Journal of Electric Propulsion

RESEARCH Open Access

AI-assisted advanced propellant development for electric propulsion

Angel Pan Du[1], Miguel Arana-Catania[2*] and Enric Grustan-Gutiérrez[1,3*]

*Correspondence:
Miguel Arana-Catania
humd0244@ox.ac.uk
Enric Grustan-Gutiérrez
E.Grustan@cranfield.ac.uk
[1]Cranfield University, College Road, Cranfield MK43 0AL, UK
[2]University of Oxford, Broad Street, Oxford OX1 3BG, UK
[3]Ignition Space, C/ Còrsega 384 6-1, Barcelona 08037, Spain

Abstract

Artificial Intelligence algorithms are introduced in this work as a tool to predict the performance of new chemical compounds as alternative propellants for electric propulsion, focusing on predicting their ionisation characteristics and fragmentation patterns. The chemical properties and structure of the compounds are encoded using a chemical fingerprint, and the training datasets are extracted from the NIST WebBook. The AI-predicted ionisation energy and minimum appearance energy have a mean relative error of 6.87% and 7.99%, respectively, and a predicted ion mass with a 23.89% relative error. In the cases of full mass spectra due to electron ionisation, the predictions have a cosine similarity of 0.6395 and align with the top 10 most similar mass spectra in 78% of instances within a 30 Da range.

Keywords Mass spectrum, Ionisation energy, Appearance energy, Machine learning, Neural networks, Multilayer perceptron, Electric thrusters

Introduction

Despite the great potential of electric propulsion, the reliance on xenon (Xe) as the primary propellant presents challenges due to its rarity and escalating costs [1]. As a consequence, a series of alternatives have been sought. The most direct approach has been using other noble gases such as krypton or argon; however, they present worse ionisation characteristics [2–6], and in the case of krypton also a rapid cost increase [7]. Other atomic alternatives have also been explored in the p-block of the periodic table, such as bismuth or iodine, and while they can have excellent performance, they have condensation issues and, in the case of iodine, potential losses due to its molecular nature [8–10]. In light of these limitations, there is a growing interest in exploring alternative molecular compounds that can effectively substitute xenon and other chemical elements as a propellant for electric propulsion (EP) while remaining stable EP applications [11]. Adamantane ($C_{10}H_{16}$) and buckminsterfullerene (C_{60}) are promising candidates, but not without drawbacks. Adamantane, along with iodine, has potential compatibility issues, such as spacecraft contamination and toxicity [3, 4] and (C_{60}) temperature stability issues; elevated temperatures result in the fragmentation of the molecule, while reduced temperatures lead to resublimation onto the engine internal surfaces [12, 13].

While the highlighted candidates present some significant issues, the list of potential molecular propellants is practically limitless. However, the virtual infinity of the candidate list and the complexity of their characteristics make a trial-and-error approach impractical, especially when facing a new component without information on its characteristics (eg, ionisation energy or molecular stability). For this reason, other methodologies are currently under exploration, such as approaches that blend empirical experimentation with computational analysis [14–24]. Moreover, quantum chemistry calculations serve as the predominant method for accurately computing various fragmentation parameters. These calculations typically require inputs related to the molecular structure and electronic configuration of the compounds under investigation [25–27]. Nonetheless, these strategies often necessitate significant temporal and computational investments, accompanied by the challenge of potential inaccuracies due to reliance on individual expertise.

To overcome similar or more complex challenges, machine learning (ML) has become a mainstay of cheminformatics, especially for drug discovery [28]. Therefore, the implementation of ML techniques holds promise for efficiently and accurately identifying ideal EP propellant candidates. By leveraging ML techniques, this project aims to streamline the selection process, reduce reliance on expensive propellants, identify viable alternatives that offer comparable performance characteristics, and kickstart the development of mission-tailored propellants by providing a tool capable of predicting the behaviour of novel molecular compounds when used as EP propellants even when the only information available is the compound structure.

Methodology

Relevant physical parameters

When facing a potential brand-new propellant, a wide array of properties can be of interest, such as density, toxicity, corrosiveness, or physical state at ambient temperature. However, this study focuses on properties relevant to the ionisation of molecular compounds.

Probably some of the obvious parameters affecting the performance of EP systems are the ionisation energy and molecular mass of the propellant [29]. To minimise the energy required for generating a high-density plasma, an ideal propellant should possess a low ionisation energy and a high ionisation cross-section [30]. The ionisation energy represents the minimum energy necessary to remove an electron from an atom or molecule, transforming it into a charged ion. Conversely, the ionisation cross-section quantifies the likelihood of an ionising collision occurring between a charged and neutral particle [29]. And while the current power levels available to EP promote heavy, easily ionised propellants (such as Xe), which deliver higher momentum, at the cost of a smaller specific impulse, the optimal molecular mass of the propellant is a new spin on the classical problem of Isp optimisation [31], depending on the mission Δv and power system mass efficiency.

Unlike atomic propellants, the possibility of splitting a molecule makes other parameters relevant. One of them is the minimum appearance energy, which represents the minimum energy required for a molecule to ionise through fragmentation. A priori, the preferred attribute for appearance energy is high, as it implies greater molecular stability, reducing the likelihood of decomposition upon ionisation. However, in some cases,

if the appearance energy is too low, the propellant molecules might disintegrate prematurely during ionisation, resulting in inefficient thrust production and potential harm to the propulsion system [12].

In addition to the stability assessment, it is critical to assess how the molecules will fragment. The ion prediction from minimum appearance energy is the expected mass of the ion resulting from the molecule fragmentation at that energy, thus giving an idea of how the molecule fragments and, consequently, the potential losses due to unused mass or low polydispersive efficiency (acceleration losses due to a diverse ion mass or specific charge population). So using Buckminsterfullerene, C_{60}, as an example, at 7.8 eV the whole molecule will ionise and at an appearance energy of 20.2 eV the ion C_{58}^{+} will appear (as will C_2) due to the molecule breaking, the mass of the ion in this case will be 696.62 Da [32]. If the ionisation energy increases, smaller ions will appear.

For these reasons, Ionisation Energy (IE), minimum appearance energy (AE), and ion prediction from the minimum AE fragmentation are the chosen parameters to test the ML algorithms for studying molecular EP propellants. These three general parameters, plus the molecular mass, provide a general idea of the quality of a compound as a propellant. As a first approximation, the propellant should be easy to ionise (low IE), stable (high AE), and in the case of creating ions also via fragmentation, the ion mass should be close to the original one (higher polydispersive efficiency).

While the previous parameters offer a general idea of the quality of a propellant for plasma thrusters, the ionisation and fragmentation phenomena are more complex, creating a distribution of charged ions. Similarly to the case of a high-power Hall Effect Thruster (HET) with multiply ionised ions, a range of specific charges in the plasma beam can result in losses that need to be accounted for [33], in particular if the specific charges are very dissimilar [34]. Mass spectrometry analysis provides valuable insight into the fragmentation patterns, which can help assess the stability and performance of potential propellants and can be used to evaluate losses due to the distribution of specific charges in the plasma beam (polydispersive efficiency) [35]. The ideal case is a monodisperse beam; failing this, lowering the coefficient of variation of the specific charge improves the polydispersive efficiency [35]. In this case, mass spectra offer the complete fragmentation profile for a given ionisation energy. Particularly, a focus is placed on electron ionisation (EI) mass spectrum (MS), due to its widespread use in the field of EP, providing valuable data for predicting the fragmentation patterns of chemical compounds [3, 10, 12].

Additionally, during the selection of alternative propellants for electric propulsion, other critical parameters should be taken into account, such as chemical compatibility with both the system [36, 37] and the thruster acceleration mechanism [29], as well as scalability with power [38]. However, these parameters are system-specific and depend on the trade-offs between propellant performance and its system engineering impacts, and while ML can also be used to optimise a whole system and identify other characteristics such as toxicity, we chose to focus on the prediction of the total mass spectra to showcase the potential of ML to help predict more complex (and system-specific) problems.

Molecular structure encoding and data

As discussed, molecule fragmentation during ionisation is a significant part of the study. Consequently, inputs related to the molecular structure and electronic configuration of the compounds under investigation are needed. Although different methods can be used to provide this information, such as bond characteristics, electronic configuration, harmonic vibrational data, or mass spectrum [21, 39–43], the molecular fingerprint is the more direct and flexible method to provide input to the ML algorithms as it does not require of extra computational or experimental steps [18, 28, 44–47]. The molecular fingerprint is a multidimensional vector containing the molecule's atomic elements and structure.

For this simplicity and directness, fingerprints are chosen as the primary input of the models, in particular, the (ECFPs), a form of molecular representation widely used in computational chemistry. They are generated through an algorithm known as the Morgan algorithm [48], which is rooted in graph theory and captures the structural features of chemical compounds. ECFPs are binary strings that encode the presence or absence of specific substructures within a molecule based on the neighbouring atoms and bonds within a defined circular radius.

The generation of these fingerprints was accomplished using the *RDKit Cheminformatics* package,[1] with a chosen fingerprint length of 4096 and a radius of 2. These specific values were chosen based on considerations of both efficiency and representational capacity.

A longer fingerprint length enhances the depiction of molecular structure by accommodating a wider array of distinctive substructures at the cost of computational requirements and memory consumption. A larger radius captures more atoms and bonds, enriching the portrayal of molecular interactions and correlations. Nonetheless, an excessive radius risks including extraneous information, potentially introducing noise to the fingerprint representation.

Furthermore, an exploration of additional inputs, as well as some adjustments to these, was undertaken to potentially enhance the models' performance. As an example, the inclusion of the mass data from the chemical compounds and the adjustment of fingerprint lengths and radii were among the investigated modifications. The efficacy of these additional inputs will be elaborated upon and scrutinised in the subsequent results section.

As is usual in ML, the choice of the dataset to train the algorithms is critical. Thanks to the ready availability of some of the aforementioned data in the National Institute of Standards and Technology (NIST) Chemistry WebBook,[2] along with its widespread use in literature, makes it an ideal choice for this study. Chemical data can often be an expensive endeavour. However, the NIST Chemistry WebBook offers a reduced version accessible to everyone, encompassing up to 72,618 compounds at present.

Ultimately, data scrubbing from the NIST database yielded mass spectra data from 21,142 distinct compounds, encompassing all available content regarding this parameter. In addition, ionisation energy data from 3,073 compounds and minimum appearance energy data from 2,148 compounds were successfully obtained. The primary drawback of open-access data collected in this database is its emphasis on relatively low-mass

[1] https://www.rdkit.org/.

[2] https://webbook.nist.gov/chemistry/.

chemical compounds (50–200 Da). Further details on the molecular weight distribution for the three databases are available in Fig. 1. Moreover, the ranges analysed for the IE and AE are 3.89–24.59 eV and 5.06–36.00 eV, respectively, and for the ion mass derived from the AE is 1–698 Da. For better visualisation, the distributions of these parameters in their datasets are represented in Fig. 2

Model architectures and training details

For this project, the chosen ML model to produce the predictions is based on neural networks, specifically utilising a (MLP) structure for all prediction scenarios, and additionally using long short-term memory (LSTM) and bidirectional long short-term memory (Bi-LSTM) networks for the case of mass spectrum prediction. The decision to use an MLP is grounded in its successful implementation in reference papers for mass spectrum prediction [47, 49], ensuring comparability with prior research and establishing a reliable and consistent approach in evaluating the current ML-based assessment system.

Nevertheless, there are critical differences between the present work and these references. Although previous research has focused on the use of machine learning models to predict mass spectra, this article extends the study further by incorporating additional predictive tasks, such as IE and minimum AE energies, thus expanding the scope of the analysis. Moreover, unlike the previously discussed research that utilised a sub-database available only for the licensed version of NIST, the current study relies exclusively on the open-access version. Therefore, straightforward comparisons cannot be made in terms

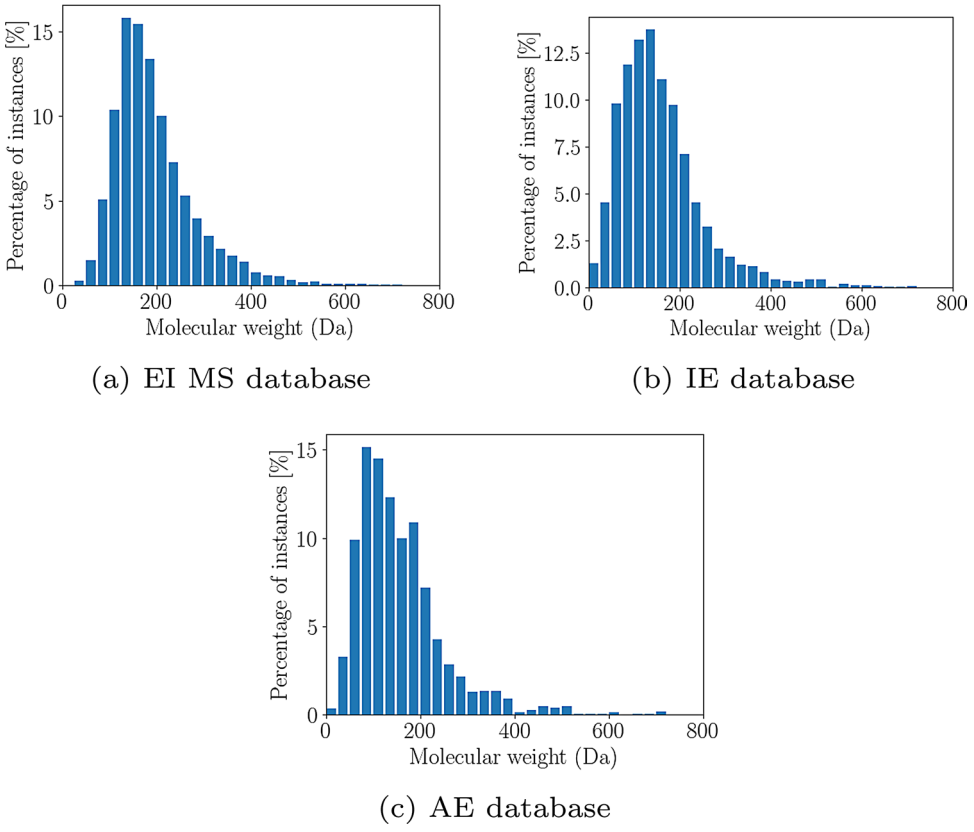

Fig. 1 Distribution of molecular weights of chemical compounds in the three databases. The histograms indicate the percentage of compounds in specified ranges of molecular weight. (**a**) EI MS database. (**b**) IE database. (**c**) AE database

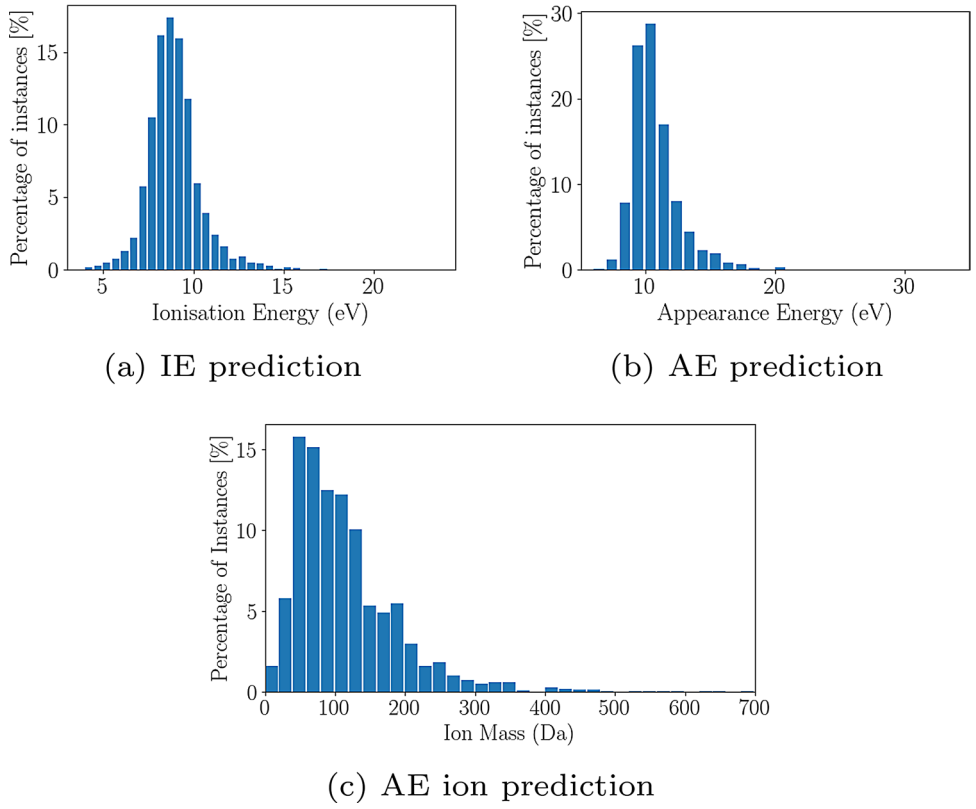

Fig. 2 Distribution of the prediction parameters in the datasets. (**a**) IE prediction. (**b**)AE prediction. (**c**) AE ion prediction

of identical evaluation metrics. However, a similar methodological configuration exists with regard to measuring model performance, as discussed in subsequent sections.

Regarding the aforementioned ML models, firstly, a multilayer perceptron [50] is a neural network composed of layers of neurons where each neuron of a layer is connected to all neurons of the following layer. The neurons represent a linear combination of the outputs provided by the connections of the previous layer followed by a non-linear function.

A long short-term memory [51] is a type of recurrent neural network. Unlike the MLP, which processes each input data point independently to produce an individual output, the latter keeps an internal memory of the processing of each input which is used for processing the following one. This allows it to relate them and thus deal with sequential data. For example, when processing a sequence of atoms, for each atom, it uses the information of all previous atoms in the sequence. The LSTM has a complex internal structure, combining multiple types of linear and non-linear functions in internal operations called gates, which allow the network to learn how to manage the internal memory for each specific task.

A Bi-LSTM [52] is a recurrent neural network which combines two LSTM networks. Each of them processes the input sequence in a different direction, starting from a different end of the sequence. This design is appropriate for sequences where each element is related to both previous and posterior elements of the sequence.

The division of the dataset was performed using a 90–5–5 randomised split, where 90% was allocated to the training set, and 5% to the validation set and to the test set,

as suggested in [47, 49]. The training set is utilised to fit the model, the validation set is employed to fine-tune hyperparameters and prevent overfitting, and the test set is reserved for evaluating the final model's performance on unseen data. For this reason, performance metrics such as loss and relative error are meant to differ between the test subset, and training and validation. Furthermore, each model training is different, meaning that the loss landscape is defined specifically by the data provided [53]. Hence, divergent behaviors can be found within the training losses from distinct predictions.

The loss function employed was SmoothL1Loss. Tuning of hyperparameters was conducted for the present studies considering various activation functions, such as leaky (ReLU), tanh, and sigmoid, different optimisers (e.g., stochastic gradient descent (SGD), Adagrad, Adadelta, and RMSprop) [54], as well as other relevant parameters for the training. The alternative proposed activation functions and optimisers were selected based on their common usage and effectiveness in machine learning literature [55, 56], and are further discussed in the corresponding subsections from their respective prediction.

Ionisation energy prediction

Like in all the following predictions, molecular fingerprints were employed as input data for the model. This is due to the accuracy of the molecular fingerprint in representing the compound's structure, a factor commonly used in quantum chemical calculations for ionisation energy determination. Additionally, insights from [46], focusing on ionisation energy prediction for volatile organic compounds (VOCs) using molecular fingerprints and neural network models, further support the effectiveness of this methodology.

Thus, the architecture from the model consists of an MLP structure, with appropriate hyperparameters properly tuned to optimise the results, whose outputs are the minimum appearance energy achieved through the electron ionisation method.

The relative error was chosen to evaluate the effectiveness of the MLP algorithm, providing a straightforward and interpretable measure of the prediction's accuracy. This value represents the mean relative error, calculated from each compound, considering the predicted parameter with respect to the real one.

Minimum appearance energy prediction

The same approach was adopted for the minimum appearance energy prediction, carefully tuning hyperparameters to optimise the results.

Unlike ionisation energy, no previous works were found regarding potential inputs or the training model. However, given the success found for the approach detailed in the preceding section, a molecular fingerprint input and a MLP architecture were chosen. Hence, the desired outcome of this architecture aims to determine the minimum appearance energy through the EI method.

Once again, the relative error was selected as the evaluation metric since the minimum appearance energy prediction, like the ionisation energy prediction, generates a single output.

Ion Mass prediction from the minimum AE fragmentation

For the prediction of the ion generated due to fragmentation at the minimum appearance energy, the lack of preceding references necessitated an innovative approach. In

this regard, the molecular fingerprint was chosen as the primary input for the training model, and the MLP as the architecture, leveraging their successful track record in the previous predictions. The fitness metric used is the relative error of the molecular mass of the predicted ion.

Mass spectrum prediction

The mass spectra prediction for chemical compounds was also based on MLP using a similar procedure to [47] and [49]. However, additional architectures were also explored to enhance the results. Among these, LSTM network and bi-LSTM network were considered due to their potential advantages in handling sequential data [57].

These studies have intimated the existence of correlations between consecutive peaks in the mass spectrum. This characteristic adjusts well with the inherent capability of this architecture to capture sequential dependencies in data. By utilising LSTM, the algorithm aims to exploit the information encoded in the sequential order of peaks, enhancing its ability to discern patterns and relationships within the mass spectrum data. This aligns with the fundamental motivation behind adopting LSTM for this specific prediction task.

In this case, the input layer receives the molecular fingerprint data, more specifically, the ECFPs, and the output layer generates the predicted mass spectra. This final layer of the model possesses an equal length to the desired output, in other words, the mass spectra vector length. This vector was structured with relative intensities, which were positioned in the vector according to their respective mass-to-charge ratio. Moreover, an attempt was made to enhance the model's performance by normalising the values with respect to the peak situated at 100%.

Following the reference papers, the cosine similarity metric was utilised to assess the performance of their model (calculating the cosine of the angle between the predicted mass spectra vector and the one in the database).

Since the objective is to design a model capable of accurately predicting the electron ionisation mass spectrum for any given molecule, the model is then employed to construct an augmented reference library comprising both predicted and experimentally measured spectra.

Subsequently, library matching is performed: the cosine similarity between each predicted spectrum and every spectrum from the augmented library is calculated and sorted from highest to lowest. Then, the rank of the correct spectrum is recorded (for each molecule used as an input during validation). The results obtained in this section are also evaluated using recall@k, where k represents values of 1, 5, or 10. This parameter measures the proportion of cases in which the correct spectrum appears within the top k ranked results. For instance, recall@1 indicates the percentage of times the correct match is ranked first, while recall@10 reflects how often it appears within the top ten. Figure 3 shows a visual depiction of this process.

Furthermore, a mass filter, similar to the one employed in the previous papers, was also integrated into the algorithm to facilitate the similarity search and analyse a more concise accuracy. The mass filter excludes species that fall outside the actual compound mass within a predetermined tolerance of 30 Da (±15 Da). This mechanism ensures that only compounds within the specified mass range are taken into consideration.

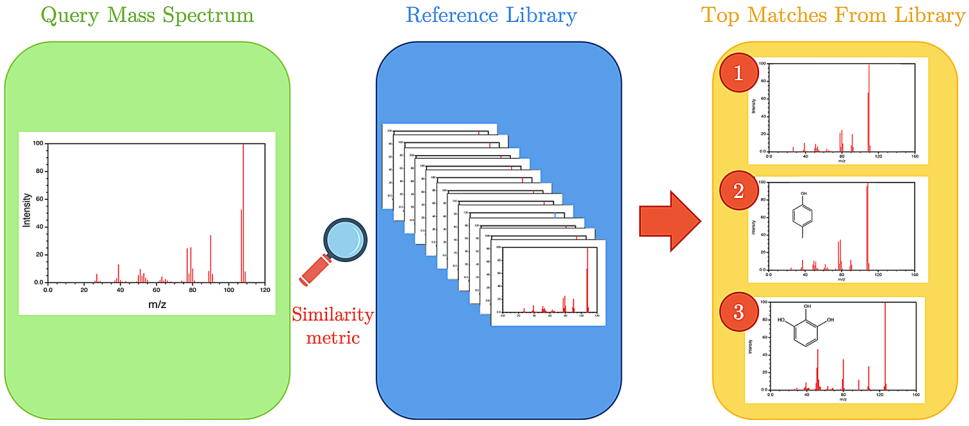

Fig. 3 Overview of the library matching. Modified from [49]

Table 1 Configuration parameters summary table for optimal prediction results

	IE	AE	AE mass	MS
MLP layers	5	5	5	5
Hidden neurons	512	64	64	4096
Epochs	22	28	16	28
Batch size	2	2	1	32
Learning rate	0.001	0.01	0.04	0.001
Dropout	0.15	0.3	0.4	0.2
Activation function	ReLU	Sigmoid	Leaky ReLU	Leaky ReLU
Optimiser	RMSProp	Adagrad	Adagrad	Adam

This strategic filtering was designed to retain a substantial number of potential candidates while minimising the reduction in the candidate list. The chosen tolerance value was determined through careful analysis, ensuring an average subset size of approximately 400 compounds within the filtered dataset. This value was selected to maintain a relative proportion to the filtered dataset length observed in the reference papers [47, 49].

Results and discussion

In this section, the results and analysis obtained from the application of the ML-based assessment system for mass spectrum prediction, ionisation energy prediction, and minimum appearance energy prediction are presented.

The summary of the optimal configuration hyperparameters found during the hyperparameter tuning of the various machine learning models is presented in Table 1.

Ionisation energy prediction results

The best predictions of the ionisation energy using the methodology of Sect. 2.4 are obtained using the configuration succinctly summarised in Table 1. Figure 4 offers the evolution of the training and validation loss after each epoch (a measurement of the training effectiveness after each training round), illustrating a consistent downward trend across all curves, successfully aligning with the expected behaviour.

The curve discrepancy and the validation loss curve noise relative to the smoother training curve could likely stem from the relatively modest database size and the potential for further training gains. Nonetheless, the overarching trajectories converge to a

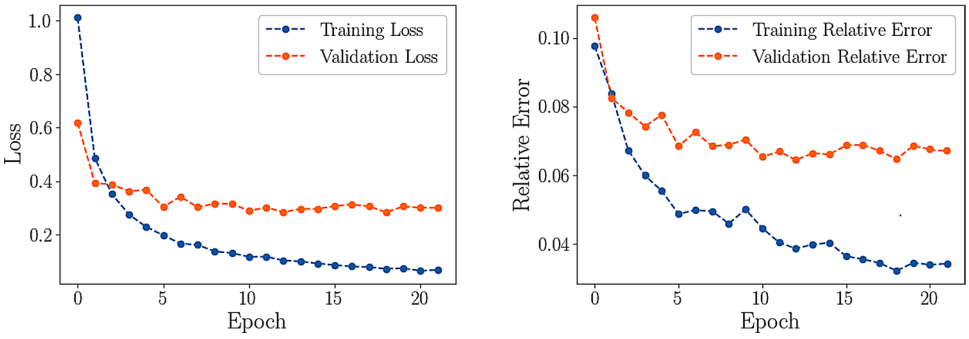

Fig. 4 Performance from the IE prediction model with the configurations from Table 1 using the training and validation sets

Table 2 Top 5 predictions from the IE prediction with the model from Table 1

	$C_8H_6F_2O_2$	$C_{13}H_{11}N$	$C_{15}H_{12}$	$C_{10}H_{13}BrO$	$C_{12}H_{16}O$
Predicted IE [eV]	8.88	8.15	7.70	8.55	7.99
Real IE [eV]	8.88	8.15	7.70	8.54	8.00

stable value for both training and validation losses and for the relative error. In fact, for this optimised model, a successful prediction of ionisation energy was accomplished, with a considerably low relative error of 6.87%. In comparison, if the average ionisation energy of the whole population was used as a predictor, the average relative error would be 18.05%.

It is important to note that the ionisation energies in the test set are representative of the whole set, falling within 69.48% of the average value, with values ranging from 3.9 to 21.6 eV. The absolute error of the predictions was analysed as a function of the energy, and no trend was found between the two.

In Table 2, the 5 predictions with the lowest relative error are presented.

Minimum appearance energy prediction results

This section delves into the outcomes of forecasting the minimum appearance energy. Firstly, attention is directed towards Table 1, which exhibits the configurations of the models that produced the least relative errors reached, pertaining to the prediction of minimum appearance energy.

Table 1 exhibits the configuration of the model that produced the lowest relative errors for the prediction of minimum appearance energy. Employing these identified hyperparameters and settings for the model, the loss curves and relative errors of Fig. 5 were obtained.

The graphs consistently portray a common trend, with all curves gradually descending and converging towards stable values. As in the previous section, the validation curves stabilise to a higher value than the training ones but with a very low relative error, but the convergence criteria must also consider the validation subset results to avoid overfitting. Moreover, there is an initial plateau in these curves, which can be potentially attributed to a less convex loss landscape with respect to the preceding section.

Ultimately, employing these identified hyperparameters and settings for the model, the attained minimal relative error for the prediction of minimum appearance energy yielded a value of 7.99%. By calculating the relative error using the mean value from the

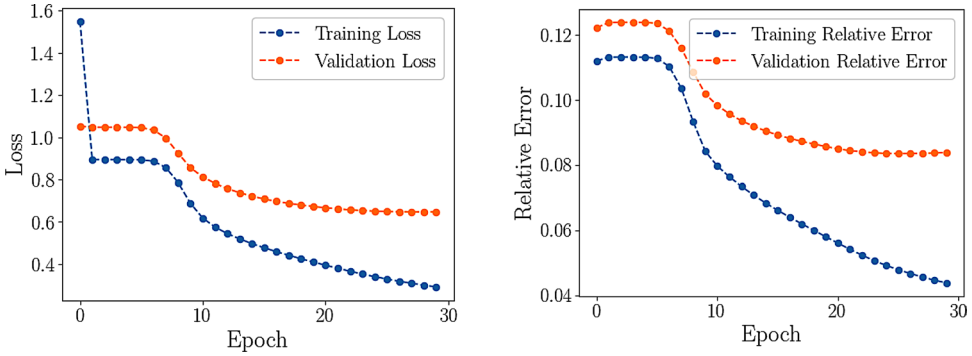

Fig. 5 Performance from the minimum AE prediction model with the configurations from Table 1

Table 3 Top 5 predictions from the AE prediction with the model from Table 1

	C_3H_5ClO	C_4H_8OS	C_7H_7F	$C_7H_{16}O_2$	C_2H_7N
Predicted AE [eV]	10.29	9.89	11.88	10.32	9.53
Real AE [eV]	10.29	9.90	11.90	10.30	9.55

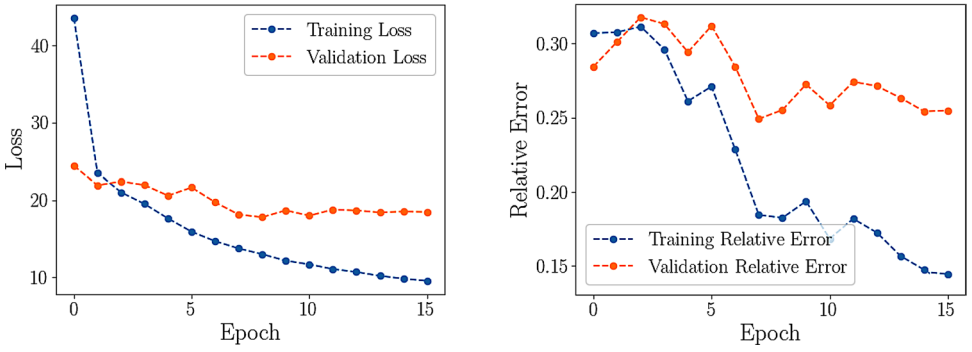

Fig. 6 Performance from the prediction model of the ion mass for minimum AE with the configurations from Table 1

dataset, the error increases to 19.1%, thereby reinforcing the accuracy of the model. All appearance energies values from the test dataset range within 44.09% of the average value, in the 7.1–18.3 eV range. In this case, the error also did not have any trend with respect to the energy. The 5 best predictions are represented in Table 3.

Ion mass prediction from the minimum AE fragmentation results

The model configurations that yielded the lowest relative errors are depicted in Table 1. Using those parameters, the training and ion mass prediction relative error curves of Fig. 6 were obtained.

In contrast to all the previous results, the ion mass prediction graphs display a comparatively higher degree of noise, ultimately converging to relatively higher error rates. This is not necessarily a surprising fact, as the algorithm needs to predict not only when the molecule will break (implicitly, the appearance energy is also necessary) but also how it will break. Therefore, the molecule's chemical structure (and its encoding) is much more critical, increasing the problem's complexity.

Since these results exhibited a less satisfactory performance, some attempts were made to enhance the model. Firstly, the normalisation of the outputs was tested, involving the

prediction of the ion mass percentage relative to the initial mass. Furthermore, other models, such as LSTM, were also implemented. However, this adjustment did not yield improvements in model outcomes.

Additionally, the incorporation of the original compound mass as an input led to some enhancements in the model. Consequently, normalising this mass input with respect to the maximum value across the database was explored, given that the remaining inputs, which are the ECFPs are represented as binary bits (0 or 1). Despite these measures, improvements remained elusive. Finally, the relative error for the associated ion mass prediction culminated at 23.89%, with the data subset presenting a wide range of values from 6 to 625 Da, that is a 98.10% variation from the average value. Taking the average value from the entire dataset, a relative error of 96.74% would be obtained, so while the relative error is higher than for the previous parameters, the algorithm still outperforms the average significantly. In Table 4, the best 5 predictions from this model are listed.

Further iterations involving variations in the fingerprint inputs, such as alterations to radii and bit string length, were also trialled but yielded no enhancements either. Thus, it is plausible that the need for additional inputs or the inherently variable nature of the parameter under prediction could cause this to be a complicated value to predict.

Mass spectrum prediction results

Finally, three different model architectures were used to predict the full mass spectra resulting from the electron ionisation of a molecule with an energy of 70 eV, the most complex problem. This specific value from the excitation energy is standardised for this mass spectrum technique. Through hyperparameter tuning, it was determined that the MLP proved to be the most effective model for predicting the mass spectrum, achieving a notable cosine similarity of 0.6395 and a recall@10 of 60.68%, with its optimal hyperparameters.

In contrast, the LSTM model yielded a cosine similarity of 0.5517 and a recall@10 of 49.62% with the best-tuned parameters, while the bi-LSTM demonstrated a cosine similarity of 0.5708 and a recall@10 of 52.36%.

The inclusion of the mass from the chemical compounds as an input and the variation of the input features, such as altering ECFPs, in terms of bit length or radii, did not yield performance improvements. Nevertheless, a notable enhancement was observed through the normalisation of relative intensities from the mass spectrum data.

Table 1, details the specific values for the different hyperparameters that yielded the optimal results.

Visual representations of the loss function's behaviour over the training epochs and the corresponding trend in cosine similarity (Fig. 7) provide insights into the convergence behaviour and predictive accuracy achieved by the model from the previous table.

Consequently, the final results, corresponding to the outcomes of the test subset, are displayed in Table 5:

Furthermore, with a mass filter to selectively focus on compounds with comparable masses, employing a tolerance of 30 Da, up to a recall@10 of 78% was achieved.

Table 4 Top 5 predictions from the ion mass for minimum AE with the model from Table 1

	C_9H_{12}	$C_{13}H_{21}NO$	C_7H_8	$C_9H_1ClN_2$	$C_7H_5NO_4$
Predicted ion mass [Da]	57.12	105.16	45.06	91.13	78.11
Real ion mass [Da]	57.67	104,05	46.01	94.45	82.06

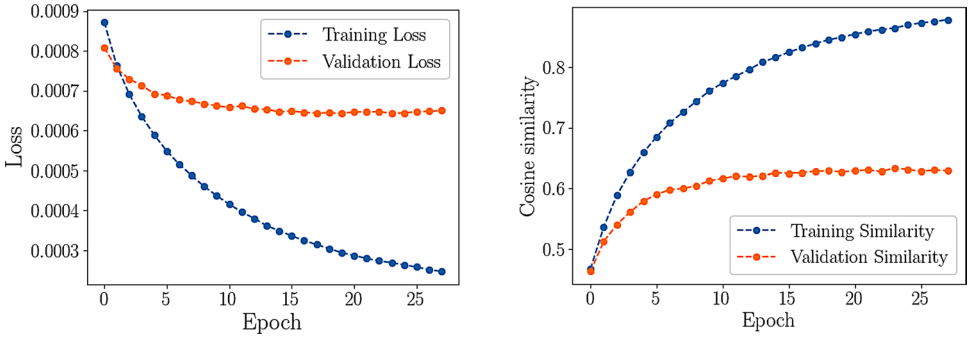

Fig. 7 Performance from the MS prediction model with the configurations from Table 1

Table 5 Results from the MS prediction with the model from Table 1

Cos. sim.	recall@1	recall@5	recall@10
0.6395	31.04%	52.89%	60.63%

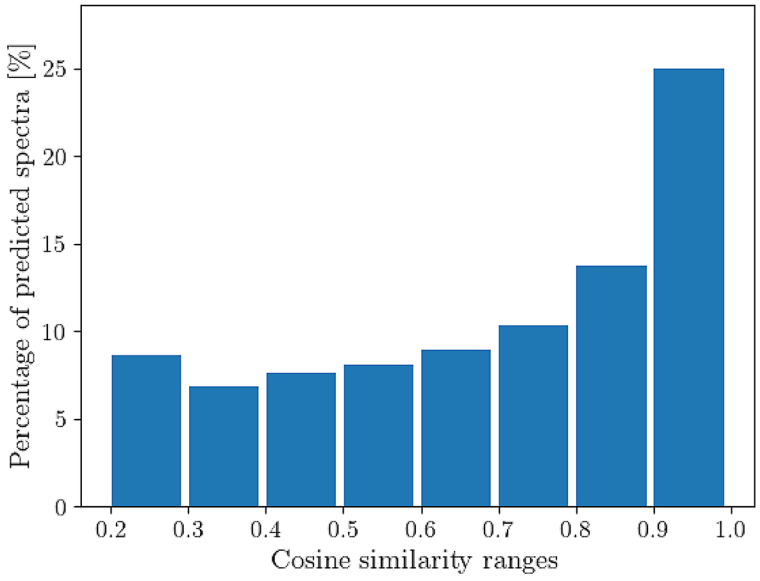

Fig. 8 Cosine similarity distribution from the test set of the MS prediction

To further gauge the efficacy of this model using the employed metrics, an additional illustrative representation is provided in Fig. 8. This graph presents the distribution of predictions from the test subset compounds, with a mean cosine similarity of 0.6395 and a standard deviation of 0.2809, and are categorised according to specific ranges of cosine similarities. Each bar in the graph represents the percentage of compounds that fall within a particular range of the similarity score.

As perceptible from the graph, a substantial frequency of predictions demonstrates cosine similarity values surpassing 0.6, amounting to more than 60% of occurrences. Furthermore, accurate predictions are discernible, as approximately one-fourth of the forecasts possess cosine similarity values exceeding 0.9.

Adamantane ($C_{10}H_{16}$), one of these cases and technologically relevant, is used to illustrate in Fig. 9 the ultimate outcomes of the prediction algorithm, and an additional

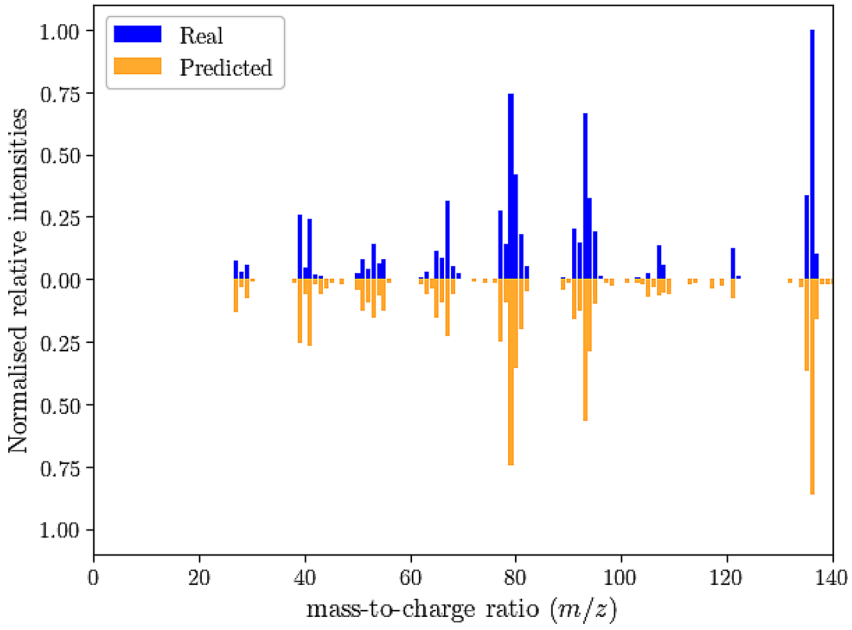

Fig. 9 Comparison between the MS prediction and the ground truth for $C_{10}H_{16}$

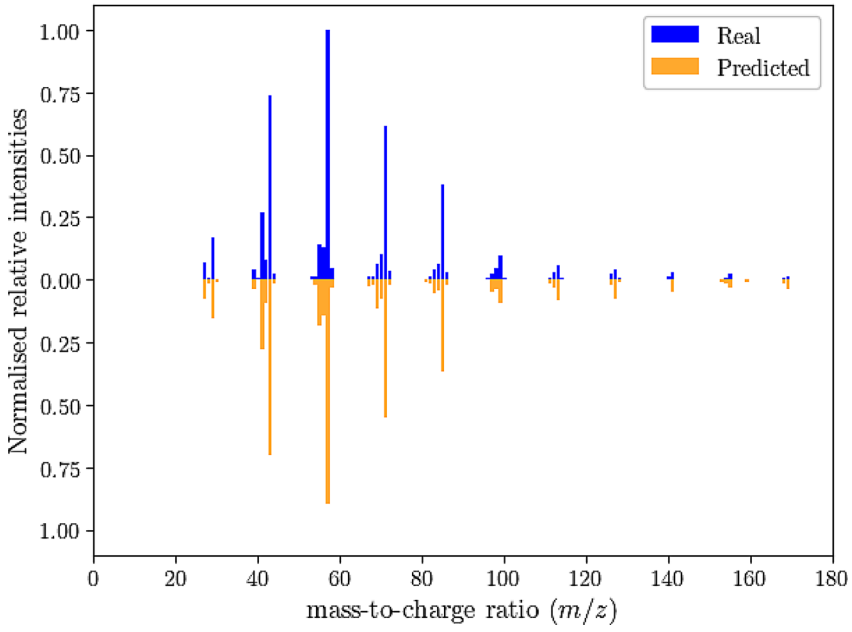

Fig. 10 Comparison between the MS prediction and the real for $C_{15}H_{32}$. Representation of the normalised relative intensities against the mass-to-charge ratio

example with pentadecane ($C_{15}H_{32}$) is also shown in Fig. 10. These compounds from their test subsets obtained cosine similarities of 0.9604 and 0.9912, respectively.

The graphical representation reveals a notable resemblance between the predicted outcomes and the actual values. While some disparities are discernible, particularly in the case of lower peaks, the accuracy of prediction is remarkable, especially in regard to the more prominent peaks. This alignment is particularly crucial as these significant peaks typically encapsulate substantial information within the mass spectrum.

Similarly to the case of ion mass prediction, the interaction between chemical bonds within the molecule reduces the fingerprint's effectiveness as a unique algorithm input. While the tuning of the fingerprint has not yielded any significant improvement, changing the algorithm (and fingerprints) depending on the number of bonds might be an improvement venue.

Furthermore, the predictive performance has the potential for enhancement through the incorporation of a graph convolutional network (GCN) for processing the ECFPs. GCN, as proposed in [49] and successfully demonstrated in [47], treats molecules as graphs, with atoms as nodes and chemical bonds as edges. This approach allows for a comprehensive representation of the entire molecular structure, potentially leading to improved predictive accuracy at a slightly higher computational cost. Nonetheless, it was not possible to recreate the same procedure from the references mentioned, due to the lack of specific sub-databases for it in the open-access version from NIST.

Finally, one of the main challenges of this section is the training database. The MS NIST database, as the name states, is used to calibrate mass spectrometers, and in consequence, its structure is not inherently optimized for training algorithms. Despite this, it remains the most suitable and comprehensive database currently available for our application. Three of the main limitations are that the spectra count the number of ions with a particular specific charge without reference to the number of ionised molecules, a similar consideration is that there is no data on ionisation efficiency (especially relevant for (especially relevant for EP), and finally, while most compounds are ionised at 70 eV, this is not always the case. Additionally, in an ideal case, the compounds should be ionised at a level representative of the ionisation dynamics where they would be used. In the case of HET, although the electron energy is not constant, it can be considered 10% of the voltage between electrodes [33], being 30 eV a good reference value [58] for standard power. However, as just discussed, in the case of high-power HET, the electron energy can be significantly higher with energies up to 100 eV being considered [58]. For Gridded Ion Engine (GIE), the electron energies tend to be lower from a few eV to 20 eV [59]. So, while the ionisation energy used to train the algorithm is on the higher end of the usual ionisation energy, it is still within the range of HET and of the same magnitude as GIE. A more refined approach would involve generating a database with various excitation energies to train the model, enabling it to calculate the spectra at the relevant energy level.

However, with this imperfect dataset, it has still been possible to demonstrate the potential of ML to predict the results of ionising a complex molecule in an electric thruster. In the future, a secondary layer of machine learning algorithms could also be implemented, combining the results from all predictions and classifying which chemical compounds are viable xenon substitutes. [60] Finally, one of the main issues of this section is the training database. The MS NIST database, as the name states, is used to calibrate mass spectrometers, and in consequence, it is not necessarily the best to train the algorithms. Three of the main issues are that the spectra count the number of ions with a particular specific charge without reference to the number of ionised molecules, a similar issue is that there is no data on ionisation efficiency (especially relevant for EP), and finally, while most compounds are ionised at 70 eV, this is not always the case. For the latter problem, a more refined approach would involve generating a database with various excitation energies to train the model.

However, with this imperfect dataset, it has still been possible to demonstrate the potential of ML to predict the results of ionising a complex molecule in an electric thruster. In the future, a secondary layer of machine learning algorithms could be implemented, combining the results from all predictions and classifying which chemical compounds are viable xenon substitutes. [60]

Conclusions

The project has implemented three different ML architectures, using the molecular fingerprint as the input, to predict the ionisation and fragmentation patterns of chemical compounds, particularly the ionisation energy, the minimum appearance energy, its associated ion mass, and the mass spectrum.

The ionisation energy was predicted with a mean relative error of 6.87%, and 7.99% for minimum appearance energy. Predictions derived from the ion mass associated with the minimum appearance energy exhibited a 23.89% mean relative error.

Finally, the prediction of mass spectra showcased a cosine similarity of 0.6395, and the predicted outcome fell within the top 10 most similar mass spectra in 60.63% of cases. Furthermore, focusing on compounds within a 30 Da range of the predicted mass, this recall@10 climbed to a notable 78%.

To build upon insights gleaned from fragmentation pattern predictions, subsequent research could involv0e the implementation of a secondary layer of machine learning algorithms. These algorithms could be trained on combined results from all predictions, thus classifying chemical compounds as viable Xenon substitutes or not. A feasible method for classification, for instance, would involve employing trade-offs, as exemplified in [60].

In summary, the project showcased the predictive capabilities of ML to understand the behaviour of chemical compounds when ionised and presented three viable implementations, underscoring the potential of artificial intelligence to drive electric propulsion propellant development.

Nomenclature

AE	Appearance energy
ECFPs	Extended circular fingerprints
EI	Electron ionisation
EP	Electric propulsion
GIE	Gridded Ion Engine
HET	Hall Effect Thruster
LSTM	Long short-term memory
ML	Machine learning
MLP	Multi-layer perceptron
MS	Mass spectrum
IE	Ionisation Energy
NIST	National Institute of Standards and Technology
ReLU	Rectified linear unit
SGD	Stochastic gradient descent

Author contributions

EGG Provided the idea, context and importance of the research, outlined the technical challenges and importances for electric propulsion and molecular propellants, and was responsible for the final edit of the manuscript. MAC provided the guidance, expertise, and contextualisation of the AI part of the research. APD carried the brunt of the coding and preparation of figures and first manuscript draft. All authors collaborated on the analysis of the results and reviewed the manuscript.

Funding

This study was carried out without any external funding.

Data availability
The training data has been obtained from the available compounds of the National Institute of Standards and Technology (NIST) Chemistry WebBook https://webbook.nist.gov/chemistry/.

Declarations

Competing interests
EGG is a founder of Ignition Space who develops algorithms for Space applications. However, the training data set is open access and the algorithms architecture fully discloses allowing to replicate the results by other scientists without a paywall.

Received: 16 September 2024 / Accepted: 22 September 2025
Published online: 07 October 2025

References

1. Welle R (1993) Xenon and krypton availability for electric propulsion-an updated assessment. In 29th Joint Propulsion Conference and Exhibit, pp 2401
2. Andreussi T, Saravia MM, Ferrato E, Piragino A, Rossodivita A, Andrenucci M, Estublier D (2017) Identification, evaluation and testing of alternative propellants for hall effect thrusters. In 35th International Electric Propulsion Conference, pp 2017–2380
3. Fazio N, Gabriel S, Golosnoy IO (2018) Alternative propellants for gridded ion engines. Space propulsion 2018 14/05/18–18/05/18), pp. 2018–00102. https://eprints.soton.ac.uk/422369/
4. Oleson S (2004) Electric propulsion for project prometheus. In 39th AIAA/ASME/SAE/ASEE Joint Propulsion Conference and Exhibit, pp 5279
5. Magaldi B, Karnopp J, Silva Sobrinho A, Pessoa R. (2022) A global model study of plasma chemistry and propulsion parameters of a gridded ion thruster using argon as propellant. Plasma 5(3):324–340. https://doi.org/10.3390/plasma5030025
6. Eckhaus AJ, Rocha AJ, Saladino AM, Jorns B (2024) Student-led design, construction, and testing of a permanent magnet hall thruster on argon propellant. AIAA SCITECH 2024 Forum 1955. https://doi.org/10.2514/6.2024-1955
7. Bryan HC, McDowell DJ, Welty AK, Kropp MT, Fujimoto MS, Hansen JK, Riley B, Thallapally P (2023 September) Cost-benefit assessment of krypton and xenon recovery from aqueous reprocessing. Technical report, Idaho National Laboratory (INL), Idaho Falls, ID (United States) https://doi.org/10.2172/2377416
8. Szabo JJ, Robin M, Hruby V (2017) Bismuth vapor hall effect thruster performance and plume experiments. In 35th International Electric Propulsion Conference, pp 1–13
9. Szabo J, Robin M, Paintal S, Pote B, Hruby V, Freeman C (2013) Iodine propellant space propulsion. In 33rd International Electric Propulsion Conference, pp 2013–2311
10. Tverdokhlebov O, Semenkin A (2001) Iodine propellant for electric propulsion-to be or not to be. In 37th Joint Propulsion Conference and Exhibit, pp 3350
11. Ling WYL, Zhang S, Hao F, Huang M, Quansah J, Xiangyang L, Ningfei W (2020) A brief review of alternative propellants and requirements for pulsed plasma thrusters in micropropulsion applications. Chin J Aeronaut 33(12):2999–3010
12. Anderson J, Fitzgerald D (1996) Fullerene propellant research for electric propulsion. In 32nd Joint Propulsion Conference and Exhibit, pp 3211
13. Scharlemann CA (2002) Theoretical and experimental investigation of c60-propellant for ion propulsion. Acta Astronautica 51(12):865–872. https://doi.org/10.1016/S0094-5765(02)00115-7
14. Hill AW, Mortishire-Smith RJ (2005) Automated assignment of high-resolution collisionally activated dissociation mass spectra using a systematic bond disconnection approach. Rapid Commun Mass spectrom 19(21):3111–3118. https://doi.org/10.1002/rcm.2177
15. Heinonen M, Rantanen A, Mielikäinen T, Kokkonen J, Kiuru J, Ketola RA, Rousu J (2008) Fid: a software for ab initio structural identification of product ions from tandem mass spectrometric data. Rapid Commun Mass spectrom 22(19):3043–3052. https://doi.org/10.1002/rcm.3701
16. Wolf S, Schmidt S, Müller-Hannemann M, Neumann S (2010) In silico fragmentation for computer assisted identification of metabolite mass spectra. BMC Bioinf 11(1):148. https://doi.org/10.1186/1471-2105-11-148
17. Wang Y, Kora G, Bowen BP, Pan C (2014) Midas: a database-searching algorithm for metabolite identification in metabolomics. Analytical Chem 86(19):9496–9503. https://doi.org/10.1021/ac5014783
18. Krettler CA, Thallinger GG (2021) A map of mass spectrometry-based in silico fragmentation prediction and compound identification in metabolomics. Briefings Bioinf 22(6):073. https://doi.org/10.1093/bib/bbab073
19. De Proft F, Geerlings P (1997) Calculation of ionization energies, electron affinities, electronegativities, and hardnesses using density functional methods. J Chem Phys 106(8):3270–3279. https://doi.org/10.1063/1.473796
20. Flad J, Stoll H, Preuss H (2008) Calculation of equilibrium geometries and ionization energies of sodium clusters up to na8. J Chem Phys 71(7):3042–3052. https://doi.org/10.1063/1.438710
21. Gasteiger J, Hanebeck W, Schulz KP (1992) Prediction of mass spectra from structural information. J Chem Inf Comp Sci 32(4):264–271. https://doi.org/10.1021/ci00008a001
22. Yang Y, Horvatovich P, Qiao L (2021) Fragment mass spectrum prediction facilitates site localization of phosphorylation. J Proteome Res 20(1):634–644. https://doi.org/10.1021/acs.jproteome.0c00580
23. Xu R, Sheng J, Bai M, Shu K, Zhu Y, Chang C (2020) A comprehensive evaluation of MS/MS spectrum prediction tools for shotgun proteomics. Proteomics 20(21–22), 1900345. https://doi.org/10.1002/pmic.201900345
24. Bremer PL, Vaniya A, Kind T, Wang S, Fiehn O (2022) How well can we predict mass spectra from structures? benchmarking competitive fragmentation modeling for metabolite identification on untrained tandem mass spectra. J Educ Chang Chem Inf And modeling 62(17):4049–4056. https://doi.org/10.1021/acs.jcim.2c00936

25. Allen F, Pon A, Wilson M, Greiner R, Wishart D (2014) Cfm-id: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. Nucleic Acids Res 42(W1):94–99. https://doi.org/10.1093/nar/gku436
26. Cautereels J, Claeys M, Geldof D, Blockhuys F (2016) Quantum chemical mass spectrometry: ab initio prediction of electron ionization mass spectra and identification of new fragmentation pathways. J. Mass spectrom 51(8):602–614. https://doi.org/10.1002/jms.3791
27. Jonkman HT, Velde GA, Nieuwpoort WC (1974) Ab initio SCF MO calculation of ionisation energies and charge distributions of TCNQ and its mono- and divalent anions. Chem Phys Lett 25(1):62–65. https://doi.org/10.1016/0009-2614(74)80332-5
28. Lo Y-C, Rensi SE, Torng W, Altman RB (2018) Machine learning in chemoinformatics and drug discovery. Drug Discov Today Today 23(8):1538–1546. https://doi.org/10.1016/j.drudis.2018.05.010
29. Choueiri EY (2009) Physics of electric propulsion. Princeton University Press
30. Mazouffre S (2016) Electric propulsion for satellites and spacecraft: established technologies and novel approaches. Plasma Sources Sci Technol 25(3), 033002. https://doi.org/10.1088/0963-0252/25/3/033002
31. Jahn RG (2006) Physics of electric propulsion. Dover Publications, Mineola, New York, USA, pp 2–11
32. Baba MS, Narasimhan TSL, Balasubramanian R, Mathews CK (1995) Ionization and fragmentation of c[sub 60]. An electron impact ionization study. J Phys Chem C; (United States) 99(10). https://doi.org/10.1021/j100010a010
33. Goebel DM, Katz I, Mikellides IG (2023) Fundamentals of electric propulsion, 2nd edn. Wiley, Hoboken, New Jersey, USA, pp 316–384
34. Lozano P, Martinez-Sanchez M (2005) Efficiency estimation of emi-bf4 ionic liquid electrospray thrusters. In 41st AIAA/ASME/SAE/ASEE Joint Propulsion Conference & Exhibit. https://doi.org/10.2514/6.2005-4388
35. Grustan-Gutierrez E, Jhuree S, Stark J (2017) Modelling of colloid thrusters for mission analysis. In Conference: International Electric Propulsion Conference 2017
36. Zhang W, Gabriel S, Stoukatch S, Eckersley S (2007) Electric propulsion assessment for esprit mission. In 2007 European Space Power Conference, ESA Publications, 9–12
37. Oleson SR, McGuire ML, Mercer CR (2004) Electric propulsion for mars sample return. J Spacecr Rockets 41(6):940–944
38. Wittenberg LJ (2019) Electric propulsion: progress, challenges, and opportunities. J Propul Power 35(4):781–795
39. Qu X, Latino DARS, Aires-de-Sousa J (2013) A big data approach to the ultra-fast prediction of dft-calculated bond energies. J cheminform 5(34). https://doi.org/10.1186/1758-2946-5-34
40. Li W, Dong H, Ma J, Li S (2021) Structures and spectroscopic properties of large molecules and condensed-phase Systems predicted by generalized energy-based fragmentation approach. Acc Chem Res 54(1):169–181. https://doi.org/10.1021/acs.accounts.0c00580
41. Ljoncheva M, Stepišnik T, Kosjek T, Džeroski S (2022) Machine learning for identification of silylated derivatives from mass spectra. J cheminform 14(1):62. https://doi.org/10.1186/s13321-022-00636-1
42. Matthews DA (2020) Eom-cc methods with approximate triple excitations applied to core excitation and ionisation energies. Mol Phys 118(21–22), 1771448. https://doi.org/10.1080/00268976.2020.1771448
43. Yerokhin VA, Patkóš V, Puchalski M, Pachucki K (2020) Qed calculation of ionization energies of 1*snd* states in helium. Phys Rev A 102, 012807. https://doi.org/10.1103/PhysRevA.102.012807
44. Klamt A, Reinisch J, Eckert F, Hellweg A, Diedenhofen M (2012) Polarization charge densities provide a predictive quantification of hydrogen bond energies. Phys Chem Chem Phys 14:955–963. https://doi.org/10.1039/C1CP22640A
45. Qiu R, Zhou W, Zheng Y, Hou H, Wang B (2023) Prediction models of the ionization coefficient and ionization cross-section based on multi-layer molecular parameters. Plasma Sci Technol 25(5). https://doi.org/10.1088/2058-6272/acac65
46. Stewart MP, Martin ST (2023) Machine learning for ionization potentials and photoionization cross sections of volatile organic compounds. ACS Earth Space Chem 7(4):863–875. https://doi.org/10.1021/acsearthspacechem.3c00009
47. Zhang B, Zhang J, Xia Y, Chen P, Wang B (2022) Prediction of electron ionization mass spectra based on graph convolutional networks. Int J Mass spectrom 475, 116817. https://doi.org/10.1016/j.ijms.2022.116817
48. Morgan HL (1965) The generation of a unique machine description for chemical structures-A technique developed at chemical abstracts service. J Chem doc 5(2):107–113. https://doi.org/10.1021/c160017a018
49. Wei JN, Belanger D, Adams RP, Sculley D (2019) Rapid prediction of electron–ionization mass spectrometry using neural networks. ACS Central Sci 5(4):700–708
50. Bengio Y, Goodfellow I, Courville A (2017) Deep learning, vol 1. MIT press, Cambridge, MA, USA,
51. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural computation 9(8):1735–1780
52. Graves A, Schmidhuber J (2005) Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Netw 18(5–6):602–610
53. Li H, Xu Z, Taylor G, Studer C, Goldstein T (2018) Visualizing the loss landscape of neural nets. In proceedings of the 32nd international conference on neural information processing systems. NIPS'18, Curran Associates Inc, Red Hook, NY, USA, 6391–6401
54. Ruder S (2017) An overview of gradient descent optimization algorithms. https://doi.org/10.48550/arXiv.1609.04747
55. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press
56. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444
57. Le X-H, Ho HV, Lee G, Jung S (2019) Application of long short-term memory (LSTM) neural network for flood forecasting. https://doi.org/10.3390/w11071387
58. Goebel DM, Katz I (2008) Fundamentals of electric propulsion, 1st edn edn. John Wiley & Sons, Ltd, Hoboken, New Jersey, USA, p 476. https://doi.org/10.1002/9780470436448.app5
59. Goebel DM, Katz I, Mikellides IG (2023) Fundamentals of electric propulsion, 2nd edn. Wiley, Hoboken, New Jersey, USA, pp 184–264
60. Giannetti V, Andreussi T, Leporini A, Gregucci S, Saravia M, Rossodivita A, Estublier D, Edwards C, Mode T (2016) Electric propulsion sytem trade-off analysis based on alternative propellant selection. In 5th Space Propulsion Conference, pp 2016–3125194

Publisher's Note