



RESEARCH ARTICLE

Artificial intelligence for ion mobility spectrometry and mass spectrometry in omics research

Insights into modifiers effects in differential mobility spectrometry: A data science approach for metabolomics and peptidomics

Stepan Stepanovic^{1,2}  | Lysi Ekmekciu¹ | Bandar Alghanem^{1,3}  |
G rard Hopfgartner¹ 

¹Life Sciences Mass Spectrometry, Department of Inorganic and Analytical Chemistry, University of Geneva, Geneva, Switzerland

²Institute of Chemistry, Technology and Metallurgy, University of Belgrade, Belgrade, Serbia

³Medical Research Core Facility and Platforms (MRCFP), King Abdullah International Medical Research Center, Riyadh, Saudi Arabia

Correspondence

S. Stepanovic and G. Hopfgartner, Life Sciences Mass Spectrometry, Department of Inorganic and Analytical Chemistry, University of Geneva, 24 Quai Ernest Ansermet, Geneva 4 CH-1211, Switzerland.

Email: stepan.stepanovic@unige.ch; gerard.hopfgartner@unige.ch

Funding information

Swiss National Science Foundation, Grant/Award Numbers: CRSII3_136282, 200021_192306

Abstract

Utilizing a data-driven approach, this study investigates modifier effects on compensation voltage in differential mobility spectrometry–mass spectrometry (DMS-MS) for metabolites and peptides. Our analysis uncovers specific factors causing signal suppression in small molecules and pinpoints both signal suppression mechanisms and the analytes involved. In peptides, machine learning models discern a relationship between molecular weight, topological polar surface area, peptide charge, and proton transfer-induced signal suppression. The models exhibit robust performance, offering valuable insights for the application of DMS to metabolites and tryptic peptides analysis by DMS-MS.

KEYWORDS

data analysis, differential mobility spectrometry, machine learning, metabolites, peptides

1 | INTRODUCTION

The integration of liquid chromatography with mass spectrometry (LC–MS) and electrospray ionization (ESI) has facilitated the identification and quantification of a large array of compounds of biological interest. LC–MS analysis is essential in omics research, aiming to comprehensively study biological molecules, such as genes, peptides/proteins, lipids, and metabolites, using high-throughput techniques to decipher complex interactions within biological systems.¹ However, the multitude of identified exogenous and endogenous compounds challenges LC–MS platforms, limited by LC's peak capacity and the difficulty in distinguishing isobaric or isomeric analytes using MS or

even MS/MS.² A popular solution is adding an additional, highly orthogonal separation dimension, either through an additional LC column with a different separation mechanism (e.g., ion exchange chromatography and Hydrophilic Interaction Liquid Chromatography) or coupling with techniques like gas chromatography. Yet these methods have their drawbacks: LCxLC^{3,4} requires complex, time-consuming parameter optimization; LCxGC⁵ demands intricate sample preparation and interface management, particularly for non-volatile analytes. Ion mobility spectrometry (IMS),^{6,7} which separates ions in the gas phase based on their mobility under an electric field, emerges as a promising alternative. Relevant IMS variations for a secondary separation dimension include Drift Tube IMS,⁸ utilizing linear mobility

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

  2024 The Authors. *Journal of Mass Spectrometry* published by John Wiley & Sons Ltd.

dependence in a constant electric field; Field Asymmetric Ion mobility Spectrometry (FAIMS),^{9–11} differentiating ions with contrasting non-linear high-field and linear low-field mobilities in alternating fields; Traveling Wave IMS,^{12,13} propelling ions through a dynamically changing electric field, focusing on induced mobility variations; and Trapped Ion Mobility Spectrometry (TIMS),^{14,15} which balances the electric field force against a gas flow (ions trapping) and then gradually decreasing the electric field (separation). All these methods enable ion resolution in space and time enabling a wide range of analytes to be separated without prior knowledge of their properties, which is crucial for untargeted profiling. It is also important to note that because of constant low field used, DTIMS allows for the direct collision cross section (CCS) measurement but TWIMS and TIMS also enable the derivation of CCS values under their operational parameters.

On the other side of ion mobility spectrum, FAIMS^{9–11} differentiates the ions with contrasting non-linear high-field and linear low-field mobilities in alternating fields. Differential mobility spectrometry (DMS),¹⁶ similar to FAIMS, applies an asymmetric electric field (separation voltage, SV) between two planar electrodes, perpendicular to ion movement. While high- and low-field waveform components have equivalent but opposite areas, differential mobility arises as ion mobility exits the linear mode in high-field conditions. Since both FAIMS and DMS introduce non-linear ion mobility, they are not so suitable for accurate CSS measurements. To mitigate differential mobility effects on ion trajectory, a compensating DC voltage (CoV) is applied to select the ion of interest.¹⁷ The CoV acts as a critical control parameter, enhancing the selectivity and specificity of the ion filtering process, which is key for its application in targeted analyses. Additionally, the DMS environment, typically using N₂ carrier gas, can be enhanced with solvent vapors like polar protic (water and alcohols), polar-aprotic (acetonitrile and acetone), or non-polar but highly polarizable (toluene) modifiers, targeting specific analyte interactions. These modifiers can be employed to fine-tune selectivity, especially valuable for separating co-eluting isomeric compounds.¹⁷

In exploring the intricacies of IMS, particularly DMS, the focus shifts to the microscopic interactions between analytes and modifiers. This transition leads us to delve into the realm of in-silico tools, crucial for gaining a deeper understanding of these interactions and correlating them with other molecular properties. The study by Ruskic and Hopfgartner¹⁸ reveals how DMS selectivity for isomeric analytes is influenced by factors like reduced mass and cluster binding energy, with findings supported by density functional theory (DFT) calculations and molecular modeling. The paper by Walker et al.¹⁹ demonstrates the use of DMS and machine learning (ML) to rapidly predict key molecular properties of drug candidates, enhancing drug discovery efficiency. Ieritano et al.²⁰ developed a ML model using Random Forest Regression to predict dispersion curves in DMS, achieving a MAE of ≤ 2.4 V, which further improved to ≤ 1.2 V with guided training. This approach significantly enhances the efficiency of DMS method development by accurately predicting ion behavior with minimal input data. A very recent paper by Stienstra et al.²¹ demonstrates the effective use of DMS and ML to predict water solubility (log S) and water-octanol partition coefficient (log P) generated by the

OPERA package, highlighting the significance of integrating both experimental DMS data and structural descriptors for improved accuracy. Bissonnette et al.²² used first-principles kinetics-based model, together with MobCal-MPI and DFT/DLPNO-CCSD(T) to study binary solvent mixtures in DMS, with conclusion that the differential mobility of ions is predominantly influenced by the solvent binding energies with a secondary contribution from solvent size. Chakraborty et al.²³ implemented convolutional neural networks and signal processing techniques like magnitude-squared coherence in DMS data analysis, achieving high accuracy in identifying pure chemicals and their mixtures, offering an efficient approach for chemical identification in various applications.

In this project, we aim to encompass a multifaceted approach to molecular analysis using advanced computational tools and data processing techniques. The general goal is to demonstrate how we can uniformly treat very different sets of data and provide an integrated omics treatment that encapsulates singly charged metabolites and multiply charged peptides. Besides gaining insight into the relationship between various important physicochemical properties, the aim was to demonstrate how we can apply data science tools even when we have very little ($n = 25$) data points and also completely diverse sets of molecular systems (metabolites and peptides) and, finally, gain clear insight into microscopic mechanisms behind analyte modifier interaction. The fully automated pipeline was created, which combines python libraries for experimental data processing and relating it with molecular structure and molecular modeling.

2 | MATERIALS AND METHODS

2.1 | Analysis of metabolites mix

A mix of 50 analytes, representative for urine and plasma metabolites (Figure S1 and Table S1), was analyzed by LC-MS on a quadrupole time-of-flight (QTOF) mass spectrometer (TTOF 6600+, Sciex, Concord, ON, Canada) and was equipped with a differential ion mobility device (SelexION, Sciex). All details about chemicals, sample preparation, LC, DMS and MS experiments can be found in published work.²⁴

2.2 | Analysis of peptides mix

One-hundred eighty-five peptides (Table S2) representing 92 proteins were selected for this study. Proteotypic peptides were synthesized under unpurified conditions to produce 20 nmol of each peptide species (JPT Peptide Technologies, Berlin, Germany).

A DMS cell (SCIEX SelexION device) was installed between the ion source and the orifice of a 5500 QTRAP. The CoV was optimized by infusing all the peptides individually by flow injection analysis (FIA) on MRM mode (three transitions for each peptide). The flow rate was set as 15 μ l/min using a HTS-PAL autosampler (CTC Analytics, Zwingen, Switzerland) and a micro-LC pump (LC-10ADVP μ , Shimadzu, Kyoto, Japan). The injection volume was 100 μ l. A SV of 3500 V was used, and the CoV was ramped from -50 to 50 V (steps of 0.2 V)

TABLE 1 Annotation of mix 50 analytes with the respective retention time, CoV values, and selected molecular properties

| N2 | Ch | IPA | Ch_IPA | EtOH | ToI | ACN | Name | Gb | MW | logP | RT | logp | TPSA | Ar |
|-----|-----|-----|--------|------|-----|-----|----------------------------|--------|-----|-------|-------|------|------|----|
| 4 | 3 | | -31 | | | -39 | L-lysine | -34.29 | 146 | -0.47 | 1.09 | -3 | 89 | 0 |
| 1 | -1 | | -38 | | | -46 | L-Histidine | -31.89 | 155 | -0.64 | 1.12 | -3.2 | 92 | 1 |
| 7 | 6 | | -16 | -41 | -20 | -29 | Carnosine | -36.07 | 226 | -1.13 | 1.13 | -4 | 121 | 1 |
| 5 | 3 | | -2 | -42 | -7 | | 1-methylhistidine | -28.59 | 169 | -0.63 | 1.14 | -3.3 | 81 | 1 |
| 9 | 8 | -21 | 6 | -7 | 3 | -31 | Glycerophosphocholine | -27.83 | 257 | -1.45 | 1.18 | -2.3 | 99 | 0 |
| 7 | 5 | | -9 | -40 | -15 | -37 | Homo-L-arginine | -41.58 | 188 | -1.16 | 1.19 | -3.7 | 128 | 0 |
| -1 | -3 | | 4 | | 8 | | L-Glutamine | -30.95 | 146 | -1.34 | 1.19 | -3.1 | 106 | 0 |
| 8 | 6 | -39 | 3 | -19 | -4 | | L-carnitine | -32.25 | 161 | -1.81 | 1.21 | -0.2 | 60 | 0 |
| 7 | 6 | -49 | -11 | -39 | -5 | -25 | N-acetylneuraminic_acid | -20.45 | 309 | -3.87 | 1.25 | -3.5 | 177 | 0 |
| -10 | 7 | | 5 | | 3 | | Creatinine | -25.85 | 113 | -1.23 | 1.25 | -1.8 | 59 | 0 |
| 2 | -1 | -10 | -21 | -2 | -22 | | Trigonelline | -26.54 | 137 | -1.13 | 1.26 | 1.2 | 44 | 1 |
| -1 | -3 | | 4 | | | | Creatine | -39.35 | 131 | -1.10 | 1.29 | -1.2 | 90 | 0 |
| -10 | -11 | | | | | | L-Proline | -30.52 | 115 | -0.18 | 1.3 | -2.5 | 49 | 0 |
| 4 | 2 | -12 | -31 | -2 | -42 | -43 | Homocitrulline | -29.27 | 189 | -0.76 | 1.34 | -3.9 | 118 | 0 |
| 3 | 1 | -50 | -37 | | | -46 | N-acetylputrescine | -33.2 | 130 | -0.14 | 1.34 | -0.7 | 55 | 0 |
| 10 | 8 | -22 | 6 | -8 | 3 | -39 | L-acetylcarnitine | -29.83 | 203 | -1.24 | 1.4 | 0.4 | 66 | 0 |
| 3 | 1 | | 10 | | | -47 | 4-guanidinobutanoic_acid | -42.13 | 145 | -0.88 | 1.69 | -1.5 | 102 | 0 |
| 2 | 1 | -41 | 0 | -41 | -7 | -30 | 3-methyladenine | -40.56 | 149 | -0.22 | 1.76 | -0.2 | 70 | 2 |
| -2 | -2 | -44 | 4 | -50 | | -19 | Urocanic_acid | -25.89 | 138 | 0.51 | 1.85 | 0 | 66 | 1 |
| 1 | -1 | -32 | -10 | -25 | 1 | | 7-methylguanine | -26.09 | 165 | -0.76 | 2.76 | -1.1 | 90 | 2 |
| -6 | -7 | -41 | | | | | Niacinamide | -25.17 | 122 | 0.18 | 2.84 | -0.4 | 56 | 1 |
| -4 | -5 | -42 | -34 | -42 | | -27 | Tyramine | -31.24 | 137 | 0.89 | 3.01 | 1.1 | 46 | 1 |
| 0 | -2 | -23 | -19 | -22 | -36 | -22 | Cotinine | -26.32 | 176 | 1.37 | 4.1 | -0.3 | 33 | 1 |
| 5 | 4 | -32 | -22 | -30 | -33 | -21 | Guanosine | -27.3 | 283 | -2.69 | 4.59 | -1.9 | 160 | 2 |
| -2 | -4 | -36 | -36 | -35 | | -38 | N-methylnicotinamide | -25.48 | 136 | 0.44 | 4.71 | 0 | 42 | 1 |
| 5 | 5 | -36 | -9 | -28 | -13 | -21 | Cyclic_AMP | -26.18 | 329 | -0.82 | 5.12 | -2.6 | 155 | 2 |
| 5 | 5 | -13 | -18 | -14 | -19 | -16 | Ethenodeoxyadenosine | -29.14 | 275 | -0.28 | 5.53 | 0.7 | 98 | 3 |
| 1 | 0 | -36 | -21 | -37 | -27 | -24 | 3-chlorotyrosine | -24.66 | 216 | 1.00 | 5.57 | -1.8 | 84 | 1 |
| 7 | 6 | | -17 | -43 | -11 | -34 | Pantothenic_acid_ | -12.98 | 219 | -1.04 | 6.82 | -1.1 | 107 | 0 |
| 0 | -2 | | -36 | | | -29 | Acetaminophen | -11.12 | 151 | 1.35 | 6.85 | 0.5 | 49 | 1 |
| 4 | 2 | | -30 | | -13 | -33 | Theobromine | -19.73 | 180 | -1.04 | 7.16 | -0.8 | 73 | 2 |
| 6 | 6 | -34 | 2 | -18 | -2 | -31 | 1-methyladenosine | -36.82 | 281 | -2.14 | 7.48 | -1 | 129 | 2 |
| 1 | -1 | | -23 | | | -30 | Isovalerylglycine | -12.52 | 159 | 0.23 | 8.18 | 1.5 | 66 | 0 |
| 7 | 6 | -40 | -15 | -34 | -32 | -20 | L-Aspartyl-L-phenylalanine | -25.08 | 280 | -0.40 | 8.33 | -3.3 | 130 | 1 |
| 2 | 0 | | -26 | | -37 | -30 | Hippuric_acid | -11.93 | 179 | 0.50 | 9.06 | 0.3 | 66 | 1 |
| 3 | 1 | | | | | | 1,3,7-trimethyluric_acid | -11.74 | 210 | -1.74 | 9.16 | -0.4 | 82 | 2 |
| 7 | 6 | | -11 | | 5 | | Chlorogenic_acid | -17.91 | 354 | -0.65 | 9.23 | -0.4 | 165 | 1 |
| 6 | 5 | -24 | 4 | -11 | -26 | -14 | 5-methylthioadenosine | -23.22 | 297 | -0.61 | 9.27 | -0.3 | 119 | 2 |
| 3 | 1 | | -37 | | -16 | -48 | Quinaldic_acid | -24.21 | 173 | 1.93 | 9.34 | 1.6 | 50 | 2 |
| 4 | 2 | | -18 | -41 | -35 | | Phenylacetyl-glycine | -9.74 | 193 | 0.43 | 10.07 | 0.7 | 66 | 1 |
| 6 | 6 | -24 | 4 | -11 | -7 | -18 | Riboflavin | -25.43 | 376 | -1.72 | 11.49 | -1.5 | 162 | 1 |
| 8 | 6 | | -18 | -39 | -27 | -23 | N-acetyl-L-phenylalanine | -10.08 | 207 | 0.82 | 12.27 | 0.6 | 66 | 1 |
| 4 | 2 | | -31 | | | | Indoleacetic_acid | -10.48 | 175 | 1.80 | 13.55 | 1.4 | 53 | 2 |
| 14 | 13 | | | | 12 | | Furosemide | -6.6 | 331 | 1.89 | 15.2 | 2 | 123 | 2 |
| 6 | 4 | | | | | | Azelaic_acid | 1.4 | 188 | 1.89 | 15.27 | 1.6 | 75 | 0 |
| 7 | 6 | | -18 | | -27 | -17 | Phloretin | -10.19 | 274 | 2.32 | 16.99 | 2.6 | 98 | 2 |

(Continues)

TABLE 1 (Continued)

| N2 | Ch | IPA | Ch_IPA | EtOH | Tol | ACN | Name | Gb | MW | logP | RT | logp | TPSA | Ar |
|----|----|-----|--------|------|-----|-----|------------------|--------|-----|------|-------|------|------|----|
| 5 | 5 | | 0 | | -12 | -1 | Cortisone | -16.15 | 360 | 1.99 | 17.07 | 1.5 | 92 | 0 |
| 6 | 6 | 6 | 6 | 6 | 7 | | Clotrimazole | -29.8 | 345 | 5.38 | 19.76 | 5 | 18 | 4 |
| 8 | 7 | -13 | 1 | -9 | 5 | -10 | Taurocholic_acid | -25 | 516 | 2.37 | 20.7 | 2.2 | 144 | 0 |
| 9 | 8 | -10 | 3 | -6 | 4 | -6 | Glycocholic_acid | -25 | 466 | 2.56 | 21.9 | 2.9 | 127 | 0 |

during optimization. The optimization of CoV for each peptide was first achieved without modifiers (only N2 as the separation gas). Then, three organic modifiers (acetonitrile, isopropanol, and methanol) were added separately at 1.5% in the separation gas to optimize the CoV for each modifier for all the peptides.

2.3 | Computational details

All DFT calculations were performed with the Amsterdam Density Functional (ADF) program package, within the Amsterdam Modelling Suite (AMS2021) modeling suite.²⁵ Initial structures were optimized with PBE DFT method²⁶ using full electron TZ2P Slater type orbitals basis, Grimme G4 dispersion correction.²⁷ The nature of the stationary points is confirmed by calculating analytical Hessians. Since the nature of protonation during ESI in DMS analyses—whether kinetically or thermodynamically controlled—remains uncertain with existing literature provides evidence supporting both mechanisms,^{28,29} we have included the later one. It is very straightforward to account for it using a simple difference between COSMO (IPA) PBE-D4 Gibbs free energy of protonated and neutral molecule as a qualitative measure of basicity and will be called Differential Proton Affinity in the rest of the manuscript.

2.4 | Data analysis pipeline

2.4.1 | Small molecules

Data in Table S1 are directly imported into pandas data frame, after which we converted molecular names to SMILES using pubchempy and then to RDKit mol objects, which were a starting point for all further property calculations. The obtained mol objects were then used to calculate MW, logP, TPSA, number of aromatic rings in a molecule, and a starting point (by providing 3D structure) for the proton affinity DFT calculations, labeled as Gb (Table 1). Thus, although this is only a 50 molecules mixture, all steps are fully automated and can be applied to datasets containing much larger number of molecules.

2.4.2 | Peptides

Data in Table S1 are directly imported into pandas data frame, after which we converted peptide sequences to their 3D structure using

alphafold, as implemented with ColabFold. The obtained .pdb files are then easily converted to RDKit mol objects, which were a starting point for all further property calculations. The obtained mol objects were then used to calculate MW, logP, and TPSA. Sum of Ser/Thr (ST) AAs, sum of Hys/Arg/Lys (basic) AAs, sum of Asp/Glu (acid), length of peptide, and ccs are calculated directly from peptide sequence. All steps are fully automated and can be applied to datasets containing a significantly larger number of molecules.

3 | RESULTS AND DISCUSSION

3.1 | DMS analysis of metabolites

In this subsection, we analyze our recent results²⁴ for a 50 analytes mixture (Figure S1 and Table 1), representative for urine and plasma metabolites, Figure 1. The data include respective LC retention time (RT) and CoV values for pure N₂, 1.5% mole ratio of cyclohexane (Ch), ethanol (EtOH), isopropanol (IPA), toluene (Tol), acetonitrile (ACN), and one binary modifier: 0.05% mole ratio IPA in Ch.

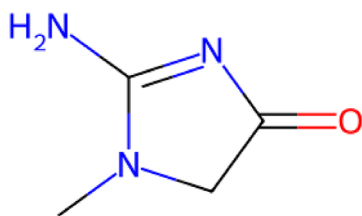
The utilization of a binary modifier is very closely related to one of the subjects of this manuscript—signal suppression with interacting modifiers. IPA of 1.5% is one of the most utilized modifiers due to its large CoV range (peak capacity). Unfortunately, it is usually coupled with analyte signal suppression (25 out of 50 metabolites could not be detected from our small molecules dataset; Figure 2), presumably due to gas-phase proton transfer reactions with IPA or a possible out-of-range CoV value. By mixing IPA with cyclohexane and lowering the concentration to 0.05%, the created binary modifier retains a reasonable peak capacity, with 45/50 analytes being detectable.¹⁷

3.2 | Relationship between various properties

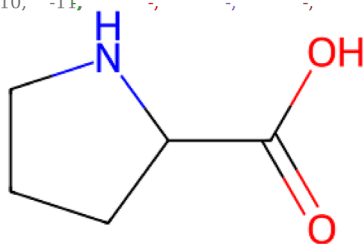
One of the most useful steps to capture various (linear) relationships between many data variables is construction of correlation matrix, Figure 3A. As the name implies, this is the matrix consisting of correlation coefficients (r), giving the direct information about the strength and direction of a linear relationship. If we are more interested about the portion of a variance explain by the variable, the coefficient of determination (r^2) gives a better insight. For that reason, we will most focus on “R² matrix,” given in Figure 3B. Beside properties seen in Table 1 and Figure 3, CCS was also calculated, using graph neural network approach implemented in SigmaCCS.³⁰ Given the exceptionally

1. Creatinine

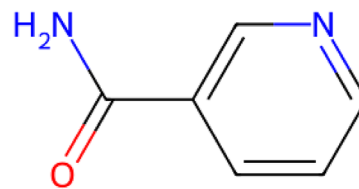
N₂, Ch, Ch_IPA, EtOH, IPA, Tol, ACN, N₂, Ch, Ch_IPA, EtOH, IPA, Tol, ACN, N₂, Ch, Ch_IPA, EtOH, IPA, Tol, ACN,
-10, 7, 5.0, -, -, 3.0, -, -10, -11, -, -, -, -, -6, -7, -, -, -41.0, -, -



2. L-Proline

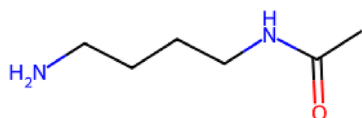


3. Niacinamide



4. N-acetylputrescine

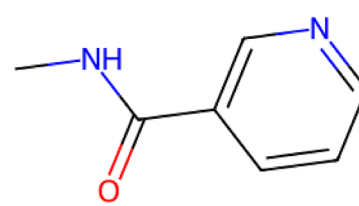
N₂, Ch, Ch_IPA, EtOH, IPA, Tol, ACN, N₂, Ch, Ch_IPA, EtOH, IPA, Tol, ACN, N₂, Ch, Ch_IPA, EtOH, IPA, Tol, ACN,
3, 1, -37.0, -, -50.0, -, -46.0, -1, -3, 4.0, -, -, -, -, -2, -4, -36.0, -35.0, -36.0, -, -38.0



5. Creatine

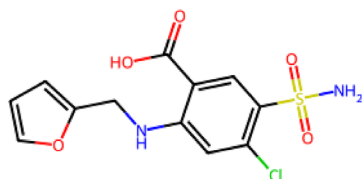


6. N-methylnicotinamide

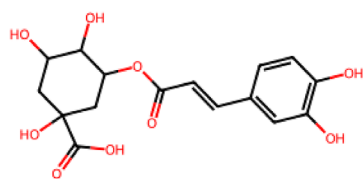


7. Furosemide

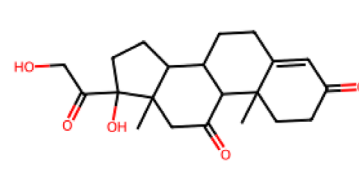
N₂, Ch, Ch_IPA, EtOH, IPA, Tol, ACN, N₂, Ch, Ch_IPA, EtOH, IPA, Tol, ACN, N₂, Ch, Ch_IPA, EtOH, IPA, Tol, ACN,
14, 13, -, -, -, 12.0, -, 7, 6, -11.0, -, -, 5.0, -, 5, 5, 0.0, -, -, -12.0, -1.0



8. Chlorogenic acid

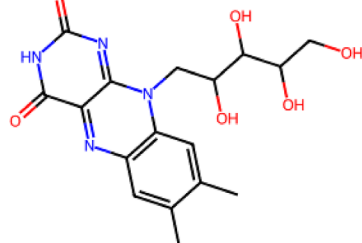


9. Cortisone

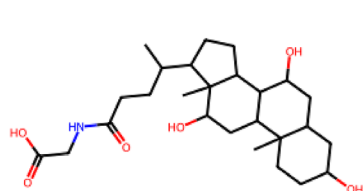


10. Riboflavin

N₂, Ch, Ch_IPA, EtOH, IPA, Tol, ACN, N₂, Ch, Ch_IPA, EtOH, IPA, Tol, ACN, N₂, Ch, Ch_IPA, EtOH, IPA, Tol, ACN,
6, 6, 4.0, -11.0, -24.0, -7.0, -18.0, 9, 8, 3.0, -6.0, -10.0, 4.0, -6.0, 8, 7, 1.0, -9.0, -13.0, 5.0, -10.0



11. Glycocholic acid



12. Taurocholic acid

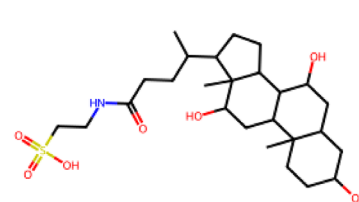


FIGURE 1 Structure of selected 12 analytes (6 smallest and 6 largest) with the respective retention time and CoV values (N₂, 1.5% mole ratio cyclohexane [Ch], ethanol [EtOH], isopropanol [IPA], toluene [Tol], acetonitrile [ACN], and one binary modifier: 0.05% mole ratio IPA in Ch)

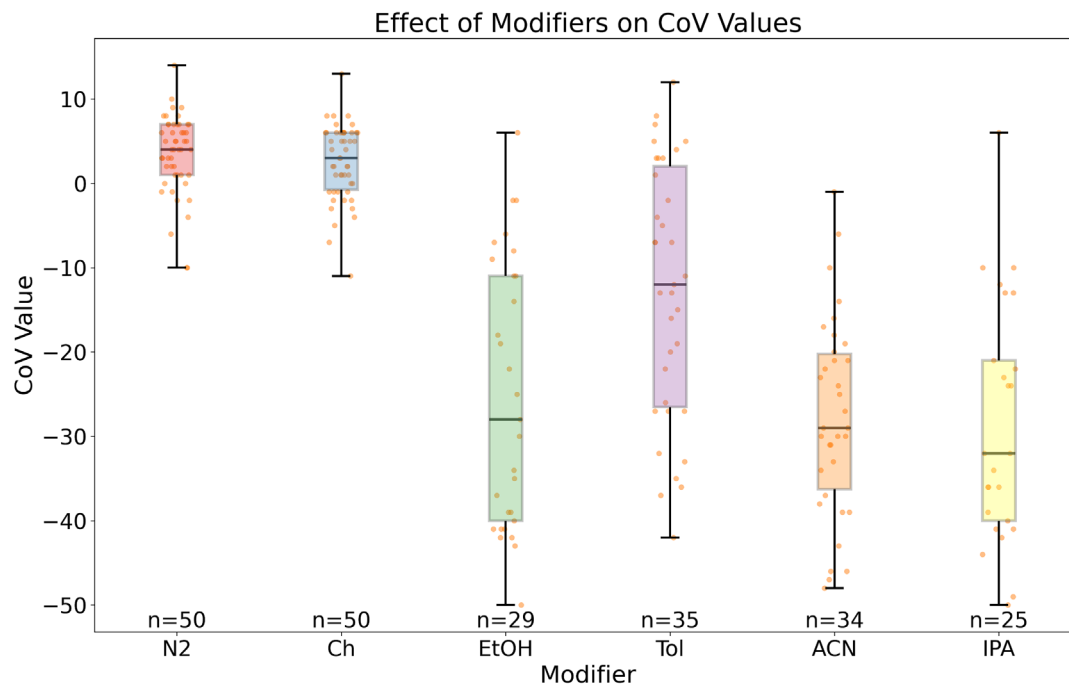


FIGURE 2 Box plot showing the effect of adding modifiers with different proton affinities (PAs) on CoV values. The modifiers were sorted based on their PA values from lowest (left) to highest (right).

high correlation with MW ($R^2 > 0.98$; Figure S2), it was concluded that, essentially, the same information is captured by CCS and is no longer discussed as a predictor for mixture 50 dataset.

Inspection of Figure 3B highlights significant variance in data. The correlation between CoV (EtOH) and CoV (IPA) stands out, with an R^2 value of 0.81. There is a distinguishable variance between Cyclohexane and N₂, showing an R^2 of 0.68. Molecular weight correlates with several parameters: CoV(N₂), CoV (cyclohexane), CoV (ACN), retention time, and TPSA. Additionally, retention time shows links to logP, molecular weight, CoV (ACN), proton affinity (Gb), and CoV (IPA).

Regarding CoV for non-clustering modifiers, the relationship is essentially linear, with one outlier (Creatinine). By removing it, R^2 goes from 0.68 to 0.97, Figure S3. This is a very strong indicator that there is a problem with CoV (cyclohexane) value for creatinine.

Strong correlation and captured variance for CoV (EtOH) and CoV (IPA) is expected since they are both alcohols differing by only one carbon atom, Figure S4.

Only moderate correlation CoV for non-clustering modifiers with MW (and CCS, Figure S2) having $R^2 = 0.4$ – 0.44 is not surprising and clearly demonstrates the point made in the introduction that the low-field limit ion mobility technologies represent much more natural choice for CCS extraction.

Observed correlations naturally lead toward an attempt to connect some of the values mentioned in a functional relationship. This task is greatly facilitated by the further reduction of initially small dataset as a consequence of signal suppression by polar modifiers (vide infra). This makes application of any more complex ML regression model, beside linear regression (LR), a task doomed to produce

large overfitting (noise captured) because of a high model complexity, very limited amount of data and highly complicated relationship between measured quantities and chemical structure. Even with LR, high generatability should not be expected since there are, for example, only 25 data points available for CoV (IPA).

Still, it is instructive to show that by only using simply calculated properties like MW, logP, and proton affinities, we capture retention time with $R^2 = 0.9$.

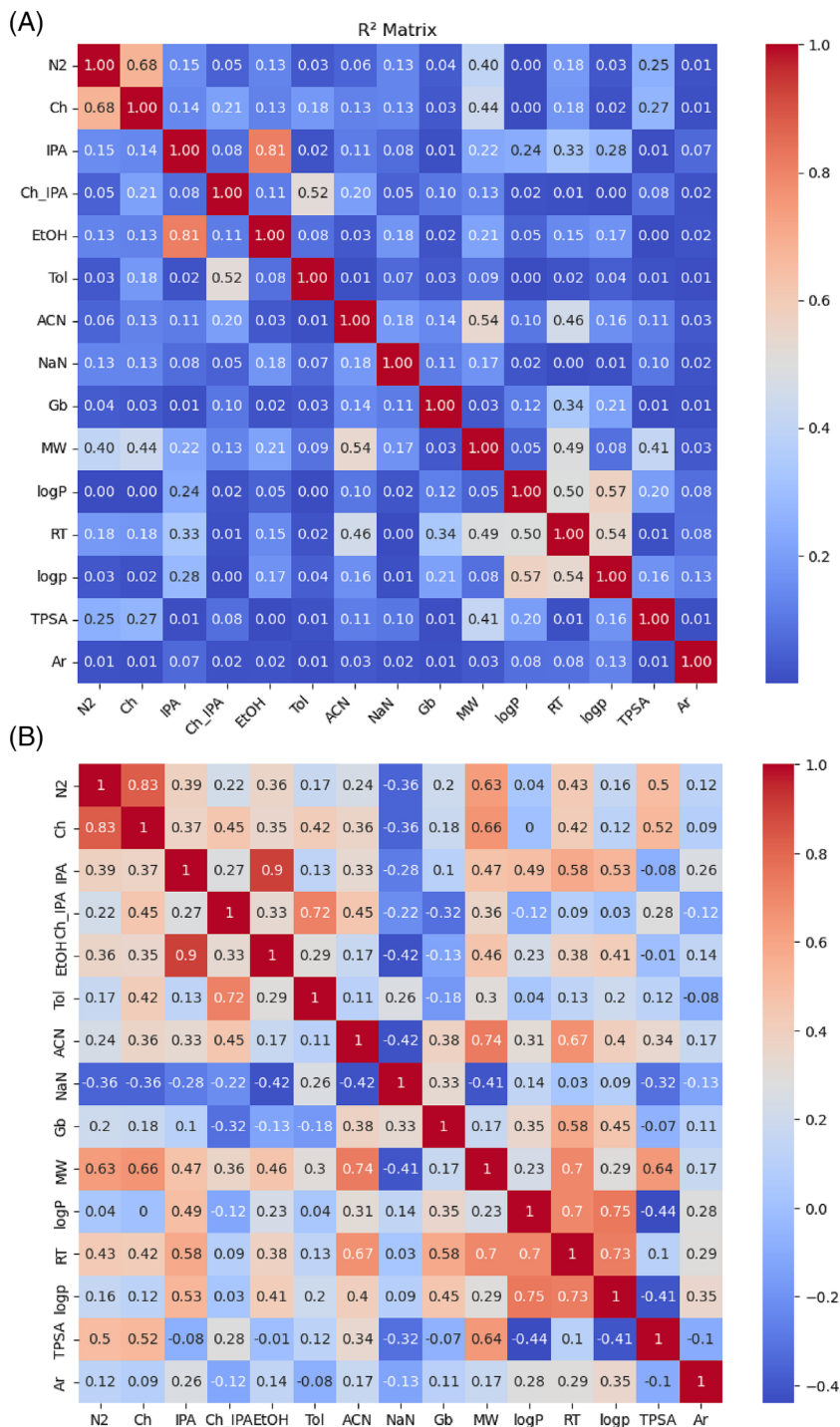
$$RT = 1.72 \cdot \log P + 0.035 \cdot MW + 0.198 \cdot Gb + 3.95, \quad R^2 = 0.9$$

If we allow the utilization of obtained CoV values, by replacing Gb with CoV (IPA), the fit improves to $R^2 > 0.92$. All of this by using no more than three features, to reduce the possibility of overfitting.

$$RT = 1.64 \cdot \log P + 0.042 \cdot MW - 0.004 \cdot \text{CoV(IPA)} - 4.1, \quad R^2 = 0.92$$

While the relationships in our study primarily highlight linear correlations, capturing non-linear patterns with such a limited dataset poses a challenge. For example, in Figure 2B, the “Ar” column, indicating the number of aromatic rings, shows negligible correlation ($R^2 \approx 0$) with CoV (Toluene). This lack of correlation is surprising, as π - π interactions are expected. However, Figure S5 reveals a trend: most molecules without aromatic rings (marked with zeros) correspond to smaller CoV (Toluene) values. An exception is homocitrulline, positioned lower in the plot, with a significantly negative CoV (Acetonitrile). This suggests homocitrulline's proton donor potential, possibly facilitating strong cation- π interactions with Toluene.

FIGURE 3 Linear relationship quantifiers among the various experimental and calculated properties from mix 50 dataset: (A) correlation matrix and (B) R^2 matrix



3.3 | Metabolites and signal suppression

After noting some simple relationships and correlations, the main part of the analysis will be to provide microscopic insight into signal suppression with polar modifiers. Out of the 50 analytes used, with non-clustering modifiers, all 50 are observed, while with IPA, cyclohexane/IPA, EtOH, Toluene, and ACN, we detect 25, 45, 29, 35, and 34, respectively.

It seems plausible to assume that the missing values are due two main factors: gas-phase proton transfer reactions with IPA and a possible out-of-range CoV value. The goal of this section is to investigate

these mechanisms and try to gain more insight using only data science perspective.

Since we cannot use CoV (ToI) on the axis and see its missing values, we decided to plot MW vs TPSA, and color the points on whether was the signal for CoV (ToI) detected (blue) or not (yellow), Figure 4A. It is clear that all the missing values are in a small MW range, which is not surprising since they are most affected with clustering/delustering mechanism. When we take a look at non-detected CoV signal for EtOH, IPA, and ACN (Figure 4B–D), similarly to CoV (ToI), we see a small MW cluster (presumably) from out of range CoV

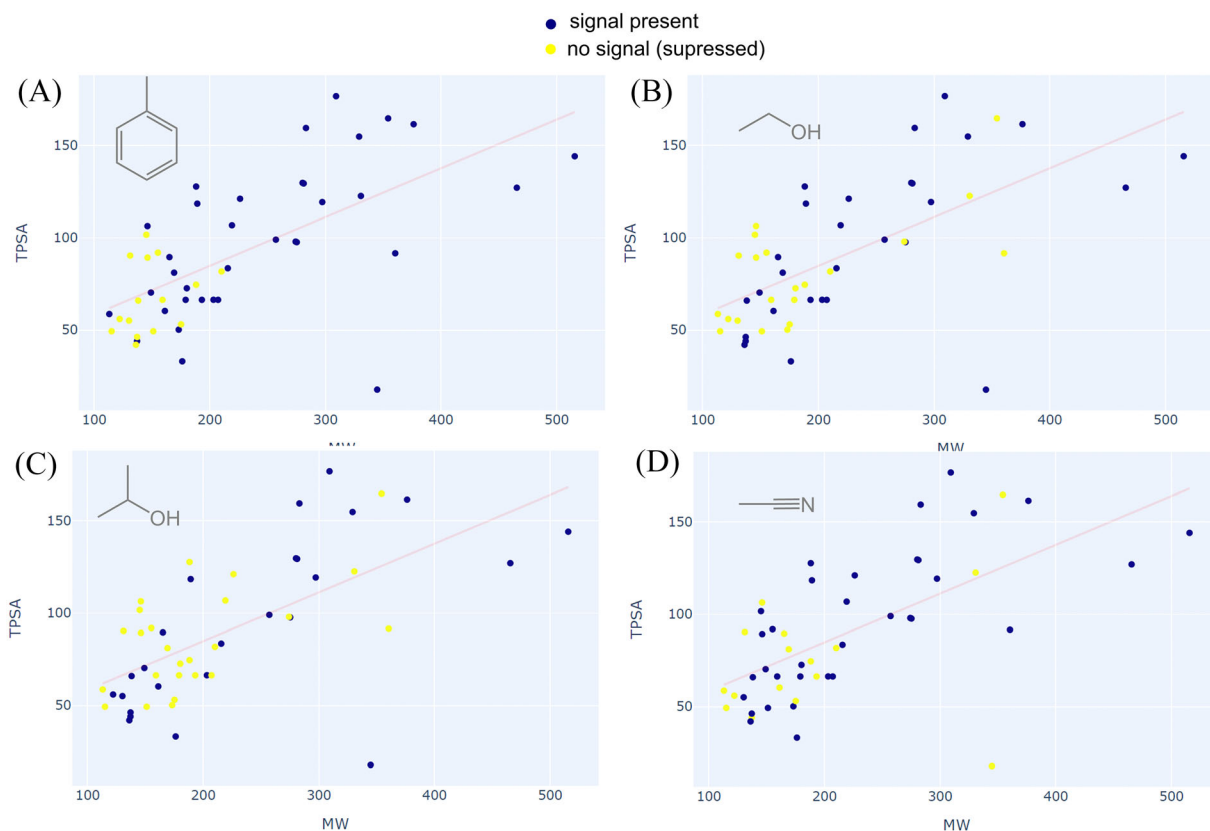


FIGURE 4 TPSA versus MW, points are colored in accordance with the presence (blue)/absence (yellow) of CoV signal. (A) Presence/absence of CoV (Tol), (B) presence/absence of CoV (EtOH), (C) presence/absence of CoV (IPA), and (D) presence/absence of CoV (ACN)

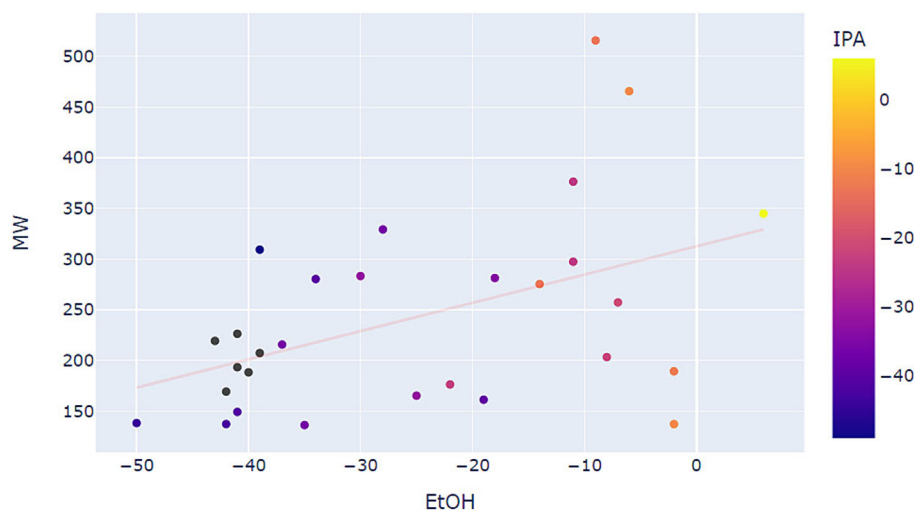


FIGURE 5 CoV (EtOH) versus MW, points are colored in accordance with CoV (IPA) signal (the gradient is explained on the legend), and the back points do not have detected CoV (IPA).

values. Furthermore, with these modifiers that have some proton acceptor properties, we also observe a second group of missing values, at medium MW values, presumably originating from proton transfer reactions with modifier. Further inspection confirms that all these systems indeed have a small basicity, as expected for proton transfer participant.

To provide the clearest evidence that the observed initial grouping of low molecular weight (MW) compounds primarily signifies

signals undetected due to being out of range, a novel approach is adopted. This involves leveraging the high correlation between the coefficients of variation for ethanol (CoV [EtOH]) and isopropanol (CoV [IPA]), alongside the observation that certain missing IPA values are actually present in EtOH data. By plotting MW against CoV (EtOH) and coloring the data points based on CoV (IPA) values, Figure 5, we establish a direct correlation between CoV (EtOH) on the x-axis and the CoV (IPA) indicated by the colors of the points.

TABLE 2 The chemical properties and optimal compensation voltage (CoV) values for a portion of 185 peptides with/without modifiers, separation voltage (SV) = 3,500 V

| | Peptide sequence | Charge | N2 | MeOH | IPA | ACN | TPSA | logP | MW | Css | Acid | NaN | ST | Basic | Acidic | Length |
|----|------------------|--------|------|------|-------|-------|------|------|--------|-----|------|-----|----|-------|--------|--------|
| 0 | VCNQIEFLNTEFK | 2 | 9.4 | 6.1 | -4.6 | -0.6 | 666 | -7 | 1583.8 | 421 | 1 | 0 | 1 | 1 | 2 | 13 |
| 1 | QAEELIQEHADQAEIR | 3 | 14.7 | | | | 983 | -14 | 2007.0 | 544 | 2 | 1 | 0 | 2 | 5 | 17 |
| 2 | DNFDIAEGVR | 2 | 11.0 | 3.8 | -17.1 | -7.2 | 545 | -8 | 1134.5 | 348 | 1 | 0 | 0 | 1 | 3 | 10 |
| 3 | VNYNFEDETVR | 2 | 10.9 | 4.4 | -11.3 | -4.2 | 658 | -9 | 1384.6 | 386 | 1 | 0 | 1 | 1 | 3 | 11 |
| 4 | HGFLEGR | 2 | 18.9 | 4.8 | -13.3 | -4.6 | 370 | -6 | 814.4 | 314 | 2 | 0 | 0 | 2 | 1 | 7 |
| 5 | LDLDQDYR | 2 | 12.1 | -1.1 | -19.8 | -7.8 | 508 | -7 | 1036.5 | 346 | 1 | 0 | 0 | 1 | 3 | 8 |
| 6 | LLEGEER | 2 | 11.2 | -2.1 | -21.7 | -8.0 | 444 | -7 | 901.5 | 321 | 1 | 0 | 0 | 1 | 3 | 8 |
| 7 | AFLPVTSPNK | 2 | 10.5 | 5.1 | -7.4 | -6.4 | 420 | -5 | 1072.6 | 343 | 1 | 0 | 2 | 1 | 0 | 10 |
| 8 | FDSVHSK | 2 | 10.8 | -5.8 | -22.5 | -10.8 | 375 | -6 | 818.4 | 307 | 2 | 0 | 2 | 2 | 1 | 7 |
| 9 | NSTIEYDGVMSK | 2 | 10.4 | 4.7 | -11.8 | -5.5 | 634 | -9 | 1383.6 | 388 | 1 | 0 | 2 | 1 | 2 | 12 |
| 10 | FENLGVSSLGER | 2 | 10.8 | 4.5 | -10.2 | -4.4 | 607 | -10 | 1306.7 | 378 | 1 | 0 | 2 | 1 | 2 | 12 |
| 11 | SFLYEIVSNK | 2 | 10.6 | 2.6 | -4.9 | -4.9 | 496 | -5 | 1198.6 | 372 | 1 | 1 | 2 | 1 | 1 | 10 |
| 12 | CTCISISNQPVNPR | 2 | 9.1 | 5.9 | -1.6 | -1.8 | 679 | -13 | 1530.7 | 413 | 1 | 0 | 3 | 1 | 0 | 14 |
| 13 | LEIPASQFCPR | 2 | 9.6 | 6.4 | -4.3 | -2.4 | 532 | -6 | 1372.7 | 395 | 1 | 0 | 1 | 1 | 1 | 12 |
| 14 | VALHILDEEDK | 3 | 16.5 | | | | 534 | -6 | 1167.6 | 435 | 2 | 1 | 0 | 2 | 4 | 10 |

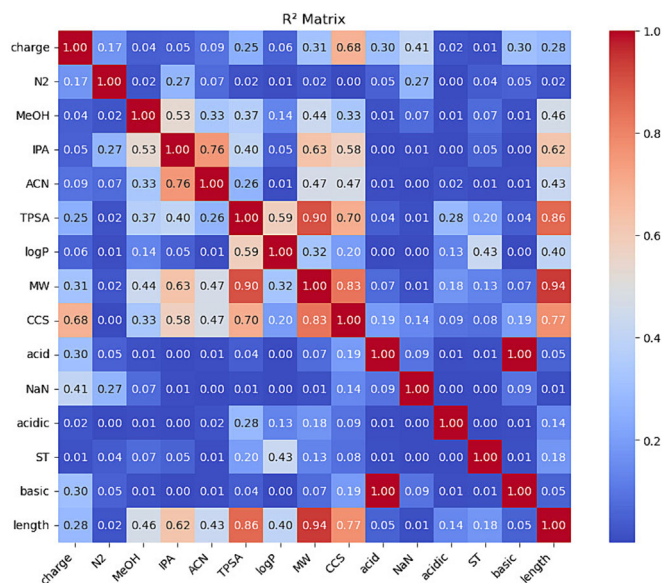


FIGURE 6 “R² matrix” for the various experimental and calculated properties from 185 peptides dataset

conclusively identified as part of the “yellow cluster” of low MW Figure 4, thereby confirming their characterization as signals not detected due to being beyond the detection range.

3.4 | DMS analysis of peptides

In section, we analyze the results for a mixture of 185 small to medium sized peptides with 6–20 amino-acids (AAs). List of all experimental data can be found in Table S2. As will be explained in the data analysis pipeline below, the steps were almost identical to the small molecule section, and the portion of the final table (because the full one is too large) with additional properties is given below (Table 2).

3.5 | Relationship between various properties

“R² matrix” for the 185 peptides is presented in Figure 6, which reveals significant variance among clusters of chemical properties. As

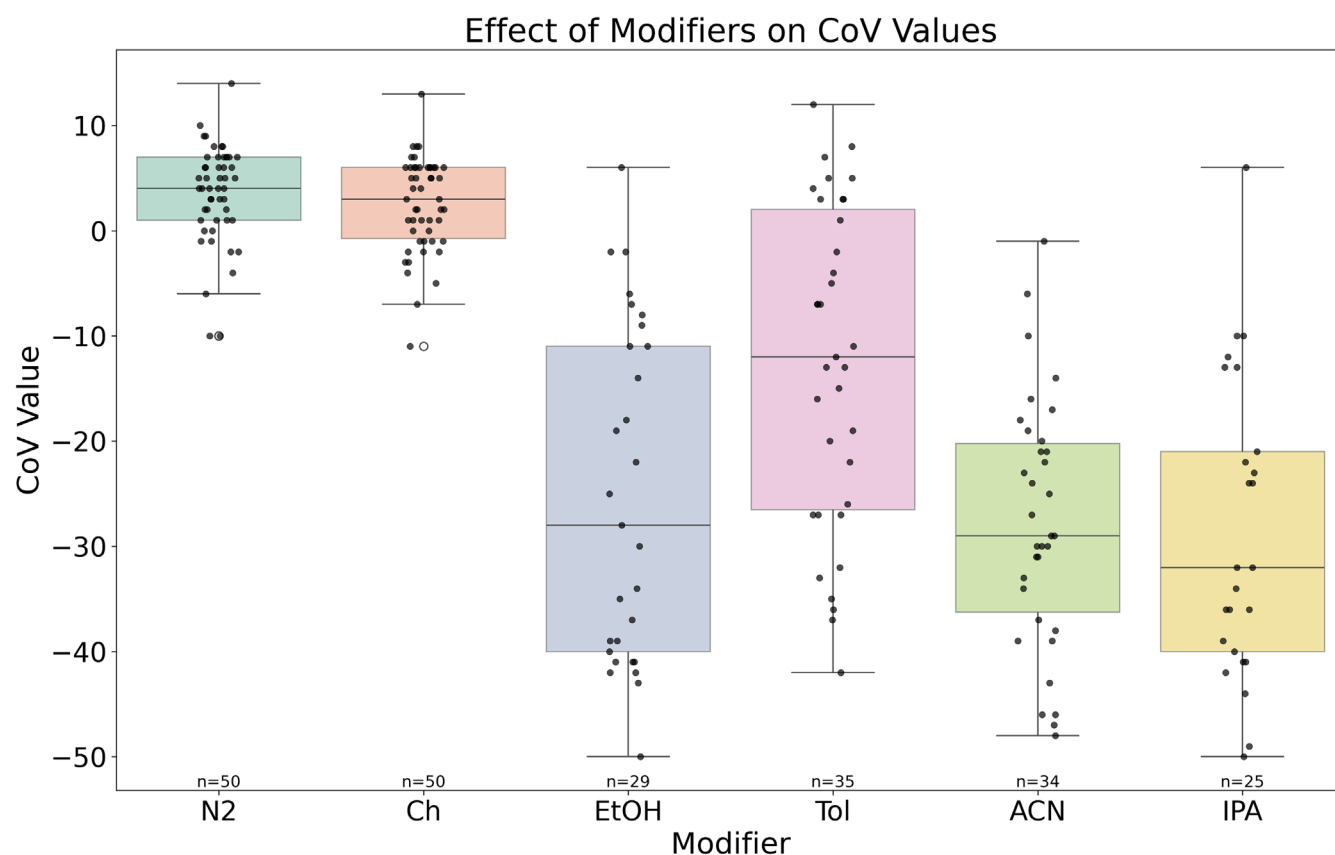
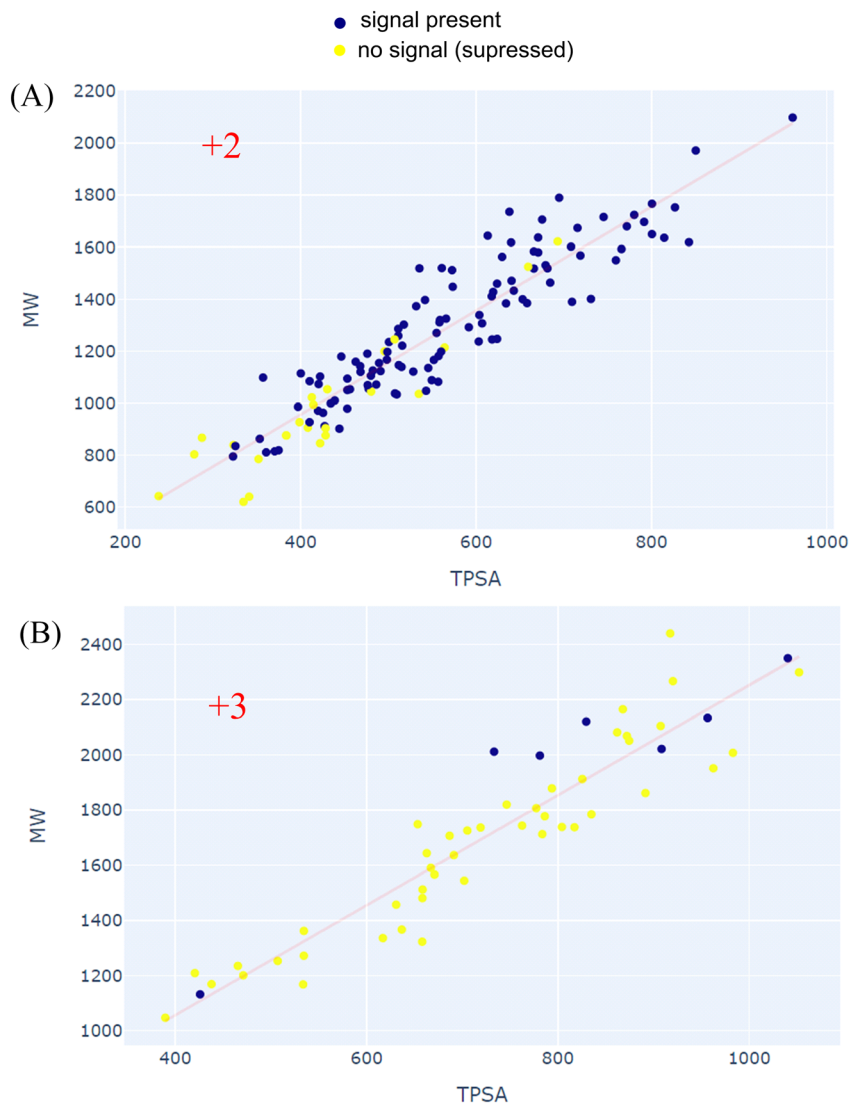


FIGURE 7 Box plot showing the effect of adding modifiers with different proton affinities (PAs) on CoV values. The modifiers were sorted based on their PA values from lowest (left) to highest (right).

Notably, at the extreme left of the plot—adjacent to the boundary of the CoV (EtOH) range—a significant number of points appear in black, indicating the absence of CoV (IPA) values. These points are

predicted, properties like CSS, MW, charge, TPSA, and peptide length are interconnected. MW, along with the previously mentioned attributes, shows a strong correlation with the CoV for all polar modifiers.

FIGURE 8 TPSA versus MW for 185 peptides, points are colored in accordance with the signal suppression, blue for MS signal present and yellow for MS signal suppressed. (A) Peptide charge = 2⁺ and (B) peptide charge = 3⁺



Interestingly, the correlation between $\text{CoV}(\text{N}_2)$ and MW, CCS, or peptide length is nearly zero. The CoV for IPA closely matches with MeOH ($R^2 = 0.53$) and ACN ($R^2 = 0.76$). Moreover, the absence of CoV values, indicated by a NaN column, correlates well with the peptide's charge ($R^2 = 0.41$).

The good correlation between MW with the CoV for all polar modifiers, as well as their good intercorrelation (IPA with both MeOH and ACN), is not surprising since all peptides essentially possess similar set of functional groups (FGs) relevant for analyte-modifier interactions.

Since the RT is not contained in the data, we will give an equation combining CoV (IPA) with MW, TPSA, and peptide charge (three very easily calculable properties) as a strong indication of interrelationship:

$$\begin{aligned} \text{CoV}(\text{IPA}) &= 0.027 \cdot \text{MW} + -0.03 \cdot \text{TPSA} + -5.998 \cdot \text{charge} + \\ &\quad - 16.19, R^2 \\ &= 0.78. \end{aligned}$$

3.6 | Signal suppression with peptides

In the first section of the manuscript, it was demonstrated that we can pinpoint the reasons behind missing CoV values to gas-phase proton transfer reactions with modifier and an out-of-range CoV value. The goal of this section is to investigate these mechanisms and see their importance with a completely different set of compounds. When we look at the CoV ranges for our peptides: N_2 (21.72 V to 7.19 V), MeOH (9.51 V to -6.93 V), IPA (-1.28 V to -23.63 V), and ACN (1.69 V to -10.81 V), Figure 7. Since CoV was ramped from -50 to 50 V (steps of 0.2 V) during optimization, it is clear that proton transfer to modifier represents a probable first step in signal suppression.

Since we have more molecules in our peptide dataset, we were able to successfully apply some important ML classification methods in order to predict whether a peptide would be detected with polar modifier or not. Here, we provide just a quick overview, detailed description, and results of all the models can be found in the

supporting information, section ML models. The features used are peptide charge, CoV(N₂), TPSA, MW, number of basic AAs (base), number of Ser and Thr residuer (ST), number of acidic sidechains (acid), and length of peptide. All features were scaled, and fivefold cross validation was used to test the accuracy with all models. The detailed grid hyperparameter search for Support Vector, Decision Tree, Random Forest, XGBoost, K-Nearest Neighbors (KNN), Logistic Regression, and Ridge classifiers is performed. The results are also compared with “Logistic Regression CV” available under scikit-learn package, which is a form of logistic regression which includes built-in cross-validation to find the optimal value of regularization parameter (C). Best results were obtained with random forests and Logistic Regression CV, but we focused our interpretation on Logistic Regression, as it provides a clearer linear relationship among the features in the built model. The Logistic Regression CV model demonstrated robust performance with an accuracy of 0.89, precision of 0.93, recall of 0.81, F1 score of 0.87, and an AUC-ROC of 0.90.

In order to check which of our features are the most important for predicting peptide CoV value suppression, we used coefficient analysis, permutation feature importance, SHAP Values, and recursive feature elimination (RFE). First three methods indicated the importance of peptide charge and three strongly related features describing peptide size: MW, TPSA, and peptide length. RFE also indicated the importance of basic amino acids count.

We will now plot all these important features to gain some visual perspective. We will use MW and TPSA as axes and present/suppressed CoV values as color, and we will make different plot for charge 2/3 Figure 8.

It can clearly be seen at Figure 8 that signal suppression mostly occurs for the 3 + charged peptides, while for the ones with charge 2+, there is a cluster at the small MW region. This indicates that the peptide charge density is crucial for the possibility of modifier to participate in proton transfer. This is not surprising; higher charge density makes proton loss much more favorable. This reason is once more confirmed when we take a look at the 3+ peptides that reached the detector; with only one exception, they are all in the MW > 2000 Da region, the one with lowest charge density. The effect of basic AAs can be seen on Figure S6, which shows that most of the 2+ peptides with suppressed signal have only one basic AA, and suppressed 3+ peptides have up to two of them. In short, the peptides with non-suppressed signal have more basic amino acids, which makes them weaker proton donors.

4 | CONCLUSIONS

This research employs a data science-based method to understand correlation between modifiers and various molecular properties in DMS and effectively pinpoints two signal suppression mechanisms for small molecules. For peptides, we use machine learning to establish a clear relationship between simple peptide properties and proton transfer-induced signal suppression. The accuracy of our ML model in determining these trends enables a deeper understanding of DMS

behavior, thereby enriching the methodology and analysis of peptides using DMS.

ACKNOWLEDGMENT

This work was supported by the Swiss National Science Foundation, Sinergia Grants CRSII3_136282 and 200021_192306. Open access funding provided by Universite de Geneve.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in the supporting information of this article.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Stepan Stepanovic  <https://orcid.org/0000-0002-4126-0891>

Bandar Alghanem  <https://orcid.org/0000-0002-3414-6580>

Gérard Hopfgartner  <https://orcid.org/0000-0002-9087-606X>

REFERENCES

- Griffiths WJ, Wang Y. Mass spectrometry: from proteomics to metabolomics and lipidomics. *Chem. Soc. Rev.* 2009;38(7):1882-1896. doi:10.1039/B618553N
- Kohler I, Verhoeven M, Haselberg R, Gargano AFG. Hydrophilic interaction chromatography–mass spectrometry for metabolomics and proteomics: state-of-the-art and current trends. *Microchem. J.* 2022; 175:106986. doi:10.1016/j.microc.2021.106986
- Pirok BWJ, Pous-Torres S, Ortiz-Bolsico C, Vivó-Truyols G, Schoenmakers PJ. Program for the interpretive optimization of two-dimensional resolution. *J. Chromatogr. A.* 2016;1450:29-37. doi:10.1016/j.chroma.2016.04.061
- Pirok BWJ, Gargano AFG, Schoenmakers PJ. Optimizing separations in online comprehensive two-dimensional liquid chromatography. *J. Sep. Sci.* 2018;41(1):68-98. doi:10.1002/jssc.201700863
- Stoll DR, Li X, Wang X, Carr PW, Porter SEG, Rutan SC. Fast, comprehensive two-dimensional liquid chromatography. *J. Chromatogr. A.* 2007;1168(1-2):3-43. doi:10.1016/j.chroma.2007.08.054
- Delafield DG, Lu G, Kaminsky CJ, Li L. High-end ion mobility mass spectrometry: a current review of analytical capacity in omics applications and structural investigations. *TRAC-Trend. Anal. Chem.* 2022; 157:116761. doi:10.1016/j.trac.2022.116761
- Eiceman GA, Karpas Z. *Ion Mobility Spectrometry*. CRC Press; 2005. doi:10.1201/9781420038972
- Cohen MJ, Karasek FW. Plasma chromatography™—a new dimension for gas chromatography and mass spectrometry. *J. Chromatogr. Sci.* 1970;8(6):330-337. doi:10.1093/chromsci/8.6.330
- Shvartsburg AA. *Differential ion mobility spectrometry: nonlinear ion transport and fundamentals of FAIMS*. CRC Press; 2008. doi:10.1201/9781420051070
- Guevremont R. High-field asymmetric waveform ion mobility spectrometry: a new tool for mass spectrometry. *J. Chromatogr. A.* 2004; 1058(1-2):3-19. doi:10.1016/j.chroma.2004.08.119
- Wu ST, Xia Y-Q, Jemal M. High-field asymmetric waveform ion mobility spectrometry coupled with liquid chromatography/electrospray ionization tandem mass spectrometry (LC/ESI-FAIMS-MS/MS) multi-component bioanalytical method development, performance evaluation and demonstration of the constancy of the compensation voltage with change of mobile phase composition or flow rate. *Rapid Commun. Mass Spectrom.* 2007;21(22):3667-3676. doi:10.1002/rcm.3264

12. Giles K, Pringle SD, Worthington KR, Little D, Wildgoose JL, Bateman RH. Applications of a travelling wave-based radio-frequency-only stacked ring ion guide. *Rapid Commun. Mass Spectrom.* 2004;18(20):2401-2414. doi:10.1002/rcm.1641
13. Pringle SD, Giles K, Wildgoose JL, et al. An investigation of the mobility separation of some peptide and protein ions using a new hybrid quadrupole/travelling wave IMS/oa-ToF instrument. *Int. J. Mass Spectrom.* 2007;261(1):1-12. doi:10.1016/j.ijms.2006.07.021
14. Fernandez-Lima F, Kaplan DA, Suetering J, Park MA. Gas-phase separation using a trapped ion mobility spectrometer. *Int. J. Ion. Mobil. Spectrom.* 2011;14(2-3):93-98. doi:10.1007/s12127-011-0067-8
15. Fernandez-Lima FA, Kaplan DA, Park MA. Note: integration of trapped ion mobility spectrometry with mass spectrometry. *Rev. Sci. Instrum.* 2011;82(12):126106. doi:10.1063/1.3665933
16. Buryakov IA, Krylov EV, Nazarov EG, Rasulev UK. A new method of separation of multi-atomic ions by mobility at atmospheric pressure using a high-frequency amplitude-asymmetric strong electric field. *Int. J. Mass Spectrom. Ion Processes.* 1993;128(3):143-148. doi:10.1016/0168-1176(93)87062-W
17. Ruskic D, Klont F, Hopfgartner G. Clustering and nonclustering modifier mixtures in differential mobility spectrometry for multidimensional liquid chromatography ion mobility-mass spectrometry analysis. *Anal. Chem.* 2021;93:6638-6645. doi:10.1021/acs.analchem.0c04889
18. Ruskic D, Hopfgartner G. Modifier selectivity effect on differential ion mobility resolution of isomeric drugs and multidimensional liquid chromatography ion mobility analysis. *Anal. Chem.* 2019;91(18):11670-11677. doi:10.1021/acs.analchem.9b02212
19. Walker SWC, Anwar A, Psutka JM, et al. Determining molecular properties with differential mobility spectrometry and machine learning. *Nat. Commun.* 2018;9(1):5096. doi:10.1038/s41467-018-07616-w
20. Ieritano C, Campbell JL, Hopkins WS. Predicting differential ion mobility behaviour in silico using machine learning. *Analyst.* 2021;146(15):4737-4743. doi:10.1039/D1AN00557J
21. Stienstra CMK, Ieritano C, Haack A, Hopkins WS. Bridging the gap between differential mobility, log S, and log P using machine learning and SHAP analysis. *Anal. Chem.* 2023;95(27):10309-10321. doi:10.1021/acs.analchem.3c00921
22. Bissonnette JR, Ryan CRM, Ieritano C, Hopkins WS, Haack A. First-principles modeling of preferential solvation in mixed-modifier differential mobility spectrometry. *J. Am. Soc. Mass Spectrom.* 2023;34(7):1417-1427. doi:10.1021/jasms.3c00117
23. Chakraborty P, Rajapakse MY, McCartney MM, Kenyon NJ, Davis CE. Machine learning and signal processing assisted differential mobility spectrometry (DMS) data analysis for chemical identification. *Anal. Methods.* 2022;14(34):3315-3322. doi:10.1039/D2AY00723A
24. Ekmekciu L, Hopfgartner G. Liquid chromatography and differential mobility spectrometry—data-independent mass spectrometry for comprehensive multidimensional separations in metabolomics. *Anal. Bioanal. Chem.* 2023;415(10):1905-1915. doi:10.1007/s00216-023-04602-0
25. Bickelhaupt FM, Baerends EJ. Kohn-sham density functional theory: predicting and understanding chemistry. *Rev. Comp. Ch.* 2000;15:1-86. doi:10.1002/9780470125922.ch1
26. Perdew JP, Burke K, Ernzerhof M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* 1996;77(18):3865-3868. doi:10.1103/PhysRevLett.77.3865
27. Caldeweyher E, Ehlert S, Hansen A, et al. A generally applicable atomic-charge dependent London dispersion correction. *J. Chem. Phys.* 2019;150(15):154122. doi:10.1063/1.5090222
28. Patrick AL, Cismesia AP, Tesler LF, Polfer NC. Effects of ESI conditions on kinetic trapping of the solution-phase protonation isomer of p-aminobenzoic acid in the gas phase. *Int. J. Mass Spectrom.* 2017;418:148-155. doi:10.1016/j.ijms.2016.09.022
29. Joyce JR, Richards DS. Kinetic control of protonation in electrospray ionization. *J. Am. Soc. Mass Spectrom.* 2011;22(2):360-368. doi:10.1007/s13361-010-0037-0
30. Guo R, Zhang Y, Liao Y, et al. Highly accurate and large-scale collision cross sections prediction with graph neural networks. *Commun. Chem.* 2023;6(1):139. doi:10.1038/s42004-023-00939-w

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Stepanovic S, Ekmekciu L, Alghanem B, Hopfgartner G. Insights into modifiers effects in differential mobility spectrometry: A data science approach for metabolomics and peptidomics. *J Mass Spectrom.* 2024;59(6): e5039. doi:10.1002/jms.5039