), the correlation between the ML model and the MOBCAL-predicted CCS values greatly increases.

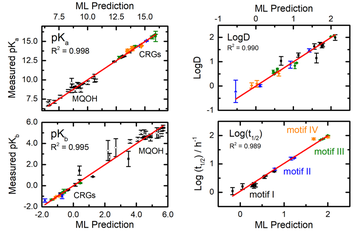


Figure 6. Other physicochemical parameters were used to train predictive models for compounds examined using the DMS data. For example, the 2-methylquinoline (and quinolin-8-ols), as well as the acrylamide CRGs, were used to build and train models to predict pka, pkb, LogD, and Log($t_{1/2}$) values (experimental values obtained by Pfizer, Groton, CT). The quality of the models will continue to improve with the inclusion of additional compounds to these data sets.

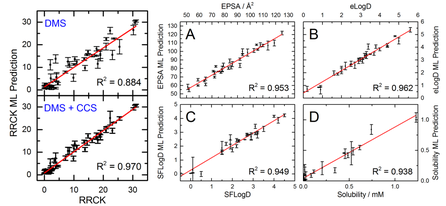
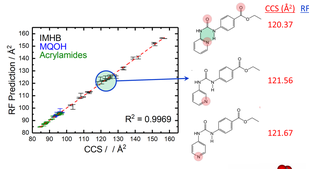


Figure 7. Improving the quality of the ML modeling by selection of additional meta data. As shown in the plots above, when an ML model was developed to predict passive cell permeability (RRCK value) using the test compounds' DMS data (upper left-most plot), a model of moderately strong correlation was the result. However, as is the case with ML methods, the inclusion of additional qualifying labels/meta data can improve the quality of the model's prediction. Here, the inclusion of the MOBCAL-predicted CCS values for each of the test compounds further improved the accuracy of the ML models for RRCK, as well as for (A) EPSA – polar surface areas, (B) eLogD – experimental octanol/water coefficient, (C) Log D, and (D) compound solubility.

RESULTS

Interesting features among the ML modeled physicochemical properties



CCS (Å²)	RF Prediction (Å²)
120.37	120.92(143)
121.56	121.61(02)
121.67	121.66(06)

Figure 8. Plot comparing the MOBCAL-predicted CCS values for three classes of compounds (IMHB-designed drugs, 2-methylquinolines (and quinoline-8-ols), and acrylamide CRGs with the ML-modeled CCS values that include DMS data. Highlighted are three IMHB isomers, whose DMS CV shifts provide clear separation (data not shown) as well as the correct rank ordering of CCS values.

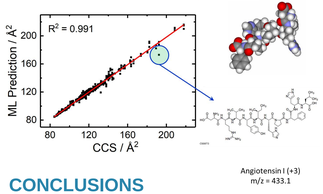


Figure 9. Plot comparing the MOBCAL-predicted CCS values for all 255 compounds examined thus far with the ML-modeled CCS values that include DMS data. While strong correlation is observed between the MOBCAL- and ML-based CCS values, outliers remain. In this highlighted case, the outlier is the triply protonated form of the peptide, Angiotensin I. Given that the vast majority of the other test compounds were singly charged species, non-peptide species, this outlier makes sense.

CONCLUSIONS

By using DMS analyses and ML modeling, we are able to obtain good predictions for the physicochemical properties of a large number of compounds. This work is on-going and will expand to a much larger cohort of compounds, including biological species, and will explore the application of more advanced ML algorithms.

REFERENCES

Brown, H. C.; Okamoto, Y. J. *Am. Chem. Soc.* **1958**, *80*, 4979−4987.

Campbell, J.L.; Zhu, M.; Hopkins, W.S. *J. Am. Soc. Mass Spectrom.* **2014**, *25*, 1583-1591.

Flanagan, M.E.; Abramite, J.A.; Anderson, D.P.; Aulabaugh, A.; Dahal U.P; Gilbert, A.M.; Li, C.; Montgomery, J.; Oppenheimer, S.R.; Ryder, T.; Schuff, B.P.; Uccello, D.P.; Walker, G.S.; Wu, Y.; Brown, M.F.; Chen, J.M.; Hayward, M.M.; Noe, M.C.; Obach, R.S.; Philippe, L.; Shanmugasundaram, V.; Shapiro, M.J.; Starr, J.; Stroh, J.; Che, Y. *J. Med. Chem.* **2014**, *57*, 10072−10079.

Goetz, G.H.; Farrell, W.; Shalaeva, M.; Sciabola, S.; Anderson, D.; Yan, J.; Philippe, L.; Shapiro, M.J. *J. Med. Chem.* **2014**, *57*, 2920−2929.

Kovačević, B.; Schorr, P.; Qi, Y.; Volmer, D.A. *J. Am. Soc. Mass Spectrom.* **2014**, *25*, 1974-1986.

Liu, C.; Le Blanc, J.C.Y.; Shields, J.; Janiszewski, J.S.; Ieritano, C.; Ye, G.F.; Hawes, G.F.; Hopkins, W.S.; Campbell, J.L. *Analyst* **2015**, *140*, 6897-6903.

Liu, C.; Le Blanc, J.C.Y.; Schneider, B.B.; Shields, J.; Federico, J.J.; Zhang, H.; Stroh, J.G.; Kauffman, G.W.; Kung, D.W.; Shapiro, M.; Ieritano, C.; Shepherson, E.; Verbuyst, M.; Melo, L.; Hasan, M.; Naser, D.; Janiszewski, J.S.; Hopkins, W.S.; Campbell, J.L. *ACS Cent. Sci.* **2017**, *3*, 101-109.

Mesleh, M.F.; Hunter, J.M.; Shvartsburg, A.A.; Schatz, G.C.; Jarrold, M.F. *J. Phys. Chem.* **1996**, *100*, 16082-16086.

Schneider, B. B.; Covey, T. R.; Coy, S. L.; Krylov, E. V.; Nazarov, E. G. *Int. J. Mass Spectrom.*, **2010**, *298*, 45-54.

Schneider, B. B.; Nazarov, E.G.; Londry, F.; Covey, T. R. *Int. J. Ion Mobil. Spec.* **2015**, *18*, 159–170.

Shvartsburg, A.A.; Jarrold, M.F. *Chem. Phys. Lett.* **1996**, *261*, 86-91.

Walker, S.W.C.; Mark, A.; Verbuyst, B.; Bogdanov, B.; Campbell, J.L.; Hopkins, W.S. *J. Phys. Chem. A.* **2018**, *122*, 3858-3865.

TRADEMARKS/LICENSING

AB Sciex is doing business as SCIEX.

Document number: [RUO-MKT-10-7830-A]