DEVELOPING MACHINE-LEARNING METHODS FOR QUANTITATION OF ORGANIC COMPOUNDS FROM ELECTRON-IONIZATION MASS SPECTROMETRY

Arnab Bose, Amrutha Raghu, and Phillip R. Westmoreland*
North Carolina State University
Raleigh, NC 27695-7905

Abstract Overview

Our work aims to develop machine-learning methods to quantify compounds from Electron Ionization Mass Spectrometry (EI-MS). This technique uses high-energy electrons (70-75 eV) to ionize and fragment the molecules of a compound entering a mass spectrometer, and the resulting spectrum is characteristic of the species. The total signal is proportional to species concentration through the electron ionization cross section, and the ratio of the total ionization cross sections of two compounds is mainly dependent on the molecular structures of the compounds. In the present study, linear regression analysis and artifical neural network analysis are performed. Molecular descriptors using the concepts of atom, bond, and group additivities are used to predict electron-ionization cross sections. Experimental measurements of EI cross sections of 372 compounds reported in the literature from 1951 to 2019 are being used as training data for the neural network.

Keywords

electron ionization, cross section, GC-MS

Introduction

Electron ionization mass spectrometry (EI-MS) is used worldwide to identify and quantify various organic and inorganic compounds. This technique uses high-energy electrons (70-75 eV) to ionize the molecules of a compound entering to a spectrometer. Once the molecule is ionized, it automatically goes through further fragmentation. During this fragmentation process, positive ions, negative ions, and radicals are generated. In time-of-flight mass spectrometry, a positive pulse of voltage propels the positive ions through a flight tube. Ions with low mass travel faster than heavier ions, and all of them are detected at the end of the flight tube, assigned masses by their arrival times, and counted. The result is a histogram where the x-axis is mass-to-charge ratio (in most of the cases the charge is unity) and the y-axis is total signals of different positive ion fragments. For an electron-ionization mass spectrum at 70 to 75 eV, the fragmentation pattern normalized to the maximum-sized signal (or base-ion signal) is characteristic of a species and its molecular features (McLafferty et al., 1993). Electron ionization is generally applied to mixtures by first resolving the individual species using gas or liquid chromatography so that a nominally pure component is introduced to the mass spectrometer. Analysis software typically reports on the similarity of the species-fragment spectrum from the experiment relative to the mass spectra of species in the mass spectral library. The quantification of total positive-ion current of an identified species is performed by using a linearity relationship (Eq. 1) with the number density N of the vaporized/gaseous species:

$$I_i^+ = Q\, I_e\, d\, N_i \tag{1}$$

where I_i^+ is the current of total positive ions detected for species i, I_e is the ionizing-electron current, d is the ionizing path length, and N_i is the number density of the species i,

which is equal to its quantity in moles, times Avogadro number (A) and volume of the ionization region (V).

Equation 1 is the characteristic equation for electron ionization by collision of a parallel electron beam with homogeneous velocities with solid spherical particles [Massey and Burhop, 1952]. For monoatomic gases, there is one characteristic length of a particle; i.e., diameter of the spherical atoms. For polyatomic gases, the spherical cross-section is viewed as an average of all the cross-sections of all characteristic lengths. The ionizing path length and the volume of the ionization chamber are instrument-specific. On the other hand, the ionizing electron current is specific to experimental conditions, such as type and temperature of ion source. The total electron ionization cross section is specific to a species and its molecular structure at a fixed energy of ionizing electrons. Thus, the ratio of total EI cross sections of two species at a fixed eV should be independent of instrument type and experimental conditions, so it can be cross-checked for interlaboratory reproducibility.

Issues with electron ionization

One major limitation of this technique is the possibility of mis-identification of compounds. Prediction of spectra for species might be a long-term answer to this issue.

Simple comparison of the experimental mass spectra with mass spectra from a mass spectral library, such as NIST/EPA/NIH Mass Spectral Library [NIST 2019], can lead to wrong identification of those compounds whose mass spectra are not available in that library. One reason might be the unavailability of an authentic sample of the compound. An example is anhydro-xylopyranose, generated during hemicellulose pyrolysis process. As this compound is commercially unavailable, its 70-eV EI mass spectrum is not available in any mass spectral library. Thus, identification and quantification of this compound possess a challenge.

There is another class of compounds which are available along with their 70-eV EI mass spectra, but due to their instability in calibration solution, they break apart into smaller species before reaching the mass spectrometer. Various sugars such as D-xylose and D-glucose exhibit this behavior in methanolic or acetonic solutions.

Issues with EI cross sections

Our first, near-term goal is to predict of total EI cross-sections of species to aid quantitation. In the past, efforts have been made to understand the total EI cross-sections of various homologous group series (or classes) of carbon-based compounds in terms of various descriptors, including carbon number, molar volume, polarizability volume, and dynamic susceptibility [Harrison et al., 1967]. Unfortunately, no general trends over all classes are observed, and no good and simple correlations have been found at 70-75 eV with these descriptors. Thus, the prediction of cross-sections with these individual descriptors is difficult. Furthermore, descriptors such as polarizability volume and dynamic susceptibility are not easily available. It makes the prediction more difficult.

A large effort has been put into developing classical and semi-classical, and quantum methods of total ionization cross-section. The major limitations of these methods are the limited experimental cross section data for complex molecules, time-consuming computations, and difficulty in implementation.

In 1983, Fitch and Sauter proposed two linear correlations for EI cross-sections of compounds relative to n-hexane at 70-75 eV. These two simple correlations are based on atom additivity, one having no hybridization and another having the hybridization of C and O. Fitch and Sauter considered the interlaboratory datasets of 179 compounds encompassing cross-section data of n-alkanes, n-alkenes, n-alkynes, cycloalkanes and -alkenes, n-phenyl-alkanes, various halides, linear aldehydes, linear ketones, two nitrogen compounds, two deuterated alkanes and H_2S. For these datasets, linear regression and cross validations were performed to find the coefficients. A major advantage of the correlations is that they are easy to implement for any electron ionization mass spectrometer if the calibration behavior of a species of known cross-section is present, such as n-hexane. However, there are certain limitations. First, these two correlations are based only on atom additivity. Thus, they provide same relative cross-sections of any two isomers with the same atom hybridizations. Moreover, due to the limitations of the datasets, hybridizations of N and S are not considered.

A possible solution

In the present study, relative EI cross-sections of total 54 compounds at 70 eV are experimentally estimated using n-hexane as the reference compound. Calibrations of various linear and heterocyclic oxygenates, alkanes and polynuclear aromatic hydrocarbons (PAHs) are performed in a two-dimensional gas chromatograph (GC x GC, Leco) followed by a time-of-flight mass-spectrometer (TOFMS, Pegasus 4D). Finally, population means of the relative cross sections (of 372 compounds) are calculated using various interlaboratory datasets. The database has relative 70-75 eV cross sections of 22 linear alkanes, 13 linear alkenes, 9 linear alkynes, 5 cycloalkanes, 1 cycloalkene, 25 phenyl hydrocarbons, 10 PAHs, 5 deuterated compounds, 13 alcohols, 7 aldehydes, 19 ketones, 10 ethers, 23 esters, 3 carboxylic acids, 1 anhydride, 20 compounds with multiple and different oxygen-based side groups, 4 linear C-H-S compounds, 4 linear C-H-N compounds, 99 halocarbons, 31 heterophenyls, 4 C-H-N heterocyclics, 5 RNA/DNA bases, 14 furan compounds, 3 dioxane compounds, 3 oxirane compounds, 3 other heterocyclic compounds, and 15 industrial gases/liquids. Two separate analyses are performed on this database, one using hybridization of atoms as descriptors (such as, H, D, F, Cl, Br, I, C sp^3, C sp^2, C sp, O sp^3, O sp^2, N sp^3, N sp^2, N sp, S sp^3, and S sp^2) and the other one using 95 various Benson-type groups as descriptors. Although the atom additivity-based

correlations cannot differentiate isomers with same atomic hybridizations, they are useful for light and small molecules where defining groups are difficult. Often the first molecule member of a homologous group-series falls in this category. However, group additivity is more detailed and sensitive to isomers.

Under each modeling project, linear regressions are performed using cross-validation. An artificial neural network (Multi-Layer Perceptron with backpropagation and RelU as activation function) with a simple architecture (1 hidden layer with two perceptron) is used in both projects to explore the non-linear behaviors using machine learning toolbox in MATLAB. Randomly chosen 70% of the data are used for training and the remaining 30% of the data are used for testing purposes. Iterations are performed until the change of each weighting factor falls below a critical value. After the modeling is completed, sensitivity analyses are performed to reduce the less contributing descriptors. After that, the modeling is repeated one more time.

Finally, predictive models are developed to verify and modify the Fitch and Sauter correlations for EI cross-sections to using atom and group additivities with linear regression models and artificial neural networks. However, deep learning will not be possible with the limited amount of dataset [A. Geron, 2017].

Conclusions

We are developing linear regression models and artificial neural network (ANN) models to predict relative ionization cross sections of various molecules by 70 eV electrons. Due to the popular usage of n-hexane as non-polar solvent and its easy availability in experimental laboratories, it is used as reference compound. A dataset of 372 compounds is used to develop atom-additivity-based and group-additivity-based linear correlations and artificial neural networks with a simple architecture.

References

McLafferty, F.W., Turecek, F. (1993). Interpretation of Mass Spectra, 4th edition. *University Science Book*. Mill Valley, CA.

Massey, H. S. W., Burhop, E. H. S. (1952). Electronic and Ionic Impact Phenomena, 1st edition*, The Clarendon Press*, Oxford, UK.

Harrison, A. G., Jones, E. G., Gupta, S. K., Nagy, G. P. (1966). Total Cross Sections for Ionization by Electron Impact. *Can. J. Chem.*, *44:16*, 1967-1973.

Fitch, W. L., Sauter, A. D. (1983). Calculation of relative electron impact total ionization cross sections for organic molecules. *Analy. Chem.*, *55:6*, 832-835.

National Institute of Standards and Technology (2019). *NIST/EPA/NIH Mass Spectral Library with Search Program (Data NIST v17, Software version 2.3)*, NIST: Gaithersburg, MD. DOI: 10.18434/T4H594

Geron, A. (2017). Hands-On Machine Learning with Scikit-Learn and TensorFlow, 1st edition. *O'Reilly Media, Inc.,* 1005 Gravenstein Highway North, Sebastopol, CA 95472, USA