Journal of Applied Physics ARTICLE scitation.org/journal/jap

Fast prediction of electron-impact ionization cross sections of large molecules via machine learning

Cite as: J. Appl. Phys. **125**, 183302 (2019); doi: 10.1063/1.5094500
Submitted: 2 March 2019 · Accepted: 21 April 2019 ·
Published Online: 8 May 2019

View Online Export Citation CrossMark

Linlin Zhong[a)]

AFFILIATIONS

School of Electrical Engineering, Southeast University, No. 2 Sipailou, Nanjing, Jiangsu Province 210096, People's Republic of China

[a)]mathboylinlin@gmail.com and linlin@seu.edu.cn

ABSTRACT

The theoretical determination of electron-impact ionization cross section (Q_{ion}) for a molecule requires *ab initio* computation, which is time-consuming for large molecules. We propose a machine learning based method to construct a model for predicting Q_{ion} of large molecules without the high-cost *ab initio* calculation. The model is learned from the data composed of the calculated Q_{ion} of the small molecules with fewer constituent atoms and the electron numbers of the corresponding molecules in a train set by a support vector machine. The radial basis function is set as a kernel function to map data to a higher dimensional space. The grid search with 5-fold cross-validation is performed to find optimal hyperparameters in the learning model. The prediction on the test sets composed of CF_4, C_3F_8, SF_6, C_6, C_6F_{12}, and $C_6F_{12}O$ shows that this data-driven model can generate well-agreed Q_{ion} and has good generalization performance.

Published under license by AIP Publishing. https://doi.org/10.1063/1.5094500

I. INTRODUCTION

The electron-impact ionization of an atom or molecule is one of the most fundamental collision processes in atomic and molecular physics as well as in many areas of application, such as plasma discharges, excimer lasers, mass spectrometry, and radiation chemistry. Many efforts have, therefore, been devoted to the study of electron-impact ionization cross sections (Q_{ion}) of atoms or molecules since the 1930s both by experimentalists and by theorists.[1] Due to the high expenditure of experiments and with the development of computational quantum chemistry, an increasing number of molecular Q_{ion} are determined by theoretical approaches, especially for the expensive gases or the gases composed of large molecules that are easy to decompose and generate noxious products.[2] For instance, the calculated Q_{ion} of the perfluoroketone (PFK) molecules $C_xF_{2x}O$ ($x = 1-5$) by the Binary-Encounter-Bethe (BEB) method and by the Deutsch–Märk (DM) method were reported recently.[3] The BEB and DM methods are two most widely used approaches for calculating Q_{ion}, both of which incorporate quantum mechanically calculated molecular structure information. However, such quantum chemical computation including geometry optimization, frequency analysis, and high-precision energy calculation is time-consuming, particularly for large molecules. Figure 1 illustrates the total central processing unit (CPU) time consumed by the software Gaussian[4] for the quantum chemical computation in determining Q_{ion} of the PFK molecules following the calculation procedure described in Ref. 3 Obviously, the consumed time increases exponentially with the size of molecules. We can predict from Fig. 1 that the computation for the very large molecules, e.g., C_6F_{12} and $C_6F_{12}O$, will be very time-consuming and even beyond the time we can accept. However, to reduce the time, the accuracy of calculation must be sacrificed in the *ab initio* methods, for example, by using low-level basis sets.

In order to obtain Q_{ion} of large molecules within an acceptable time, a low-cost machine learning based method is proposed in this work. In the rest of the paper, the proposed method for the prediction of Q_{ion} via machine learning is first described, including the introduction of the calculation procedure, the construction of a dataset, and the parameters of the machine learning algorithm. The prediction method is then validated through the comparison between the predicted Q_{ion} and the experimental results for CF_4 and C_3F_8. The generalization of the method is also shown in the demonstration of the prediction of Q_{ion} for SF_6, C_6, C_6F_{12}, and $C_6F_{12}O$. Finally, the summary of the work is given.

II. PROPOSED METHOD

A. Calculation procedure

Machine learning is the study and construction of computer algorithms that can learn from data.[5] The ability of these data-driven

Published under license by AIP Publishing.

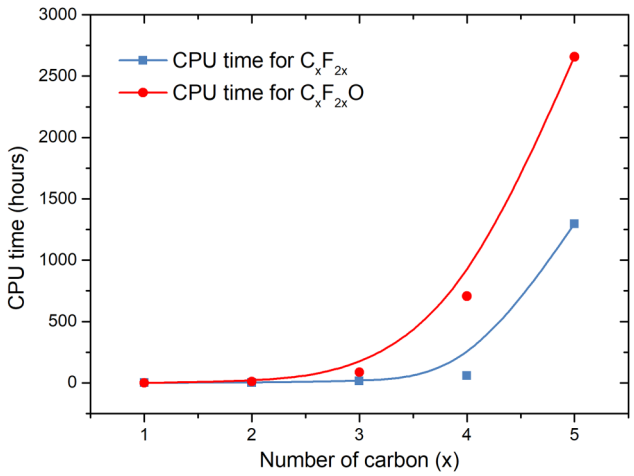

FIG. 1. Total CPU time for the quantum chemical computation in determining Q_{ion} of the molecules C_xF_{2x} and $C_xF_{2x}O$ (x = 1–5).

algorithms has led to a wide range of applications in science and technology, such as autonomous vehicle control, natural language translation, face recognition, medical diagnosis, and chemical syntheses.[5–8] We found that regression-based machine learning can also be applied for the prediction of molecular Q_{ion}. Q_{ion} describes the collisions between rapidly moving electrons and slow heavy molecules. Q_{ion} of a molecule is usually obtained by summing up the contributions arising from the ejection of an electron from the different molecular orbitals.[9] As a result, Q_{ion} of a molecule depends on the incident electron energy and the molecular structure information, such as molecular orbital composition, differential dipole oscillator strength, and binding energy and kinetic energy for each subshell.[9,10] Such molecular information can be obtained by *ab initio* calculation, which, however, is time-consuming for large molecules as described above. Through the investigation of Q_{ion} for large quantities of molecules, we found

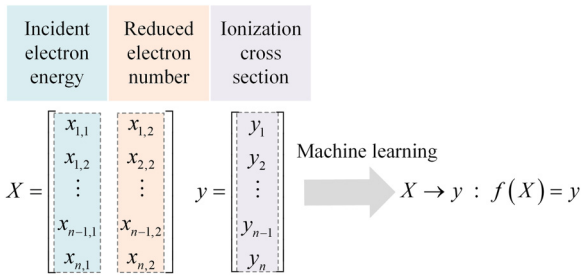

FIG. 3. Construction of attribute vector ***X*** and target vector *y* as required by the machine learning model which learns the functional relationship $f(\boldsymbol{X}) = y$ from these train data.

that the high-cost *ab initio* computation can be avoided in certain circumstances in which only the information of small molecules is needed to predict Q_{ion} of large molecules.

We use CF_4 as an example to demonstrate the machine learning based method for calculating Q_{ion} without the direct *ab initio* computation for the target molecule. After this demonstration, we will show the application of this method to other larger molecules including C_3F_8, SF_6, C_6, C_6F_{12}, and $C_6F_{12}O$. As illustrated in Fig. 2, the whole calculation procedure is divided into two parts. The first part is devoted to the calculation of Q_{ion} for the small molecules, which have fewer constituent atoms than the target large molecule. Data of these small and large molecules are referred to as train and test sets, respectively, in data-driven algorithms. For instance, to predict Q_{ion} of CF_4, CF_x (x = 1−3) are taken as a train set and CF_4 as a test set. The second part in the procedure is the construction of a machine learning based model which is learned from the data in a train set and has the ability to predict Q_{ion} of the molecule in a test set. Owing to this data-driven model, the *ab initio* computation for the large molecule in the test set is avoided. It is notable that in order to measure the generalization performance of the learning model on the test set, we also calculate Q_{ion} of large molecules based on *ab initio* theories in this work.

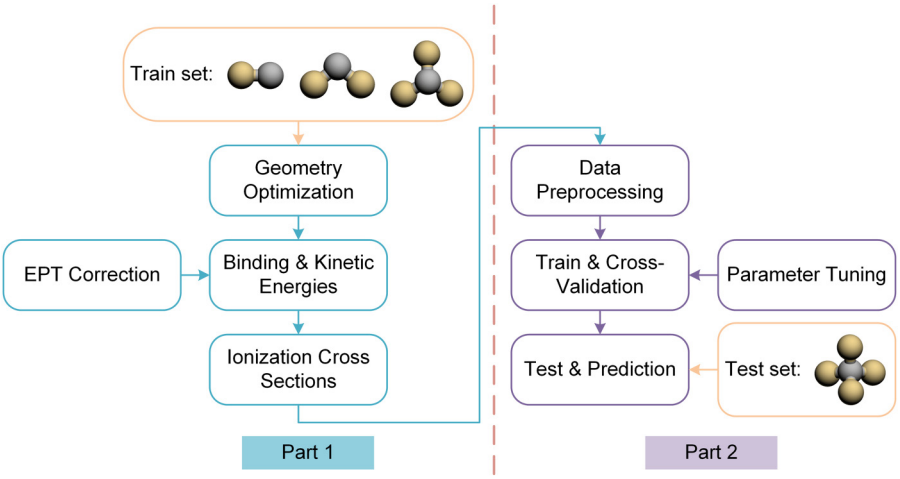

FIG. 2. Calculation procedure for the prediction of electron-impact ionization cross sections of large molecules (using CF_4 as an example) via machine learning.

Published under license by AIP Publishing.

B. Construction of dataset

As described above, there are several theoretical approaches to determining molecular Q_{ion}. Following our previous work,[2,3] the BEB model developed by Kim and Rudd is used in this work to calculate Q_{ion} of small molecules in the train set. A four-step calculation scheme is designed to obtain the high-precision binding energies, kinetic energies, and ionization potential.[3] The chemical structures of the molecules in the train set are first optimized by the hybrid density functional theory (DFT) method with the Austin–Petersson–Frisch functional including dispersion (APF-D). The split-valence triple-zeta correlation-consistent basis set with added diffuse functions, namely, aug-cc-pvtz, is chosen to build the quantum mechanical wave functions. Based on the optimized molecular structures, the binding and kinetic orbital energies of the molecules are calculated with APF-D/aug-cc-pvtz model chemistry. To improve the energy calculation, the electron propagator theory (EPT) with the outer valence Green's function (OVGF) approximation is applied to calculate the binding energies for the molecular orbitals with valence electrons. Lastly, the complete basis set (CBS) method is used to calculate the energies of the molecules and their positively charged ions. The ionization potential is determined as the difference between the energies of a molecule and its cation.

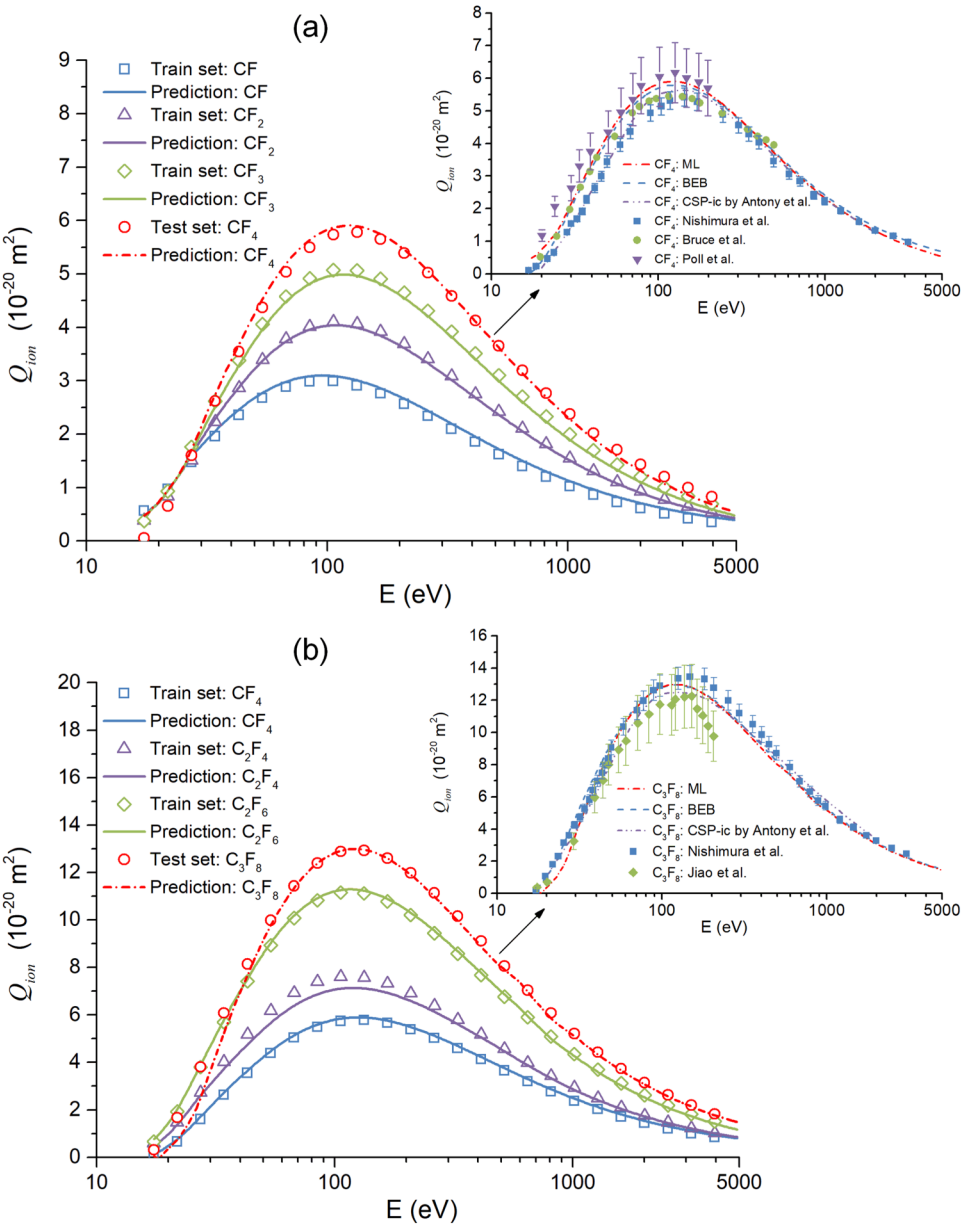

FIG. 4. Prediction of electron-impact ionization cross sections of CF_4 and C_3F_8 by the machine learning (ML) based method and comparison with other theoretical and experimental results. The CSP-ic method was performed by Antony *et al.*[13] The experiments were conducted by Nishimura *et al.*,[14] Bruce *et al.*,[15] Poll *et al.*,[16] and Jiao *et al.*[17]

Published under license by AIP Publishing.

It should be noted that the machine learning based model in this work does not require that the ionization cross sections in the train dataset be calculated by the *ab initio* method. However, due to experimental errors, the experimental data compiled from different sources or measured through different approaches may not be consistent with each other. In order to achieve better prediction performance, the molecular information generated by the *ab initio* method with the same basis set and the Q_{ion} calculated by the same method are recommended to be used to construct the train dataset.

It is known that Q_{ion} of a molecule depends on many parameters, e.g., incident electron energy, kinetic energy, and binding energy. If $\boldsymbol{X}$ denotes these parameters and y the values of Q_{ion}, the theoretical methods for calculating Q_{ion}, such as the BEB model, give analytic expressions $f(\boldsymbol{X}) = y$ which describe the relationship between the above parameters and the corresponding Q_{ion}. However, the machine learning based method does not give analytic expressions but learns the similar relationship between $\boldsymbol{X}$ and y. This functional relationship $\boldsymbol{X} \rightarrow y$ is learned from the data of the molecules in the train set. Such a learning process is also called training. As shown in Fig. 3, $\boldsymbol{X}$ and y are usually presented in the form of matrix and vectors, respectively. After the training, a good machine learning model has the ability to predict Q_{ion} of molecules which are not in the train set. This ability is known as generalization performance.

In order to prepare $\boldsymbol{X}$ and y, Q_{ion} calculated in part 1 of the calculation procedure (as described in Fig. 2) are first preprocessed as required by the machine learning model shown in Fig. 3. 500 points of the incident electron energy from an ionization potential up to 5000 eV are sampled in the form of logarithmic increment. The sampling of ionization cross sections in the logarithmic coordinate is more uniform than that in the linear coordinate, which will improve the model performance in the later training and predicting procedures. For each molecule, these 500 incident electron energies are taken as the first attribute in the training vector, i.e., the first column in the matrix $\boldsymbol{X}$. This is natural because Q_{ion} of a molecule is dependent on the incident electron energy. However, this attribute is not enough to feature Q_{ion} of different molecules. We found that the number of electrons a molecule possesses can be taken as another attribute in the training vector, i.e., the second column in the matrix $\boldsymbol{X}$. Consequently, after the data preprocessing, a 500×2 matrix $\boldsymbol{X}$ for attribute values and a 500×1 vector y for Q_{ion} are constructed for each molecule. According to our practical experience, the presentation of Q_{ion} in the unit of $10^{-20}\,\mathrm{m}^2$ and the reduction of electron numbers by a factor of 0.01 are recommended in the construction of the dataset, because this makes an optimal numeric balance between each attribute. After that, a machine learning algorithm is designed to learn and estimate automatically the unknown regression functional relationship between attribute matrix $\boldsymbol{X}$ and target values y. In this work, a support vector machine (SVM) is used.

C. Parameters of the machine learning algorithm

SVM is a popular machine learning algorithm for classification, regression, and other learning tasks.[11] By comparison with artificial neutral networks (ANNs), SVM has solid theoretical foundations based on the Vapnik–Chervonenkis (VC) theory, and its generalization performance does not depend on the dimensionality of the input space.[12] However, the quality of the SVM algorithm depends on a proper setting of SVM hyperparameters, e.g., penalty parameter C and kernel function parameters. The main issue for applying SVM is how to perform parameter tuning for a given dataset. In our work, the radial basis function (RBF) is set as a kernel function $K(x_i, x_j)$ to map data to a higher dimensional space,[11]

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \tag{1}$$

where x_i and x_j are training vectors, and γ is the kernel parameter.

There are two hyperparameters for SVM with RBF kernel: C and γ. The performance of SVM is strongly affected by these two parameters. It is not known beforehand which C and γ are best for our model, and consequently, some kind of parameter search must be done. Following the recommendation by Chang and Lin,[11] we adopt a straightforward strategy, i.e., grid search for the parameter tuning of C and γ. Going by practical experience, C and γ are gridded as exponentially growing sequences (i.e., 2^0 to 2^7 for C and 2^{-7} to 2^0 for γ in this work) and each pair of (C, γ) will be tried. To prevent the overfitting problem, a 5-fold cross-validation procedure is used in the grid search. Cross-validation is one of the validation techniques for assessing how the results of a statistical analysis will generalize to an independent dataset. One round of cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the training set) and validating the analysis on the other subset (called the validation set). In our work, the training data are divided into 5 subsets of equal size. Sequentially, one subset is tested using the model trained on the remaining 4 subsets. Each instance of the whole training set is predicted once. The average prediction accuracy is taken as the performance measurement of the model trained using the given (C, γ). Various pairs of (C, γ) are tried and the one with the best cross-validation accuracy is picked as the best (C, γ).

III. VALIDATION OF THE METHOD

Finally, the SVM model with the best hyperparameters is used to predict molecular Q_{ion} both in the train and in the test sets at given incident electron energies. All the calculations concerning SVM are implemented using the open source Python library scikit-learn. The machine learning model is constructed based on the data in the train set. A good learning model should present good performance not only on a train set but also, and more importantly, on a test set.

TABLE I. Mean squared error (MSE) and root mean squared error (RMSE) between machine learning estimates and the calculated values of Q_{ion} by the BEB method for CF_4, C_3F_8, SF_6, C_6, C_6F_{12}, and $C_6F_{12}O$.

	CF_4	C_3F_8	SF_6	C_6	C_6F_{12}	$C_6F_{12}O$
MSE ($\times 10^{-40}\,\mathrm{m}^4$)	0.018	0.15	0.014	0.14	0.027	0.056
RMSE ($\times 10^{-20}\,\mathrm{m}^2$)	0.13	0.39	0.12	0.37	0.16	0.24

Published under license by AIP Publishing.

As shown in Fig. 4(a), our model predicts good values of Q_{ion} for CF_4 in the test set in comparison with the theoretical values by the present BEB method and the complex spherical potential-ionization contribution (CSP-ic) method.[13] The mean squared error (MSE) and the root mean squared error (RMSE) between SVM estimates and the calculated values of Q_{ion} by the BEB method for CF_4 are 0.018 ($\times 10^{-40}$ m^4) and 0.13 ($\times 10^{-20}$ m^2), respectively, as listed in Table I. We also compare the machine learning prediction for the

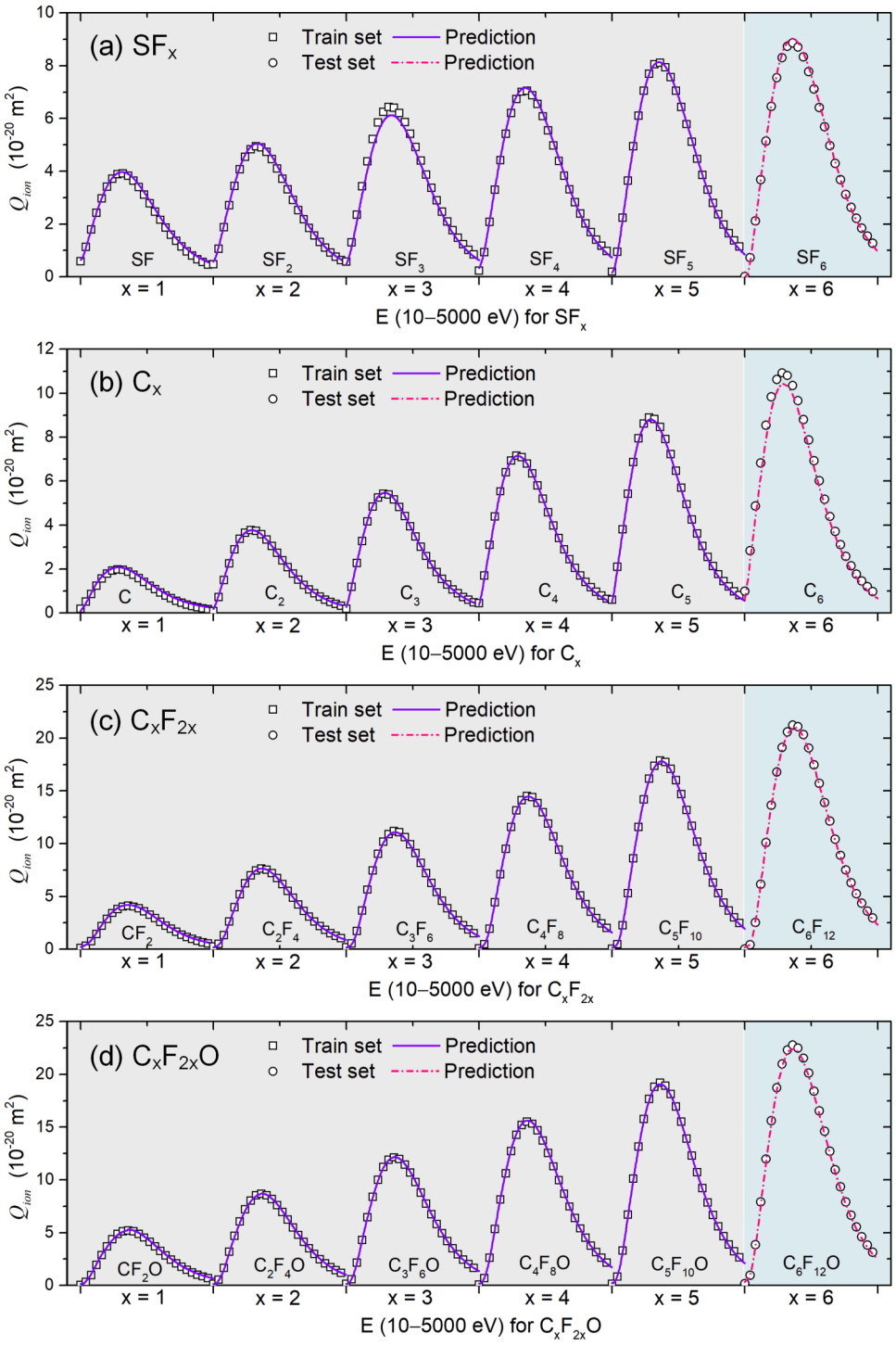

FIG. 5. Prediction of electron-impact ionization cross sections of (a) SF_x, (b) C_x, (c) C_xF_{2x}, and (d) $C_xF_{2x}O$ (x = 1–6) by the machine learning based method. For each molecule, the incident electron energy E is discrete in logarithmic increment from the ionization potential up to 5000 eV.

Published under license by AIP Publishing.

Q_{ion} of CF_4 with the experimental results by Nishimura *et al.*,[14] Bruce and Bonham,[15] and Poll *et al.*[16] The discrepancy is observed in the low electron energy range, which, however, is acceptable, considering the uncertainty of the experiments.

In Fig. 4(b), Q_{ion} of C_3F_8 is predicted by learning the information from the train set composed of CF_4, C_2F_4, and C_2F_6. It can be seen that the machine learning prediction for C_3F_8 agrees well with the results calculated by the BEB and CSP-ic methods[13] and the results measured by Nishimura *et al.*[14] and Jiao *et al.*[17] In general, Q_{ion} predicted by the machine learning method is closer to the results by the BEB model than the one predicted by the other methods. This is because the machine learning based model is trained by inputting the data in the train set in which all the Q_{ion} are calculated by the BEB method in this work. This also means that the training model learns some knowledge or information that the BEB model has.

IV. GENERALIZATION OF THE METHOD

In order to check the generalization of the prediction method described above, we apply this method to predict Q_{ion} of four large molecules SF_6, C_6, C_6F_{12}, and $C_6F_{12}O$.

For each large molecule, the small molecules with fewer constituent atoms including SF_x, C_x, C_xF_{2x}, and $C_xF_{2x}O$ (x = 1–5) are put in each corresponding train set. SF_6, C_6, C_6F_{12}, and $C_6F_{12}O$ are put in each test set accordingly. The BEB method is used to calculate Q_{ion} of the molecules in the train sets based on the results of the quantum chemical computation. The grid search with 5-fold cross-validation is also performed to find optimal hyperparameters (C, γ). Figure 5 shows the comparison of the estimates and the calculated values of Q_{ion} by the BEB method on both train and test sets. The prediction performance on the test sets is generally good as observed from both Fig. 5 and Table I.

V. SUMMARY

In summary, the data-driven method we present in this work demonstrates the power of machine learning for the prediction of molecular Q_{ion}. The model learns from the calculated Q_{ion} and electron numbers of small molecules and, thus, has the ability to predict Q_{ion} of large molecules without the heavy-cost *ab initio* computation for these large molecules. The prediction on the test sets shows that this data-driven model has good generalization performance. This means that by constructing a proper train set, this method can be easily applied for the prediction of unavailable Q_{ion} for other large molecules.

ACKNOWLEDGMENTS

This work was supported in part by the Natural Science Foundation of Jiangsu Province (Grant No. BK20180387).

REFERENCES

[1]K. Becker and V. Tarnovsky, Plasma Sources Sci. Technol. **4**(2), 307 (1995).
[2]L. Zhong, X. Wang, and M. Rong, Phys. Plasmas **25**(10), 103507 (2018).
[3]L. Zhong, J. Wang, X. Wang, and M. Rong, Plasma Sources Sci. Technol. **27**(9), 095005 (2018).
[4]M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, Williams F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, Gaussian 16, Revision B.01, Gaussian, Wallingford, CT, 2016.
[5]D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, and A. G. Doyle, Science **360**(6385), 186–190 (2018).
[6]M. I. Jordan and T. M. Mitchell, Science **349**(6245), 255–260 (2015).
[7]E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder, and E. Olivetti, Chem. Mater. **29**(21), 9436–9444 (2017).
[8]M. H. S. Segler, M. Preuss, and M. P. Waller, Nature **555**(7698), 604–610 (2018).
[9]H. Deutsch, K. Becker, S. Matt, and T. Märk, Int. J. Mass Spectrom. **197**(1), 37–69 (2000).
[10]Y.-K. Kim and M. E. Rudd, Phys. Rev. A **50**(5), 3954–3967 (1994).
[11]C. C. Chang and C. J. Lin, ACM Trans. Intell. Syst. Technol. **2**(3), 27 (2011).
[12]V. Cherkassky and Y. Ma, Neural Netw. **17**(1), 113–126 (2004).
[13]B. K. Antony, K. N. Joshipura, and N. J. Mason, J. Phys. B At. Mol. Opt. Phys. **38**(3), 189–205 (2005).
[14]H. Nishimura, W. M. Huo, M. A. Ali, and Y. K. Kim, J. Chem. Phys. **110**(8), 3811–3822 (1999).
[15]M. R. Bruce and R. A. Bonham, Int. J. Mass Spectrom. Ion Process. **123**(2), 97–100 (1993).
[16]H. U. Poll, C. Winkler, D. Margreiter, V. Grill, and T. D. Märk, Int. J. Mass Spectrom. Ion Process. **112**(1), 1–17 (1992).
[17]C. Q. Jiao, A. Garscadden, and P. D. Haaland, Chem. Phys. Lett. **325**(1–3), 203–211 (2000).

Published under license by AIP Publishing.