Journal of Physics D:
Applied Physics

PAPER • **OPEN ACCESS**

Machine learning-based estimator for electron impact ionization fragmentation patterns

To cite this article: Kateryna M Lemishko *et al* 2025 *J. Phys. D: Appl. Phys.* **58** 105208

OPEN ACCESS
IOP Publishing
Journal of Physics D: Applied Physics
J. Phys. D: Appl. Phys. **58** (2025) 105208 (15pp)
https://doi.org/10.1088/1361-6463/ada37e

Machine learning-based estimator for electron impact ionization fragmentation patterns

Kateryna M Lemishko[1], Gregory S J Armstrong[1], Sebastian Mohr[1], Anna Nelson[1], Jonathan Tennyson[1,2,*] and Peter J Knowles[3]

[1] Quantemol Ltd, 320 City Rd, London EC1V 2NZ, United Kingdom
[2] Department of Physics and Astronomy, University College London, London WC1E 6BT, United Kingdom
[3] School of Chemistry, Cardiff University, Main Building, Park Place, Cardiff CF10 3AT, United Kingdom

E-mail: j.tennyson@ucl.ac.uk

Received 1 July 2024, revised 4 December 2024
Accepted for publication 22 December 2024
Published 9 January 2025

Abstract

Numerous measurements and calculations exist for total electron impact ionization cross sections. However, knowing electron impact ionization fragmentation patterns is important in various scientific fields such as plasma physics, astrochemistry, and environmental sciences. Partial ionization cross sections can be calculated by multiplying total ionization cross sections with branching ratios for different fragments, which can be deduced from ionization mass spectra. However, the required mass spectrometry data is frequently unavailable. A machine learning-based method to predict mass spectra is presented. This method is used to estimate partial electron impact ionization cross sections using the predicted mass spectra and the appearance thresholds for the ionic fragments. As examples, ammonia and the C_2F_5 radical are considered: branching ratios derived from the predicted mass spectra and Binary-Encounter Bethe (BEB) total ionization cross sections are used to predict the fragmentation pattern for each species. The machine learning algorithm can also be used to predict mass spectroscopy fragmentation patterns. While effective, the method has key limitations: it does not account for light fragments such as H^+, whose peaks are absent in the training data, and its validity is restricted to electron impact energies below 100 eV to minimize the contribution of double ionization, which is not accounted for by the BEB model. Although BEB cross sections are used in this work, the method is not reliant on BEB and can be applied to any set of total ionization cross sections, including experimental measurements.

Keywords: machine learning, mass spectrometry, branching ratios, electron impact ionization

* Author to whom any correspondence should be addressed.

1. Introduction

Electron impact ionization cross sections are the subject of intense research interest due to their significance across various scientific domains, including plasma physics [1–4], planetary science [5]. astrochemistry, astrophysics [6], and environmental sciences [7, 8].

Obtaining precise measurements of ionization cross sections, especially for unstable particles like radicals, poses challenges in experimental settings. However, various theoretical methods have been proposed for estimating total ionization cross sections. The most commonly used methods are semi-empirical or approximate; these include the Binary Encounter Bethe (BEB) approximation by Kim and Rudd [9], the DM method introduced by Deutsch *et al* [10, 11], and the spherical-complex optical-potential model by Joshipura *et al* [12]. These methods give generally reliable predictions of the total ionization cross section [13, 14]. There are also completely *ab initio* methods of computing electron impact ionization cross sections for molecules [15–21]; these methods have largely concentrated on small molecules, not least because they are much more computationally expensive than methods such as BEB. However, these *ab initio* methods could also be used to give the total ionization cross sections.

While total electron impact ionization cross sections are valuable, there is often a need to determine partial ionization cross sections. Partial cross sections can be expressed as the total cross section multiplied by branching ratios for the production of distinct fragments. Branching ratios can be inferred from ionization mass spectra, assuming that the ratios of charged fragments resulting from electron impact ionization align with the observed fragments in the mass spectra at the energy when the spectrum was obtained. This implies that the reference energy is sufficiently high for the obtained branching ratios to stabilize.

There have been various attempts to predict fragmentation patterns after impact ionization on theoretical grounds [5, 22–26]; while these studies have provided estimated fragmentation patterns from first principles, an alternative reliable semi-empirical procedure was developed by Hamilton *et al* [27], and subsequently adopted by others [28, 29]. Graves *et al* [25] tested various procedures for predicting fragmentation patterns including the first principles method of Huber *et al* [23] and Hamilton *et al*'s method. They found that the method of Hamilton *et al* [27] gave much more reliable results. This is unsurprising as Hamilton *et al*'s procedure is semi-empirical and involves the use of measured mass spectra to predict the required fragmentation patterns. In principle, combining mass spectroscopy with an accurate total cross section can provide the complete solution to fragmentation problem [30, 31], however there are a number of issues. First, mass spectroscopy data is usually only available for a single electron collision energy; for the library of mass spectroscopy data provided by the National Institute of Standards and Technology (NIST) [32] we use here, this energy is 70 eV. Hamilton *et al* [27] resolved this problem by computing threshold energies for each ion of interest, then using simple curves which (a) go to zero at this threshold, (b) reproduce the observed fragmentation pattern at 70 eV and (c) are normalized to the total (BEB) cross section. We adopt this approach here.

Second, while the NIST Chemistry WebBook database is robust, it does not include many potential species whose partial ionization cross sections are of interest to researchers, particularly in the case of unstable species such as radicals. To address this issue, we have developed a machine learning model trained on mass spectra from the NIST Chemistry WebBook. This model allows us to predict mass spectra for a broader range of chemical species, including those not covered in the experimental database.

Third, many mass spectra, including those provided by NIST, are insensitive to the presence of light fragments, such as H^+ ions. As a consequence, it should be noted that the machine learning model predicts zero intensity for the peak corresponding to H^+ ions, based on the training data characteristics. Therefore, when using mass spectra predicted by the machine learning model for branching ratio calculations, we cannot account for the H^+ fragment. This limitation must be considered when using machine learning-predicted mass spectra for determining partial ionization cross sections. If mass spectrometry data does not include all the required fragmentation patterns, Huber *et al*'s method [23] can be used to provide information on the missing fragments.

Machine learning has gained popularity across various scientific domains, including computational chemistry and plasma modeling. In computational chemistry, machine learning models have been used to predict a wide array of molecular properties, such as atomization energies [33, 34], dipole moments [35], partial charges [36], atomic forces [37], hydration free energies [38], and ionization energies [39]. In plasma physics, machine learning approaches have been used in a variety of applications. For instance, they have been used to solve the stationary Boltzmann equation of electrons in weakly ionized plasmas [40]. Additionally, in plasma processing, neural networks have been utilized to estimate sputtered particle distributions [41], predict plasma operating features associated with the sputtering process [42], and predict plasma etch data [41, 43, 44]. An artificial neural network (ANN) has been also implemented to predict cross sections as functions of energy based on swarm transport data [45]. In addition, machine learning methods have been utilized to predict electron impact ionization cross sections. Specifically, total ionization cross sections were predicted using a support vector machine model trained on BEB estimations for small molecules [46]. Moreover, a recently introduced effective approach involves a simple neural network trained on a small dataset, which predicts total ionization cross sections based on input counts of C, O, N, and H atoms in molecules [47]. Another machine learning-based approach has been proposed to estimate rate coefficients of heavy species collisions, which can supply unknown reaction rate data in plasma chemistries, facilitating the creation of complete chemistry sets without relying on guesswork [48].

As mentioned earlier, partial ionization cross sections can be estimated using fragmentation patterns information inferred from electron ionization mass spectrometry data. A previously reported machine learning-based method for mass spectra prediction introduced the use of Graph Convolutional Network layers to extract structural features from chemical species [49]. In the present work, we aimed to develop a machine learning approach to infer electron impact ionization fragmentation patterns from ionization mass spectra predictions using conventional machine learning algorithms and simple neural networks trained on easily accessible input data.

2. Methods

2.1. Machine learning

In this work, we tested various machine learning algorithms, exploring both conventional supervised learning regressors and deep learning techniques. Overall, we evaluated five machine learning algorithms, including four conventional ones: Random Forest regression [50], XGBoost regression [51], k-Nearest Neighbors (KNN) regression [52], Ridge regression [53], and a simple multilayer feed-forward neural network algorithm [54].

2.1.1. Random Forest. Random Forest is widely recognized as a powerful and robust machine learning algorithm suitable for both regression and classification tasks [50]. At its core, it operates as an ensemble learning technique, constructing a multitude of decision trees during training and generating predictions based on the average (for regression tasks) or majority vote (for classification tasks) of the individual trees. This approach enhances predictive accuracy and reduces overfitting by aggregating the predictions from diverse and independently trained decision trees.

Decision trees, the foundation of the Random Forest algorithm, operate by recursively dividing datasets into two subsets, creating a binary tree structure down to the leaf nodes. Each leaf node represents a specific range in the feature space. The creation of decision nodes follows a greedy approach, starting from the root and utilizing the CART algorithm (Classification and Regression Tree). CART is responsible for determining the decision feature and threshold at each node. Throughout this process, CART identifies the feature and threshold that lead to more homogeneous nodes in the case of classification and aims to minimize error in the resulting nodes for regression tasks [55].

In this work, the Random Forest model was implemented using the RandomForestRegressor class from the *scikit-learn* library [56].

2.1.2. Extreme Gradient Boosting (XGBoost). XGBoost is an implementation of the Gradient Boosted Trees regressor, which shares a fundamental principle with the Random Forest algorithm. Like Random Forest, it combines numerous weak-learning trees to create a robust regressor. However, in contrast to Random Forest, where many trees are built on different subsets of the training dataset simultaneously, the Gradient Boosting algorithm adopts a sequential approach. Here, trees are added one after another, and each new tree is trained on the residual errors of its predecessor. This iterative process refines the model by focusing on areas where the previous predictions were less accurate.

In this work, the XGBoost model was implemented using the *XGBoost* library [51]. XGBoost, known for its efficiency and scalability, optimizes the Gradient Boosting technique by incorporating regularization methods and parallel computation, making it a versatile and powerful tool for predictive modeling tasks across diverse datasets.

2.1.3. kNN. The kNN regressor [52] follows a different approach compared to the Gradient Boosted Trees regressor. Instead of relying on an ensemble of decision trees, kNN is a non-parametric method that makes predictions based on the majority vote or average of the k-nearest data points in the feature space.

In the kNN regressor, the prediction for a given data point is determined by examining its proximity to other data points in the training dataset. The ‘k’ in kNN represents the number of nearest neighbors considered in the prediction. These neighbors are identified based on a distance metric, e.g. Euclidean distance, in the feature space.

Unlike the sequential nature of training in gradient boosting, the kNN regressor does not involve a training phase in the traditional sense. Instead, it stores the entire training dataset and makes predictions on new data points by finding the KNN. In this model, ‘k’ serves as a hyperparameter , and its optimal value needs to be fine-tuned.

In this work, KNN regressor model was implemented using KNeighborsRegressor class from the *scikit-learn* library [56].

2.1.4. Ridge regressor. The Ridge Regression model adopts a parametric approach to address multicollinearity and overfitting in linear regression [53]. It introduces a regularization term, often denoted as the L2 norm, to the linear regression objective function (e.g. mean squared error), penalizing large coefficients. This regularization term controls the magnitudes of the coefficients, preventing them from becoming excessively large. During the training phase, the model learns the optimal values for the regression coefficients. Simultaneously, the hyperparameter ‘alpha,’ which determines the strength of the regularization, is tuned. The tuning of ‘alpha’ takes place during both the training and validation phases to strike a balance between fitting the training data well and preventing overfitting. Ridge Regression’s regularization mechanism mitigates the impact of multicollinearity and enhances the model’s generalization performance on new, unseen data.

To implement Ridge Regression in this work, the Ridge class from the *scikit-learn* library [56] was used.

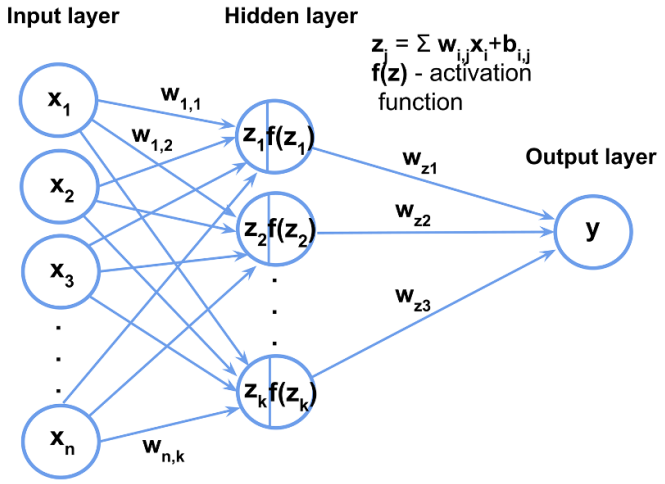

Figure 1. A representation of a simple multilayer perceptron network with one hidden layer.

2.1.5. Multilayer perceptron (MLP). MLP is the most basic architecture of an artificial neural network (ANN) [54]. MLP is a feedforward neural network that consists of multiple layers of nodes organized into an input layer, one or more hidden layers, and an output layer (figure 1). The input layer receives input features forwarded into the network. Hidden layers, positioned between the input and output layers, consist of nodes that utilize mathematical functions to map data. Ultimately, the output layer generates the model's output, providing essential values for tasks such as classification or regression.

MLP is a fully connected network, meaning that each node from one layer connects to all the nodes from the following layer. These connections are characterized by weights. Additionally, hidden and output nodes possess a component known as bias, serving as a threshold to fine-tune predictions by influencing the node output. The goal of the model training is to find the optimal set of these weights and biases. In the hidden layers, data undergoes transformation through summation and activation. Summation involves calculating the sum of products between node outputs from the previous layer and the weights of the connections, along with their corresponding biases. The resulting weighted sum of inputs at each node is then subjected to an activation function. Activation functions play a crucial role in introducing non-linearity to the model. Commonly used activation functions include sigmoid (or logistic activation), tanh (or hyperbolic activation), and Rectified Linear Unit (ReLU). The selection of the activation function depends on the nature of the prediction task. In this work, MLP model was implemented using *PyTorch* python package [57].

2.2. Data acquisition

The critical importance of data for the success of machine learning models cannot be emphasized enough [58]. Firstly, the quantity of data is crucial because it allows the model to identify patterns and relationships within the processed information. An adequate data volume ensures that the model encounters a diverse range of scenarios, enhancing its ability to generalize well to new, unseen data. Essentially, a larger dataset provides a broader perspective, empowering the model to provide informed predictions. Similarly, data quality plays a critical role in training robust and trustworthy machine learning models. Clean, reliable data mitigates the risk of the model learning from noise or biased patterns, which could otherwise compromise its performance.

Ionization mass spectra for chemical compounds with molecular masses up to 300 Da were collected from the NIST WebBook [59] and individual publications [60–62]. The dataset was curated by excluding non-unique spectra and noisy data, resulting in a total of 6550 unique data entries. Each entry in the mass spectrum data was transformed into a vector with a length of 300, where each element represented the peak intensity at $\frac{m}{z}$ values from 1 to 300. To achieve uniformity, each spectrum was padded with zeros to reach the desired length. Additionally, to ensure consistency, all peak intensities were normalized by dividing them by the sum of intensities across all peaks.

2.3. Feature engineering

To prepare input features for machine learning model training, the chemical structures of all compounds in the dataset were acquired in SMILES format [63]. Subsequently, the RDKit open-source tool was used to extract information about atoms and molecules based on the obtained SMILES [64]. This information encompassed various details, such as atom types, valences, degrees, formal charges, aromaticity, hybridization types, and the number of unpaired electrons across bonds within a compound. Moreover, descriptors representing information about individual bonds in compounds were extracted, incorporating details about bond types (single, double, triple), the count of conjugated bonds, the number of bonds forming part of a ring, and information about bond chirality. Additionally, general molecular properties like molecular mass, total number of atoms, and total number of bonds were used among other input features.

The feature values extracted from the chemical structures were normalized using the *StandardScaler* class from sklearn.preprocessing module [56]. This common preprocessing technique ensures that the data distribution has a mean of 0 and a standard deviation of 1, providing a standardized scale for the features.

2.4. Model training and testing

Predicting mass spectra can be approached as a multi-output regression task, where the target variable for each instance is a vector of length 300, which can be represented as $y =$

$[y[1]], y[2], \ldots, y[300]]$. Each element $y[i]$ in the vector corresponds to the peak intensity at a specific mass to charge, $\frac{m}{z}$, value with $1 \leqslant \frac{m}{z} \leqslant 300$. These intensities are treated as individual outputs of the model, with predictions learned directly from the training dataset without relying on any assumptions about the fragmentation process. During inference, for an unknown molecule, the SMILES representation is accepted as input, processed to generate numerical features, and the entire intensity vector is predicted by the trained machine learning model.

In this work, the cosine similarity between normalized real and predicted spectra was used to evaluate the performance of each machine learning model. The cosine similarity score is defined as follows:

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

where:

y_{real} represents the real mass spectrum,
y_{pred} represents the predicted mass spectrum,
i is the index of a peak in the mass spectrum, where $i = \frac{m}{z} - 1$,
$m_{\max}$ is the maximum $\frac{m}{z}$ value with a non-null intensity in the mass spectrum.

The dataset was partitioned into two distinct subsets using random sampling, with 80% of the total data being dedicated to model training and validation, while the remaining 20% was exclusively allocated for testing purposes.

The hyperparameters for Random Forest, XGBoost, kNN and Ridge Regression models were obtained through 5-fold cross validation. In this approach, the dataset is divided into five subsets, or folds. The model is trained five times, each time using four folds for training and the remaining fold for validation. Throughout this iterative process, each fold served as the validation set exactly once, and the final performance metric was averaged over these iterations. This technique ensures robust assessment of model performance while mitigating the risk of overfitting.

2.5. Prediction post-processing to account for different isotopes

The refinement of the final model-generated mass spectra involved a specialized post-processing algorithm to account for different isotopes. This algorithm systematically extracted elemental compositions from predicted fragments by parsing their SMILES representations. Following this, it calculated the probabilities of different isotopic compositions based on known natural abundances.

Subsequently, the algorithm adjusted the intensities of the predicted mass spectra. This adjustment was informed by the calculated probabilities of isotopic variants for each fragment. Higher probabilities were accorded greater weight in the intensity refinement process.

2.6. Total ionization cross sections calculation

The total ionization cross section for C_2F_5 was calculated by using the BEB method [9] implemented in Quantemol-Electron Collisions (QEC) *version* 1.2 [65]. QEC is a user-friendly expert system designed for *ab initio* electron-molecule calculations. QEC is built upon the UKRmol+ suites of codes [66] which are optimized to produce fast and reliable electron-scattering calculations. The code is integrated with the molecular electronic structure code MOLPRO [67] which provided description of the molecular target including its orbitals and symmetry; within QEC target specific input for the BEB calculation is provided by MOLPRO.

According to BEB, the total ionization cross section σ_{BEB} is given by,

$$\sigma_{\text{BEB}} = \frac{S}{t+u+1}\left[\frac{1}{2}\left(1-\frac{1}{t^2}\right)\ln t + 1 - \frac{1}{t} - \frac{\ln t}{1+t}\right], \quad (2)$$

where $t = T/B, u = U/B$, and $S = 4\pi a_0^2 N (R/B)^2$. a_0 is the Bohr radius, and R is the Rydberg energy. B, U, and N are the binding energy, the kinetic energy, and the occupation number, respectively, for the sub-shell; these values are determined at the Hartree–Fock [14], in our case using MOLPRO. If the kinetic energy T of the incident electron is less than B, then $\sigma_{\text{BEB}} = 0$.

While the BEB method generally provides reliable estimates for single ionization cross sections, with typical uncertainties of 10%–15% [9], its accuracy can vary. Discrepancies of up to 25%–30% [68–70] have been reported for certain species and energy ranges. Furthermore, the BEB method does not account for double or higher-order ionization processes, which may lead to underestimations of the total ionization cross section in cases where these processes play a significant role.

2.7. Partial ionization cross sections calculation

Partial ionization cross sections can be estimated by multiplying the BEB total ionization cross section with the respective branching ratios:

$$\sigma_i(T) = \Gamma_i(T) \cdot \sigma_{\text{BEB}}(T) \quad (3)$$

where:

$\sigma_i(T)$ is the partial ionization cross section value at energy T,
$\Gamma_i(T)$ is the branching ratio at energy T, and
$\sigma_{\text{BEB}}(T)$ is the total ionization cross section value at energy T.

Ionization mass spectrometry data can be used to find the branching ratios at a given energy T using the following equation [25, 71]:

$$\Gamma_i(T) = \begin{cases} 0, & T < D_i \\ \Gamma_i\left(T^{\text{ref}}\right)\left[1 - \left(\frac{D_i}{T}\right)^{\gamma}\right], & T \geqslant D_i. \end{cases} \quad (4)$$

Here, $\Gamma_i(T^{\text{ref}})$ is the branching ratio at the reference energy T^{ref}, where $T^{\text{ref}} = 70$ eV, D_i is the appearance threshold of species i due to dissociation of the parent ion (i.e. the lowest energy for which a particular fragment ion can be formed), and γ is a parameter that controls how quickly the asymptotic value of the branching ratio is reached. γ was determined to be 1.5 ± 0.2 by Janev and Reiter [71]. Therefore a value of 1.5 was adopted in this work.

At 70 eV, the branching ratios approach stable values with minimal variation, indicating near energy-independent behavior. This value is widely used as a standard ionization energy in the mass spectrometry community, as it provides sufficient energy to initiate fragmentation in the dominant channels for most molecules.

It is important to note that while the current method provides reliable estimates for partial ionization cross sections by multiplying total cross sections with branching ratios, it does not account for variations in the shapes of energy-dependent cross section curves for different cations compared to their parent cations. These variations can cause shifts in the peaks of cross section curves, which are not accounted for in the present methodology.

Alternatively, partial ionization cross sections can also be calculated using the modified BEB (m-BEB) model, as done in [26, 28]. This method adjusts the binding energies of molecular orbitals to reflect appearance energies and derives partial cross sections by scaling total cross sections with experimentally determined branching ratios. Unlike the current approach, the m-BEB model assumes energy-independent branching ratios, which are normalized to sum to 1 at a reference energy, typically 70 eV. This approach might yield better results for certain species, especially when high-quality experimental branching ratio data are available.

2.8. Appearance thresholds calculations

The appearance threshold of an arbitrary fragment A^+ can be calculated, assuming the dissociation reaction $AB + e^- \rightarrow A^+ + B + e^-$ where B is the lowest-energy isomer, as

$$D_{A^+} = E\left(A^+\right) + E(B) - E\left(AB^+\right) + \text{IP} \qquad (5)$$

where $E(X)$ is the energy of species X and IP is the ionization energy of species AB. *Ab initio* electronic structure calculations were used to obtain the ion fragmentation energies, and the systematic error associated with computation of ionization energies is avoided by using an empirical value for IP.

We have approximated the species energies by calculating their zero-temperature values, optimizing the geometry and computing harmonic vibrational frequencies to obtain a zero-point energy to be added to the electronic energy.

For geometry optimization and vibrational frequencies, spin-unrestricted Kohn–Sham Density-Functional Theory with the PBE functional [72] was used. For the equilibrium geometry energies, coupled-cluster theory with single and double excitations, perturbative inclusion of connected triple excitations, CCSD(T), with a spin-restricted Hartree Fock [73], was used. Calibration studies with systematic sequences of basis sets, and with explicitly-correlated methods, resulted in selection of the CCSD(T)-F12A ansatz [74, 75] using the appropriate triple-zeta-quality basis set, cc-pVTZ-F12 [76].

3. Results and discussion

In our exploration of machine learning models for predicting mass spectra vectors, we evaluated the performance of diverse algorithms through cross-validation, using training and validation subsets.

A crucial step to optimize the predictive capabilities of these models involved hyperparameter tuning, where specific parameters were adjusted for each algorithm. In the case of the Random Forest regressor, the optimal values for parameters such as the number of trees in the forest, the minimum number of samples required to split an internal node, and the minimum number of samples required to be at a leaf node were determined. For the XGBoost regressor, the hyperparameter tuning process focused on finding the optimal values for parameters like the number of boosting rounds, the learning rate (step size), the maximum depth of the trees, etc. Similarly, the kNN regressor underwent hyperparameter tuning, where the optimal number of neighbors to consider for predictions and the choice of distance metric (e.g. Euclidean distance) were fine-tuned. In the case of the Ridge regressor, hyperparameter tuning involved determining the optimal value for the regularization strength (alpha). In the case of the MLP model, the hyperparameter tuning involved identifying the optimal number of hidden layers and nodes, setting the learning rate (lr) for optimization, and determining the ideal number of training epochs. This exploration of optimal parameter configurations ensured that the models were fine-tuned to deliver robust and effective predictions across both the training and validation data subsets.

Table 1 presents a comparative analysis of mean cosine similarity scores achieved by different models, averaged across five distinct train and validation data subsets.

Here, the Base model serves as a benchmark algorithm. It adopts a straightforward approach, generating predictions by averaging the spectrum vectors derived from the entire training dataset. The validation cosine similarity is computed as the average cosine similarity obtained through a meticulous 5-fold cross-validation process.

Notably, both the Random Forest regressor and XGBoost regressor showed considerably high mean cosine similarity scores exceeding 0.97 on the training set. However, their performance on the validation set, although still quite good, exhibited a decline, indicating potential overfitting on training data. The kNN regressor, Ridge regressor, and MLP displayed slightly worse performance, with the MLP achieving the highest similarity on the training set. It is important to highlight that the MLP model underwent a distinct validation

Table 1. Performance comparison of the machine learning models tested on the training and validation data subsets.

	Mean cosine similarity	
Machine learning model	Training	Validation
Base model	0.302	0.301
Random Forest regressor	0.973	0.719
XGBoost regressor	0.968	0.743
kNN regressor	0.773	0.577
Ridge regressor	0.667	0.570
Multilayer perceptron	0.977	0.604

Table 2. Summary of mean and median test cosine similarities for different machine learning models.

ML Model	Mean Test Cosine Similarity	Median Test Cosine Similarity
Base model	0.289	0.220
Random Forest regressor	0.723	0.775
XGBoost regressor	0.753	0.808
Ridge regressor	0.592	0.596
kNN regressor	0.588	0.614
Multilayer perceptron	0.638	0.665
Ensemble model	0.763	0.814
Ensemble model + post-processing	0.847	0.882

process, using a single validation set due to its computational intensity.

Overall, our optimized Random Forest and XGBoost regression models yielded outstandingly good mean cosine similarities between real validation mass spectra and their corresponding predictions.

3.1. Performance on test data

After evaluating the models on the validation data, we proceeded to assess their performance on an independent set of completely unseen test mass spectra. Table 2 summarizes the mean and median values of cosine similarities between real test mass spectra and their corresponding predictions made by various machine learning models.

Overall, the mean and median cosine similarities obtained for the test data closely align with the values averaged across five validation subsets during the cross-validation process. This alignment suggests that all the selected models demonstrate strong generalization abilities.

Figure 2 compares the distributions of cosine similarities produced by different models on the test data with the distribution of test cosine similarities obtained using the Base model. All five machine learning models significantly outperformed the Base model, which yielded a mean cosine similarity of 0.29. As shown in figure 2, the Random Forest regressor and XGBoost regressor algorithms performed exceptionally well on the test data, achieving notably high mean cosine similarities of 0.723 and 0.753, respectively. It is noteworthy that the cosine similarity distributions for these two models significantly deviate from a normal distribution. To ensure a more robust comparison, we also considered the median test cosine similarities. It was found that for more than 50% of test instances, cosine similarities between real and predicted mass spectra were higher than 0.775 for the Random Forest regressor model and 0.808 for the XGBoost regressor model.

The MLP model performed slightly less effectively on the test data, achieving a mean test cosine similarity of 0.638 and a median of 0.665.

Lastly, the kNN regressor resulted in a mean test cosine similarity of 0.588 and a median of 0.614, while the Ridge regression model yielded a mean test cosine similarity of 0.592 and a median cosine similarity of 0.596.

3.2. Ensemble model

To improve prediction accuracy, we assembled an ensemble model by combining three individual models—the Random Forest regressor, XGBoost regressor, and MLP—that consistently provided the best predictions across five validation sets and the test set. In this ensemble model, predictions are generated as the sum of weighted predictions from each constituent model. The optimal weights allocated to the individual models were obtained through a 5-fold cross-validation process, yielding the following weight distribution: 0.7 for XGBoost, 0.2 for Random Forest, and 0.1 for MLP.

In figure 3(A), a comparison is presented between the distributions of individual cosine similarities for the real mass spectra from the test dataset and the predictions obtained using the ensemble model and our top-performing isolated model, the XGBoost regressor. Notably, the ensemble model surpasses the XGBoost regressor with a median cosine similarity of 0.814 and a mean of 0.763 (see table 2).

To further refine the predictive ability of our ensemble model, we implemented a post-processing step aimed at optimizing the accuracy of the generated mass spectra. This post-processing algorithm operates on the outputs of the ensemble model, refining the intensity predictions for each fragment based on a consideration of its elemental composition and the associated probabilities of specific isotopic compositions based on known natural abundances of different isotopes (see Methods section).

Figure 3(B) compares the distribution of test cosine similarities generated by our ensemble model with and without prediction post-processing. Upon adjusting the predicted intensities to account for known natural isotope abundances, the median cosine similarity on the test data significantly increased to 0.882, highlighting a notable improvement compared to the 'naked' ensemble model, which yielded a median cosine similarity of 0.814. Figure 4 presents examples of various predicted spectra alongside their corresponding experimental spectra, which were obtained from the NIST WebBook [59] using electron ionization at 70 eV, for varying values of

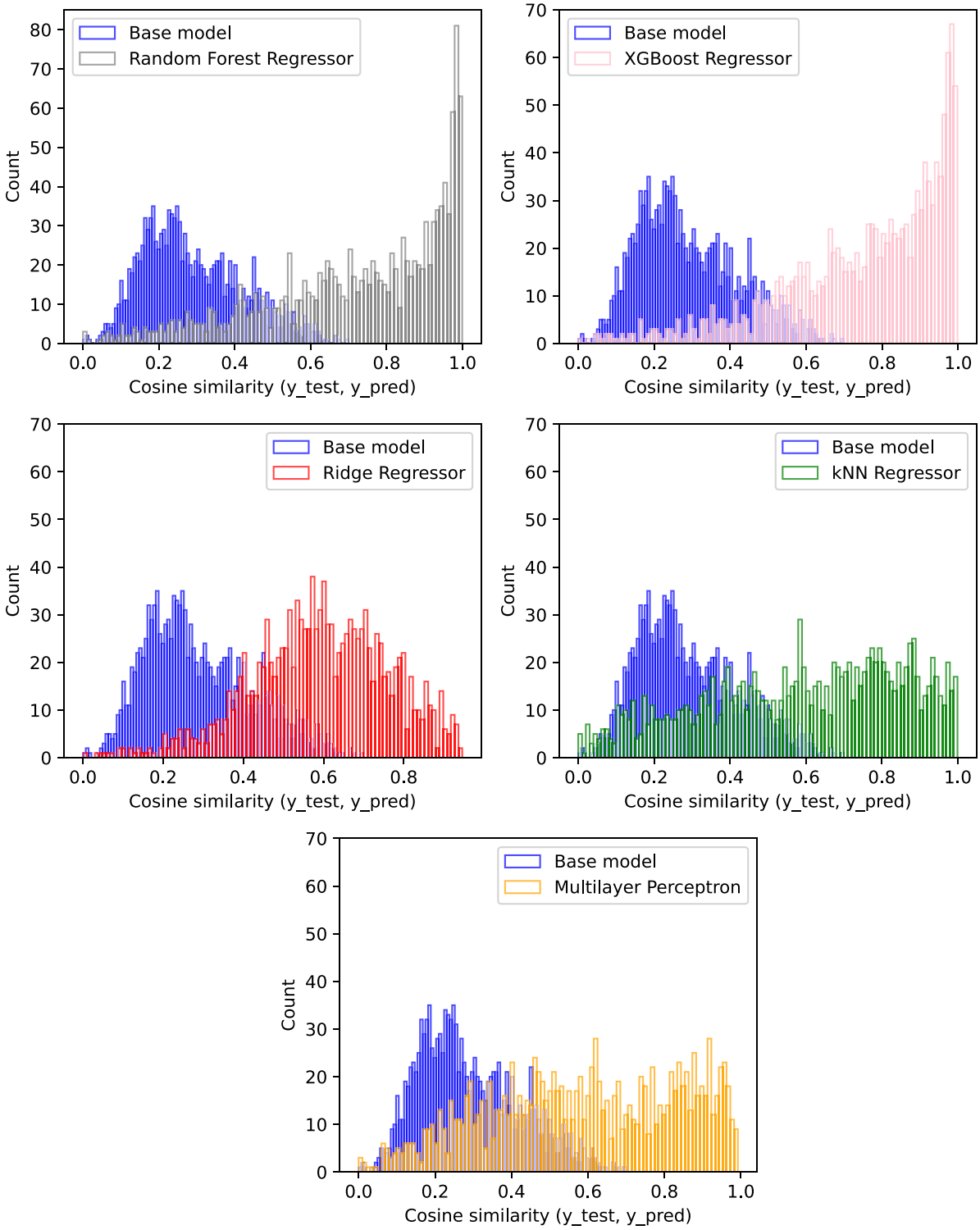

Figure 2. Comparison of distributions of cosine similarities between real and predicted test spectra for different machine learning models.

cosine similarity. Notably, for around 70% of the test cases, the cosine similarity exceeded 0.8. This enhancement emphasizes the efficacy of our post-processing step in refining the agreement between predicted and experimental mass spectra. These findings underscore the substantial improvement achieved by our approach.

3.3. Partial ionization cross sections for NH_3 and C_2F_5

We used our machine learning model's predictions to estimate the partial electron impact ionization cross sections of a well-studied molecule, NH_3, and a radical species, C_2F_5. The predicted mass spectra for both species are displayed in figure 5. In the case of the ammonia molecule, the experimental mass

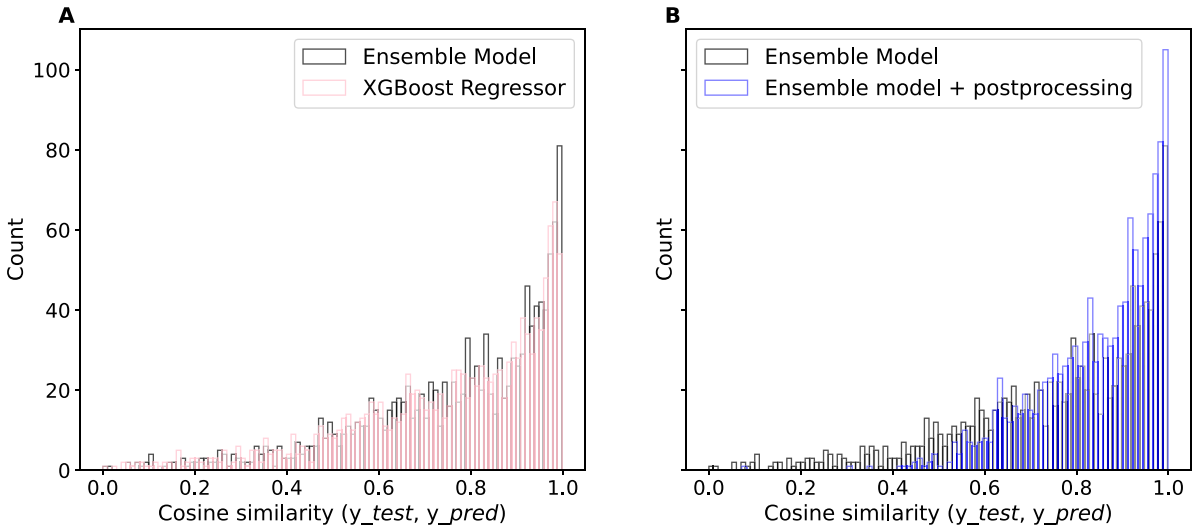

Figure 3. Comparison of distributions of cosine similarities between real and predicted test spectra: (A) distributions of test cosine similarities obtained with the ensemble model vs. XGBoost Regressor. (B) A comparison of test cosine similarity distributions obtained with the ensemble model with and without post-processing.

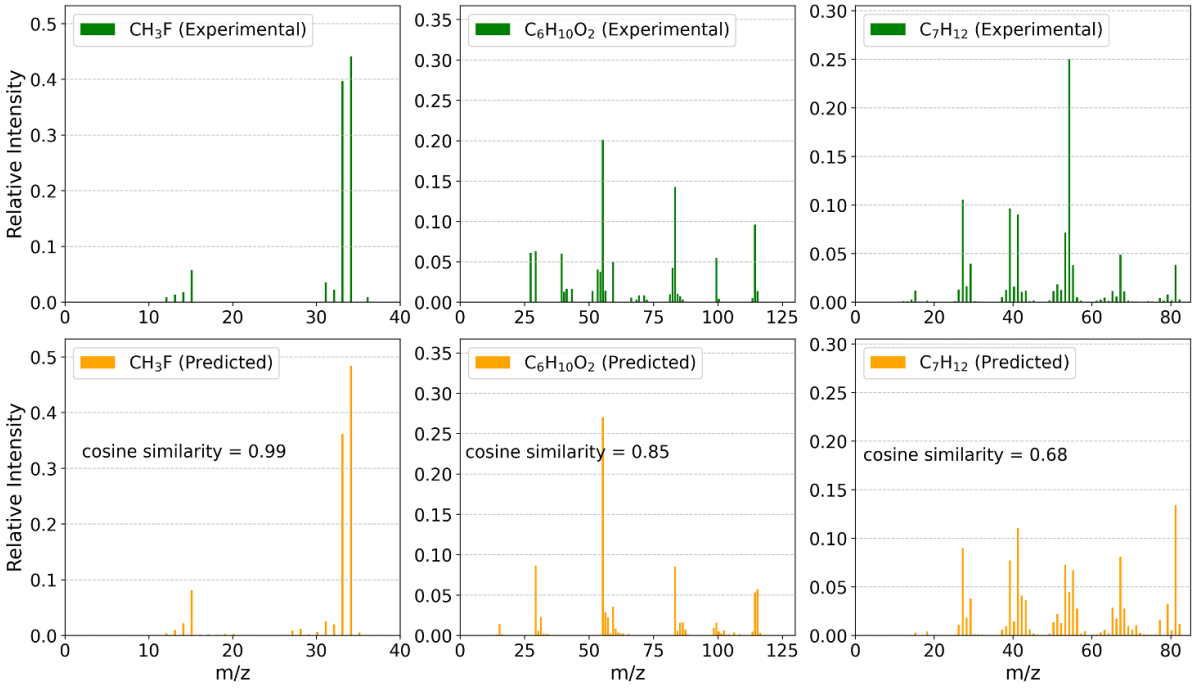

Figure 4. Examples of real and predicted spectra with varying values of cosine similarity. The experimental spectra were obtained from the NIST WebBook [59].

spectrum is available [32]. Figure 6 presents a comparison of the scaled relative intensities of the predicted versus experimental mass spectra. To enable direct comparison, both spectra were normalized by dividing each intensity by the sum of all intensities within the respective spectrum. Additionally, table 3 provides the corresponding relative intensity values for NH_3.

Partial ionization cross sections were determined by multiplying the BEB total ionization cross section by the respective branching ratios. Energy-dependent branching ratios were derived from the branching ratio values at a reference energy and the corresponding fragment appearance thresholds, as described in the Methods section, see equation (4).

It is important to note that while the BEB method is effective for estimating single ionization cross sections, it does not account for double or higher-order ionization processes. Due to this limitation, the scope of this work is restricted to estimating ionization cross sections up to 100 eV; for practical applications in plasma physics the energy region just above threshold (i.e. below 70 eV) is the important one. This energy range

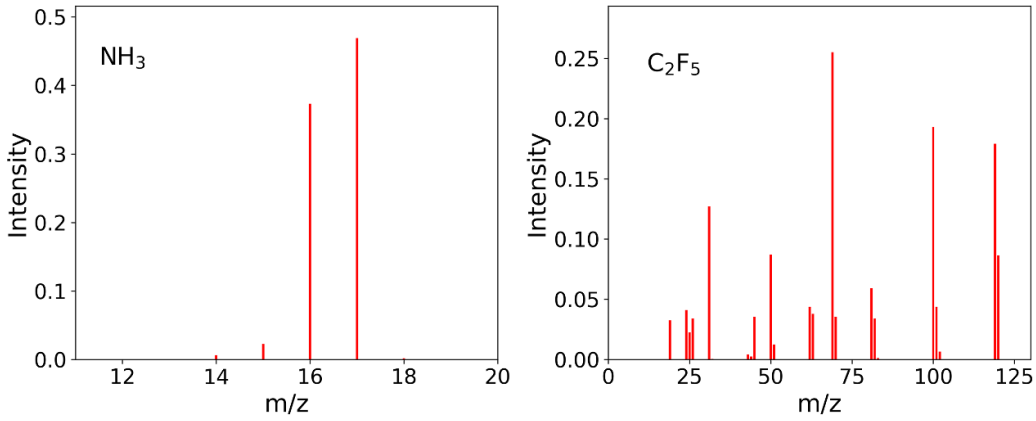

Figure 5. Predicted mass spectra for NH_3 (left) and C_2F_5 (right).

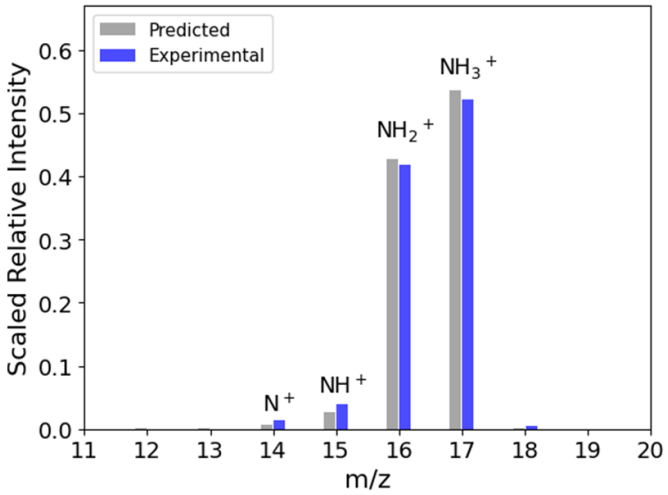

Figure 6. Scaled experimental vs. predicted mass spectra for NH_3.

Table 3. Comparison of relative fragment intensities, branching ratios, and appearance thresholds for NH_3 at 70 eV.

	Relative Intensity			
Fragment	Experimental	Predicted	Branching Ratio	Appearance Threshold (eV)
NH_3^+	0.5125	0.5365	0.538	10.25
NH_2^+	0.4185	0.4269	0.428	15.70
NH^+	0.0393	0.0260	0.026	22.80
N^+	0.0140	0.0076	0.008	26.60
H^+	—	—	—	27.50

minimizes the contribution of double ionization, ensuring the validity of the BEB method for calculating total ionization cross sections and, from these, partial ionization cross sections for the singly ionized fragments.

Table 3 presents the branching ratios at a reference energy of 70 eV for different fragment ions formed from NH_3, as inferred from the mass spectrum predicted by our machine learning model.

Using equation (4), we calculated energy-dependent branching ratios based on the predicted branching ratios at 70 eV and the experimental appearance thresholds for the production of corresponding cations reported in the literature [77]. It is important to note that the predicted experimental mass spectra do not account for H^+, so we were unable to calculate a branching ratio for this fragment due to this limitation. The experimental values of appearance thresholds are shown in table 3. The total ionization cross section for NH_3 was calculated using the BEB method as described in the Methods section. A comparison between the calculated and experimental ionization cross sections for NH_3 is presented in figure 7. Overall, the calculated cross sections are in good agreement with the experimental data.

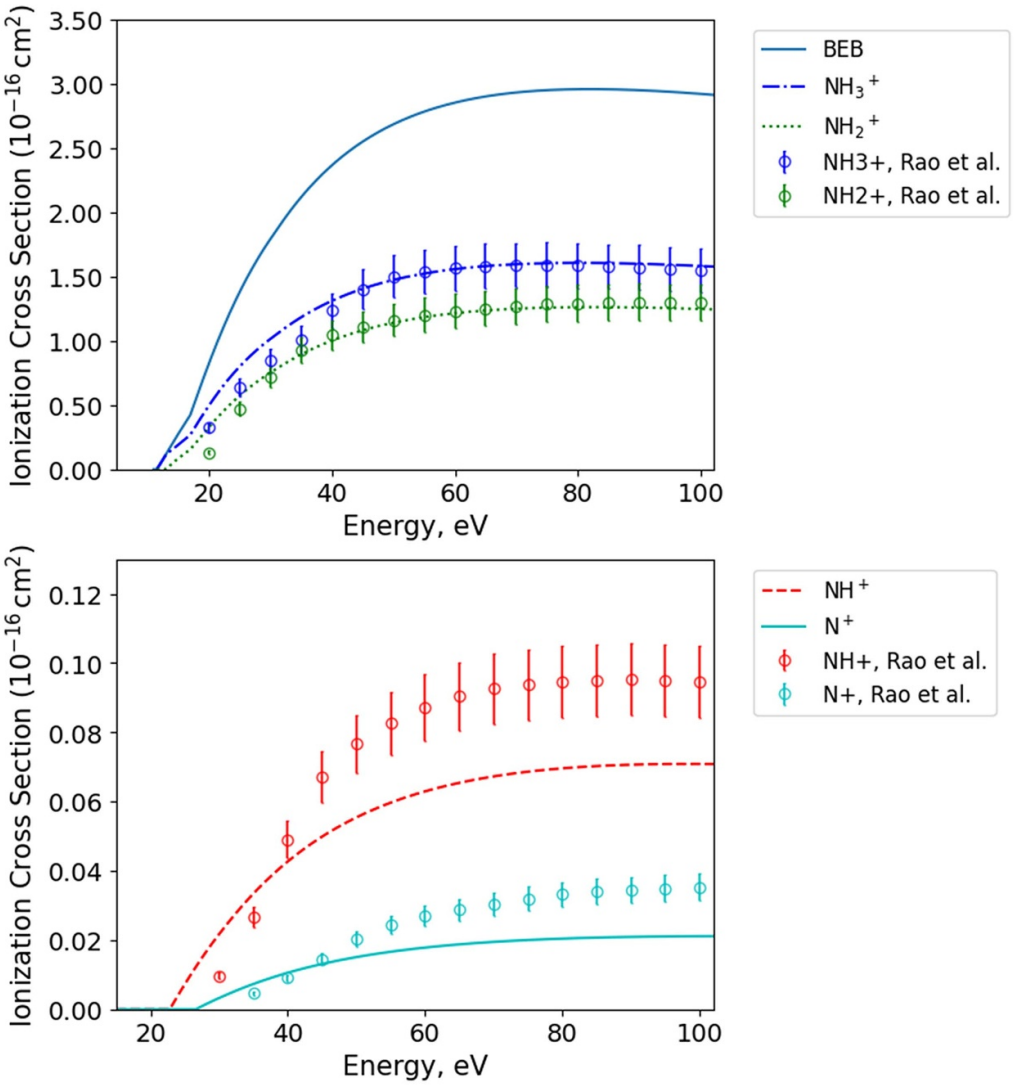

Figure 7. Total and partial electron impact ionization cross sections for NH_3. Solid and dashed curves represent cross sections calculated using machine learning predictions and symbols denote experimental data [77].

The experimental ionization cross section data for NH_3, used for comparison with our theoretical predictions, were obtained from [77]. The experimental setup described in that work, including its resolution and efficiency of ion collection, was designed to ensure accurate and reliable measurements. It featured a time-of-flight mass spectrometer (TOFMS) and a quadrupole mass spectrometer (QMS), both optimized for high-resolution ion detection. The TOFMS provided high temporal resolution by measuring ion time-of-flight, while the QMS selected ions based on their mass-to-charge ratio, minimizing photon interference. Ion collection was achieved using a 100 V cm^{-1} electric field between parallel extraction grids, ensuring efficient capture of ions within the desired energy range. Electron impact energy was precisely controlled with a multichannel analyzer.

Next, we calculated the total and partial ionization cross sections for the C_2F_5 radical. Table 4 shows the branching ratios at a reference energy of 70 eV for fragment ions formed from C_2F_5, derived from the mass spectrum predicted by our machine learning model, along with the calculated fragment appearance thresholds. The fragment appearance threshold energies were calculated as described in the Methods section (see equation (5)) and are based on the following dissociation reactions:

$$C_2F_5^+ \rightarrow CF_3^+ + CF_2 \quad (1),$$
$$C_2F_5^+ \rightarrow CF_2^+ + CF_3 \quad (2),$$
$$C_2F_5^+ \rightarrow CF^+ + CF_4 \quad (3),$$
$$C_2F_5^+ \rightarrow C^+ + CF_3 + F_2 \quad (4),$$
$$C_2F_5^+ \rightarrow F^+ + C_2F_4 \quad (5),$$
$$C_2F_5^+ \rightarrow C_2F_4^+ + F \quad (6),$$
$$C_2F_5^+ \rightarrow C_2F_3^+ + F_2 \quad (7),$$
$$C_2F_5^+ \rightarrow C_2F_2^+ + F_2 + F \quad (8),$$
$$C_2F_5^+ \rightarrow C_2F^+ + 2F_2 \quad (9),$$
$$C_2F_5^+ \rightarrow C_2^+ + 2F_2 + F \quad (10).$$

Table 4. Branching ratios for fragments formed from C_2F_5 at 70 eV predicted by the machine learning model and calculated fragment appearance thresholds.

Fragment	Branching ratio	Appearance threshold (eV)
CF_3^+	0.221	14.84
$C_2F_4^+$	0.169	16.59
$C_2F_5^+$	0.157	12.67
CF^+	0.110	14.59
CF_2^+	0.076	17.20
$C_2F_3^+$	0.052	20.77
$C_2F_2^+$	0.049	24.98
C_2^+	0.044	37.21
F^+	0.042	23.89
C_2F^+	0.040	29.32
C^+	0.039	26.36

Using the calculated fragment appearance thresholds, we computed energy-dependent branching ratios with equation (4). We then obtained the partial ionization cross sections by multiplying the BEB cross section by the respective branching ratios. The calculated total and partial ionization cross sections for C_2F_5 are presented in figure 8 and compared with the available experimental data [78].

The BEB cross section shows reasonable agreement with the experimental data up to approximately 40 eV, after which the calculated values exceed the experimental measurements. This discrepancy aligns with the observations reported by Gupta *et al* [79]. The total ionization cross section reported in [78] was calculated as the sum of the partial cross sections for the two dominant channels, with contributions from all other fragments either ignored or considered negligible. This approach only gives a lower limit experimental values and is one of the reasons our BEB predictions are higher than the observations.

The calculated ionization cross section for $C_2F_5^+$ is consistent with the experimental data. Additionally, our approach accurately predicts the CF_3^+ channel as the most dominant. However, the calculated cross section values for this channel are lower than the experimental values by a factor of 2–3, likely in part due to the significant contributions from other channels predicted by our model. Moreover, it is important to note that the branching ratios used to calculate partial ionization cross sections are derived from machine learning-predicted mass spectra. Machine learning models occasionally overestimate or underestimate target variables due to inherent limitations in the training data and generalization process. These factors may contribute to the observed discrepancies between the predicted and experimental data.

The experimental ionization cross section data for C_2F_5, used to validate our theoretical results, were obtained from [78]. Their setup involved a fast-beam apparatus with a Colutron ion source to generate $C_2F_5{}^+$ ions, which were neutralized via charge transfer in a Xe gas cell and crossed with a well-characterized electron beam (5–200 eV, 0.5 eV FWHM). Fragment and parent ions were separated using an electrostatic hemispherical analyzer and detected by a channel electron multiplier (CEM). Absolute cross sections were calibrated using established methods to ensure reliable and accurate measurements.

4. Conclusions

While total electron impact ionization cross sections hold significant value, understanding partial ionization cross sections is equally important. Despite the availability of several theoretical methods for estimating total ionization cross sections, calculating partial ionization cross sections presents a more challenging task. Partial ionization cross sections can be calculated by multiplying total ionization cross sections with branching ratios for different fragments, which, in turn, can be deduced from ionization mass spectra. Unfortunately, the required mass spectrometry data is frequently unavailable.

In summary, our proposed approach for mass spectra prediction combines three powerful machine learning models—XGBoost, Random Forest, and MLP—in an ensemble model to enhance the accuracy of mass spectra predictions. Through an optimization process involving 5-fold cross-validation, we determined the optimal weights for each constituent model, allocating 0.7 to the XGBoost model, 0.2 to the Random Forest model, and 0.1 to the MLP model. The resulting ensemble model consistently outperformed our top-performing isolated model, the XGBoost regressor.

To further enhance predictive capabilities of our ensemble model, we introduced a post-processing step tailored to correct the model outputs for natural isotope abundances. This refinement process significantly improved the agreement between predicted and experimental mass spectra. Notably, our enhanced model achieved a median cosine similarity of 0.882 on the test data, surpassing the ensemble model without post-processing, which yielded a median cosine similarity of 0.814. Furthermore, the post-processed ensemble model demonstrated substantial performance across a majority of test cases. In around 70% of instances, the similarity between real and predicted spectra exceeded 80%. These results collectively demonstrate the robustness and efficacy of our approach

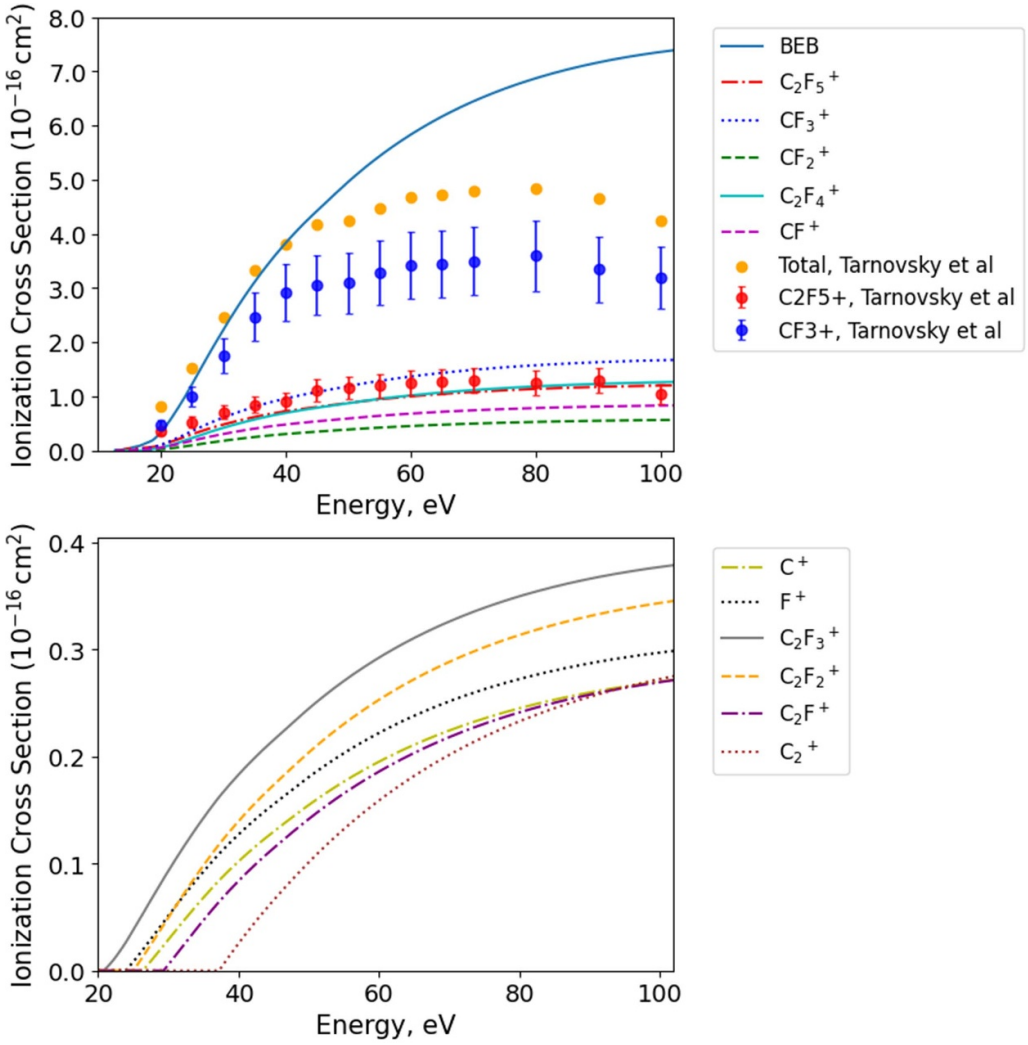

Figure 8. Total and partial electron impact ionization cross sections for C_2F_5. The solid and dashed curves represent cross sections calculated using machine learning model predictions; the symbols correspond to experimental data [78].

in improving the accuracy of mass spectra predictions by combining the strengths of various models in the ensemble and optimizing predictions through targeted post-processing.

We demonstrate that our machine learning predictions can be used to estimate partial ionization cross sections, as exemplified by our calculations for both the well-studied NH_3 molecule and the more challenging C_2F_5 radical. However, it is important to highlight key limitations: the approach is valid primarily for electron impact energies below 100 eV, as the BEB model does not account for double ionization processes, which become significant at higher energies. Additionally, the machine learning predictions are limited by the absence of light fragments such as H^+ in the training data, resulting in zero intensities for their peaks. This must be considered when using the predicted mass spectra for branching ratio calculations.

It is worth noting that although BEB total ionization cross sections are used in this work, our machine learning-based fragmentation model is not inherently reliant on the BEB model. The method is versatile and can be applied to any set of total ionization cross sections, including experimental measurements.

Our plan is to use this methodology to provide electron impact fragmentation cross sections which will be placed in the Quantemol Data Base (QDB) [80] and, in due course, an updated version of QEC which will provide fragmentation patterns alongside the BEB total ionization cross sections which are already provided by QEC.

Finally, although the focus of this work is on the providing important electron impact fragmentation cross sections for use in modelling, our machine learning algorithm actually learns mass spectroscopy fragmentation patterns. This means that it can be used as a means to predict such patterns for molecules yet to be studied using mass spectroscopy.

Data availability statement

The cross section data generated for in this study are freely available from the Quantemol Data Base (QDB) [80].

The data that support the findings of this study are available via the following URL/DOI: www.quantemoldb.com/.

Acknowledgment

Development of QEC was supported by STFC Grant ST/R005133/1.

Conflict of interest

K M L, G A, S M and A N work for Quantemol Ltd and J T is a Director of Quantemol Ltd; Quantemol are interested in utilizing the results of this study. P J K declares no conflict of interest.

ORCID iDs

Kateryna M Lemishko https://orcid.org/0000-0003-1659-2920
Gregory S J Armstrong https://orcid.org/0000-0001-5949-2626
Anna Nelson https://orcid.org/0009-0001-7741-5488
Jonathan Tennyson https://orcid.org/0000-0002-4994-5238
Peter J Knowles https://orcid.org/0000-0003-4657-6331

References

[1] Kim H-T, Lim J-S, Kim M-S, Oh H-J, Ko D-H, Kim G-D, Shin W-G and Park J-G 2015 *Microelectron. Eng.* **135** 17
[2] Asundi R K and Craggs J D 1964 *Proc. Phys. Soc.* **83** 611
[3] Montague R G, Harrison M F A and Smith A C H 1984 *J. Phys. B: At. Mol. Phys.* **17** 3295
[4] Munjal H and Baluja K L 2007 *J. Phys. B: At. Mol. Phys.* **40** 1713
[5] Irikura K K 2017 *J. Res. Natl Inst. Stand. Technol.* **122** 28
[6] Itikawa Y 2017 *J. Phys. Chem. Ref. Data* **46** 043103
[7] Radoiu M and Hussain S 2009 *J. Hazard. Mater.* **164** 39
[8] Shih M, Lee W-J, Tsai C-H, Tsai P-J and Chen C-Y 2002 *J. Air Waste Manage. Assoc.* **52** 1274
[9] Kim Y-K and Rudd M E 1994 *Phys. Rev.* A **50** 3954
[10] Deutsch H, Becker K, Matt S and Mark T D 2000 *Int. J. Mass Spectrom. Ion Process.* **197** 37–69
[11] Deutsch H, Becker K and Mark T D 2000 *Eur. Phys. J.* D **12** 283–7
[12] Joshipura K N and Antony B K 2001 *Phys. Lett.* A **289** 323
[13] Zhou W, Wilkinson L, Lee J W L, Heathcote D and Vallance C 2019 *Mol. Phys.* **117** 3066–75
[14] Graves V, Cooper B and Tennyson J 2021 *J. Chem. Phys.* **154** 114104
[15] Gorfinkiel J D and Tennyson J 2004 *J. Phys. B: At. Mol. Phys.* **37** L343–50
[16] Gorfinkiel J D and Tennyson J 2005 *J. Phys. B: At. Mol. Phys.* **38** 1607–22
[17] Colgan J and Pindzola M S 2012 *Eur. Phys. J.* D **66** 284
[18] Ali E and Madison D 2019 *Phys. Rev.* A **100** 012712
[19] Scarlett L H, Jong E, Odelia S, Zammit M C, Ralchenko Y, Schneider B I, Bray I and Fursa V D 2023 *At. Data Nucl. Data Tables* **151** 101573
[20] Randazzo J M, Marante C, Chattopadhyay S, Schneider B I, Olsen J and Argenti L 2023 *Phys. Rev. Res.* **5** 043115
[21] Falkowski A G, Bettega M H F and Lima M A P 2024 *Phys. Rev.* A **110** 022808
[22] Ásgeirsson V, Bauer C A and Grimme S 2017 *Chem. Sci.* **8** 4879–95
[23] Huber S E, Mauracher A, Süß D, Sukuba I, Urban J, Borodin D and Probst M 2019 *J. Chem. Phys.* **150** 024306
[24] Li C, Chin C-H, Zhu T and Hui Zhang J Z 2020 *J. Mol. Struct.* **1217** 128410
[25] Graves V, Cooper B and Tennyson J 2021 *J. Phys. B: At. Mol. Phys.* **54** 235203
[26] Goswami K, Luthra M, Bharadvaja A and Baluja K L 2022 *Atoms* **10** 101
[27] Hamilton J R, Tennyson J, Huang S and Kushner M J 2017 *Plasma Sources Sci. Technol.* **26** 065010
[28] Goswami K, Arora A K, Bharadvaja A and Baluja K L 2021 *Eur. Phys. J.* D **75** 228
[29] Shanmugasundaram S, Agrawal R and Gupta D 2024 *J. Chem. Phys.* **160** 094310
[30] Ellis-Gibbings L K, Fortune W G, Cooper B, Tennyson J and Price S D 2021 *Phys. Chem. Chem. Phys.* **23** 11424
[31] Ellis-Gibbings L K, Cooper B, Tennyson J and Price S D 2022 *J. Phys. B: At. Mol. Phys.* **55** 124001
[32] Wallace W E 2016 NIST mass spectrometry data center *Mass Spectra NIST Chemistry WebBook, NIST Standard Reference Database Number* vol 69, ed P J Linstrom and W G Mallard (National Institute of Standards and Technology) p 20899
[33] Rupp M, Tkatchenko A, Müller K-R and von Lilienfeld O A 2012 *Phys. Rev. Lett.* **108** 058301
[34] Faber F A, Hutchison L, Huang B, Gilmer J, Schoenholz S S, Dahl G E, Vinyals O, Kearnes S, Riley P F and von Lilienfeld O A 2017 *J. Chem. Theory Comput.* **13** 5255–64
[35] Pereira F and Aires-de Sousa J 2018 *J. Cheminform.* **10** 43
[36] Bleiziffer P, Schaller K and Riniker S 2018 *J. Chem. Inf. Model.* **58** 579–90
[37] Li Z, Kermode J R and De Vita A 2015 *Phys. Rev. Lett.* **114** 096405
[38] Zhang Z-Y, Peng D, Liu L, Shen L and Fang W-H 2023 *J. Phys. Chem. Lett.* **14** 1877–84
[39] Liu Y and Li Z 2023 *J. Chem. Inf. Model.* **63** 806–14
[40] Kawaguchi S, Takahashi K, Ohkama H and Satoh K 2020 *Plasma Sources Sci. Technol.* **29** 025021
[41] Krüger F, Gergs T and Trieschmann J 2019 *Plasma Sources Sci. Technol.* **28** 035002
[42] Salimian A, Haine E, Pardo-Sanchez C, Hasnath A and Upadhyaya H 2022 *Coatings* **12** 953
[43] Kim B, Im S and Yoo G 2021 *Electronics* **10** 49
[44] Rietman E and Lory E 1993 *IEEE Trans. Semicond. Manuf.* **6** 343–7
[45] Stokes P W, Cocks D G, Brunger M J and White R D 2020 *Plasma Sources Sci. Technol.* **29** 055009
[46] Zhong L 2019 *J. Appl. Phys.* **125** 183302
[47] Harris A L and Nepomuceno J 2023 preprint
[48] Hanicinec M, Mohr S and Tennyson J 2023 *J. Phys. D: Appl. Phys.* **56** 374001
[49] Zhang B, Zhang J, Xia Y, Chen P and Wang B 2022 *Int. J. Mass Spectrom.* **475** 116817
[50] Ho T K 1995 *Proc. 3rd Int. Conf. on Document Analysis and Recognition* vol 1 pp 278–82
[51] Chen T and Guestrin C 2016 *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* pp 785–94
[52] Cover T M and Hart P E 1967 *IEEE Trans. Inf. Theory* **13** 21–27
[53] Hoerl A E and Kennard R W 1970 *Technometrics* **12** 55–67
[54] Haykin S S 1994 *Neural Networks: A Comprehensive Foundation* (Prentice Hall PTR)
[55] Wu X *et al* 2008 *Knowl. Inf. Syst.* **14** 1–37
[56] Pedregosa F *et al* 2011 *J. Mach. Learn. Res.* **12** 2825–30

[57] Paszke A, Gross S, Massa F, Lerer A, Bradbury J and Chanan G 2019 *Advances in Neural Information Processing Systems* 32
[58] Halevy A, Norvig P and Pereira F 2009 *IEEE Intell. Syst.* **24** 8–12
[59] Linstrom P J and Mallard W G 2001 *J. Chem. Eng. Data* **46** 1059–63
[60] Westmore J B and Fisher K J W G D 1999 *Int. J. Mass Spectrom.* **182–183** 53–61
[61] De Ridder J J and Dijkstra G 1968 *Org. Mass Spectrom.* **1** 647–57
[62] Lewis J and Johnson B F G 1968 *Acc. Chem. Res.* **1** 245–56
[63] Weininger D 1988 *J. Chem. Inf. Comput. Sci.* **28** 31–36
[64] Rdkit: Open-source cheminformatics (available at: www.rdkit.org)
[65] Cooper B *et al* 2019 *Atoms* **97** 7
[66] Mašín Z, Benda J, Gorfinkiel J D, Harvey A G and Tennyson J 2019 *Comput. Sci. Commun.* **249** 107092
[67] Werner H-J, Knowles P J, Manby F R, Black J A, Doll K, Heßelmann A, Kats D and Köhn A 2020 *J. Chem. Phys.* **152** 144107
[68] Ali M A, Kim Y-K, Hwang W, Weinberger N M and Rudd M E 1997 *J. Chem. Phys.* **106** 9602
[69] Hwang W, Kim Y-K and Rudd M E 1996 *J. Chem. Phys.* **104** 2956
[70] Nishimura H, Huo W M, Ali M A and Kim Y-K 1999 *J. Chem. Phys.* **110** 1234–40
[71] Janev R K and Reiter D 2004 *Plasmas* **11** 780–829
[72] Perdew J P, Burke K and Ernzerhof M 1996 *Phys. Rev. Lett.* **77** 3865–8
[73] Knowles P J, Hampel C and Werner H-J 1993 *J. Chem. Phys.* **99** 5219–27
[74] Adler T B, Knizia G and Werner H-J 2007 *J. Chem. Phys.* **127** 221106
[75] Knizia G, Adler T B and Werner H-J 2009 *J. Chem. Phys.* **130** 54104
[76] Peterson K A, Adler T B and Werner H-J 2008 *J. Chem. Phys.* **128** 084102
[77] Rao M V V S and Srivastava S K 1992 *J. Phys. B: At. Mol. Phys.* **25** 2175–87
[78] Tarnovsky V, Deutsch H and Becker K 1999 *J. Phys. B: At. Mol. Phys.* **31** L573–6
[79] Gupta D, Choi H, Song M Y, Karwasz G P and Yoon J S 2017 *Eur. Phys. J. D.* **71** 88
[80] Tennyson J *et al* 2022 *Plasma Sources Sci. Technol.* **31** 095020