Machine learning estimator for electron impact ionisation fragmentation patterns

K M Lemishko[1], S Mohr[(*)1], A Nelson[1], J Tennyson[2]

[1] *Quantemol Ltd, 320 City Rd, London EC1V 2NZ, United Kingdom*
[2] *Department of Physics Astronomy, University College London, Gower St., London WC1E 6BT, United Kingdom*
[(*)] s.mohr@quantemol.com

There are many measurements and calculations of total electron impact ionisation cross sections. However, it is often desirable to know the cross sections of the dissociation products resulting from ionisation. Partial ionisation cross sections can be derived from the total ionisation cross section by multiplying it by branching ratios for the production of distinct fragments. Branching ratios, in turn, can be inferred from ionisation mass spectrometry data, assuming that the ratios of charged fragments resulting from electron impact ionisation align with the observed fragments in the mass spectra at the energy where the spectrum was obtained [1].

Unfortunately, the required mass spectrometry data are not always readily available. A viable approach to generate reasonable mass spectra estimations for species lacking experimental data on fragmentation patterns is to use existing mass spectrometry data to train a machine learning model. Machine learning has been previously used in plasma modelling, including for estimation of rate coefficients of heavy species collisions [2] and prediction of total ionisation cross sections [3, 4].

We have developed a machine learning regression model to estimate ionisation mass spectra for the subsequent inference of electron impact ionisation fragmentation patterns. The model was trained on over 6500 instances of mass spectra of species with molecular masses up to 300 Da obtained from the NIST WebBook [5]. The model features included data describing readily available general properties of species, along with data characterising species atoms and bonds, extracted using the RDKit open-source tool [6]. The final prediction algorithm is a voting regressor that combines three optimised machine learning regressors: a random forest regressor [7], a gradient-boosted trees regressor [8], and a multilayer perceptron regressor [9] (Fig. 1).

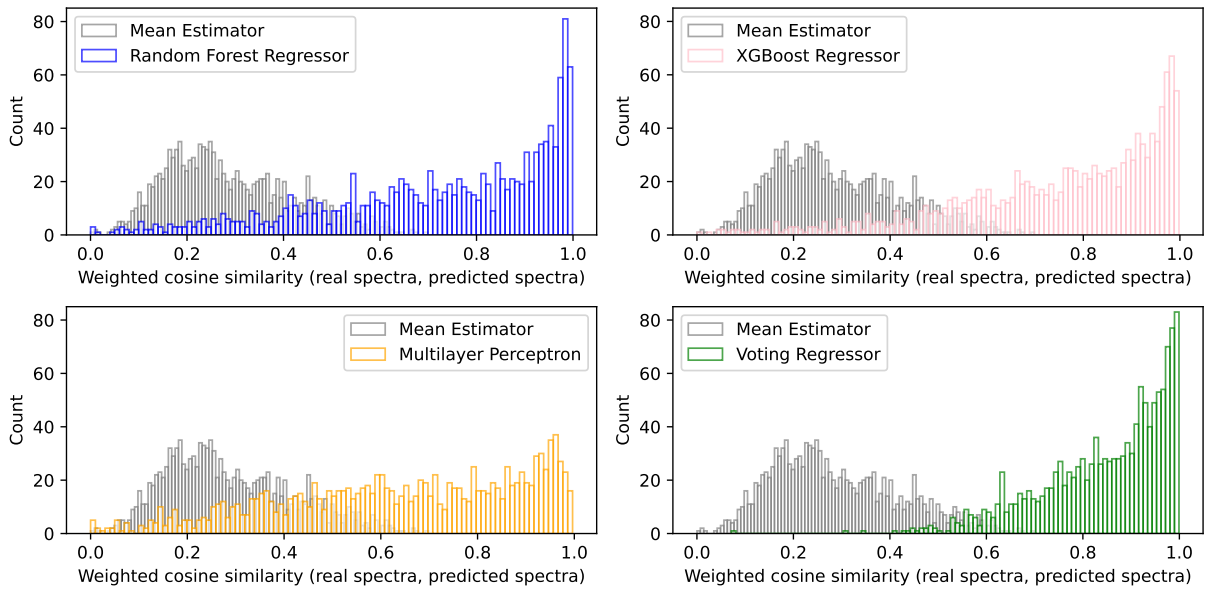

Fig. 1: Comparison of distributions of real and predicted spectra similarities obtained with various machine learning models and the mean estimator for test species.

As a metric of model accuracy, we have used weighted cosine similarity of real and predicted spectra vectors, defined as:

$$cosine\_similarity(y_{\text{real}}, y_{\text{pred}}) = \frac{\sum_{i=0}^{m_{\max}-1} y_{\text{real}}[i] \cdot y_{\text{pred}}[i]}{\sqrt{\sum_{i=0}^{m_{\max}-1} (y_{\text{real}}[i])^2} \cdot \sqrt{\sum_{i=0}^{m_{\max}-1} (y_{\text{pred}}[i])^2}}, \quad (1)$$

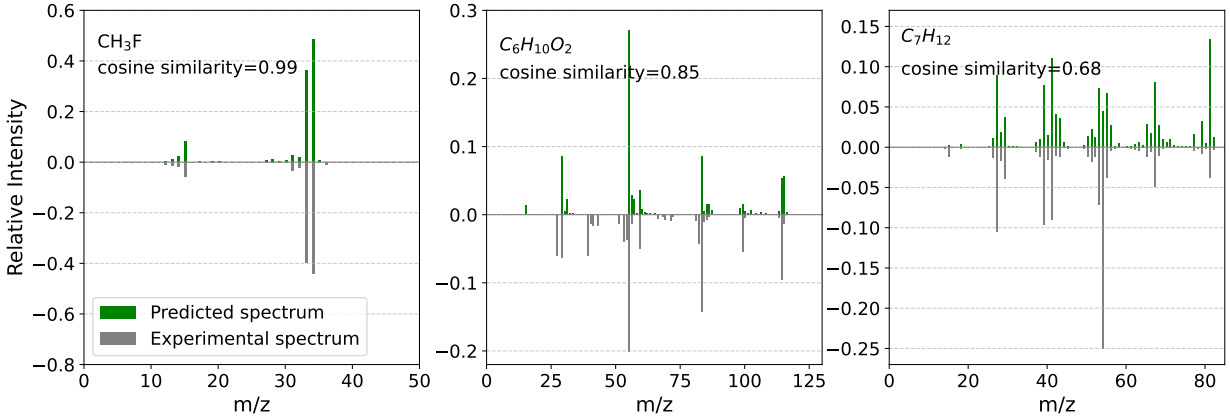

Fig. 2: Examples of real and predicted spectra with different values of cosine similarity

Fig. 2 presents examples of various predicted spectra alongside their corresponding experimental spectra for varying values of cosine similarity. An evaluation of the final prediction algorithm was performed on a dataset containing over 1300 test mass spectra. It was found that for around 70% of the test species, the similarity between the predicted and experimental spectra exceeded 80%. These results demonstrate that the developed machine learning approach can estimate ionisation mass spectra with a reasonable degree of accuracy.

[1] T. D. Mark, *Beiträge aus der Plasmaphysik* **22, 3** (1982) 257–294.
[2] M. Hanicinec, S. Mohr, J. Tennyson, *J. Phys. D: Appl. Phys.* **56** (2023) 374001 (24pp).
[3] L. Zhong, *J. Appl. Phys.* **125** (2019) 183302.
[4] A. L. Harris, J. Nepomuceno, *preprint* (2023).
[5] W.E. Wallace, *Gaithersburg, MD: National Institute of Standards and Technology* (2016) 20899.
[6] RDKit: Open-Source Cheminformatics, *http://www.rdkit.org*
[7] T.K. Ho, *Proceedings of 3rd international conference on document analysis and recognition* **1** (1995) 278—282.
[8] T. Chen, C. Guestrin, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016) 785–794.
[9] S. Haykin, *Prentice Hall PTR* (1994)