ACS EARTH AND SPACE CHEMISTRY

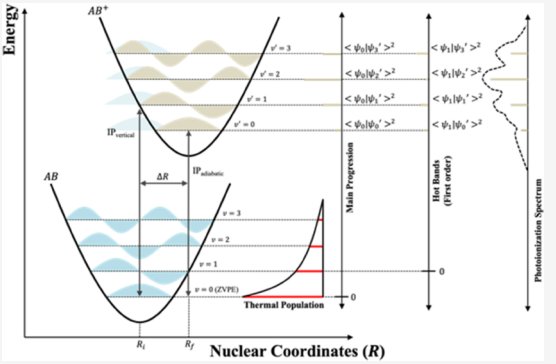
http://pubs.acs.org/journal/aesccq
Article

Machine Learning for Ionization Potentials and Photoionization Cross Sections of Volatile Organic Compounds

Matthew P. Stewart and Scot T. Martin*

Cite This: ACS Earth Space Chem. 2023, 7, 863−875
Read Online

ABSTRACT: Molecular ionization potentials (IP) and photoionization cross sections (σ) can affect the sensitivity of photoionization detectors (PIDs) and other sensors for gaseous species. This study employs several methods of machine learning (ML) to predict IP and σ values at 10.6 eV (117 nm) for a dataset of 1251 gaseous organic species. The explicitness of the treatment of the species electronic structure progressively increases among the methods. The study compares the ML predictions of the IP and σ values to those obtained by quantum chemical calculations. The ML predictions are comparable in performance to those of the quantum calculations when evaluated against measurements. Pretraining further reduces the mean absolute errors (ε) compared to the measurements. The graph-based attentive fingerprint model was most accurate, for which ε_{IP} = 0.23 ± 0.01 eV and ε_{σ} = 2.8 ± 0.2 Mb compared to measurements and computed cross sections, respectively. The ML predictions for IP correlate well with both the measured IPs (R^2 = 0.88) and with IPs computed at the level of M06-2X/aug-cc-pVTZ (R^2 = 0.82). The ML predictions for σ correlated reasonably well with computed cross sections (R^2 = 0.66). The developed ML methods for IP and σ values, representing the properties of a generalizable set of volatile organic compounds (VOCs) relevant to industrial applications and atmospheric chemistry, can be used to quantitatively describe the species-dependent sensitivity of chemical sensors that use ionizing radiation as part of the sensing mechanism, such as photoionization detectors.



KEYWORDS: *ionization potential, photoionization cross section, machine learning, photoionization detector, volatile organic compounds*

1. INTRODUCTION

Ionization potential (IP; unit, eV = electronvolt; 1 eV = 1.6 × 10^{-19} J) and photoionization cross section (σ; unit, Mb = megabarn; 1 Mb = 10^{-18} cm^2) affect the quantitative response of some optical sensors to species concentration. Volatile organic compounds (VOCs), a large and diverse group of chemical species important to atmospheric chemistry, have concentrations that span several orders of magnitude.[1] Although a photoionization detector (PID) can sense a broad array of VOCs, its response remains qualitative without knowledge of molecular IP and σ values. Quantum chemical methods exist for calculating the IP and σ values of VOCs, but the computational cost becomes prohibitive for large molecules and basis sets.[2]

In recent years, machine learning (ML) techniques have facilitated the application of quantum chemical methods for predicting molecular properties to address this computational challenge. Quantitative structure−property relationships (QSPR) derived by machine learning (ML) have screened properties of prospective drug molecules,[3,4] fuels,[5,6] and polymers.[7,8] These methods implicitly approximate quantum chemical calculations by a coarse-grained basis set.[9] As such, the methods have low accuracy for predicting properties that depend on electronic structure, which is problematic because IP and σ are both highly sensitive to electronic structure. The ML-based QSPRs thus do not predict IP and σ accurately enough for practical use. More explicit ML methods developed using deep learning approaches, such as Chemception[10] and graph-based neural networks (GNNs),[11] incorporate spatial dependencies representing the bonding among atoms. In this way, they can incorporate quantum chemical elements that approach the accuracy of direct quantum chemistry calculations.

Accurate predictions of IP and σ values across a broad range of VOCs are a baseline need for employing PIDs to quantify atmospheric VOC concentrations. PIDs have high sensitivity, broad selectivity, and fast response time. However, when used outside of a laboratory environment consisting of individual and known gaseous species, the sensor output can be qualitative rather than quantitative due to unknown species-dependent responses. These responses depend primarily on the IP and σ

Received: January 11, 2023
Revised: March 16, 2023
Accepted: March 17, 2023
Published: April 6, 2023

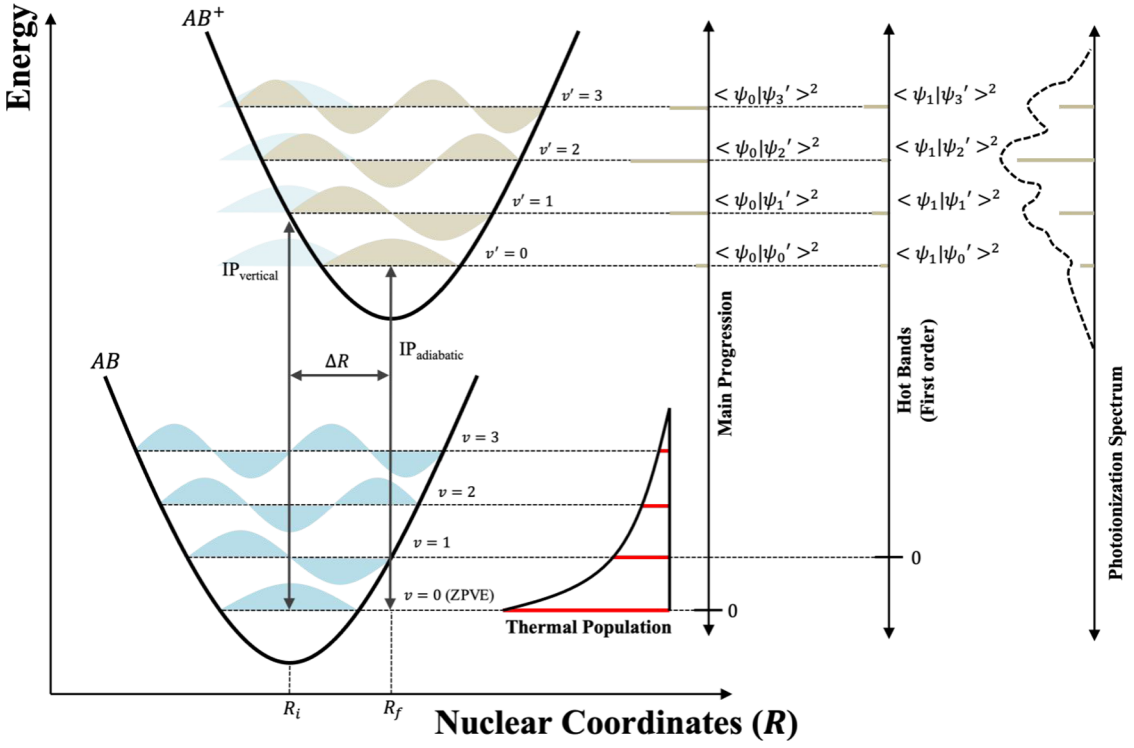


Figure 1. Diagram of quantum energy levels and the transitions among them. Nuclear position coordinate (R) varies along the abscissa. Energy varies along the ordinate. Vertical and adiabatic transitions associated with the ionization potential are indicated. The adiabatic IP accounts for the relaxation of the nuclear coordinates in the excited state ($R_i \rightarrow R_f$), leading to a lower energy than that of the vertical IP. Terms AB and AB^+ refer to the molecular and ionized forms of a species. Quantum number (ν) refers to a vibrational state. The vibrational component of the wavefunction is denoted by an integer subscript, and the electronic component by the prime symbol. Other symbols are defined in the main text. The extent of wavefunction overlap between the ground and excited state is denoted by the overlap of shaded regions at each vibrational state. An example of a resulting photoionization spectrum is shown on the right-hand side. It represents the overlap integral of wavefunctions (ψ) in correspondence to the transitions of the main progression as well as the hot bands. From this spectrum, the value of the photoionization cross section (σ) at a photon energy can be extracted.

values of the molecules, as follows. The photon energy used in the sensor must exceed the ionization potential of the molecule, and the photoionization cross section of the molecule at that photon energy proportionally affects the sensor output. Due to the many different types of molecules produced through atmospheric VOC oxidation mechanisms, determining IP and σ by laboratory measurements of each gas species is not feasible. The development of accurate prediction methods is thus needed.

Herein, ML-based approaches are applied for predicting the IP and σ values of atmospherically relevant VOCs at a photon energy of 10.6 eV (117 nm). The most widely employed commercial photoionization detectors operate at this energy. A dataset of 1251 IP values measured in the laboratory using photoelectron spectroscopy or photoionization mass spectrometry was used in the analysis. For the associated molecular species, quantum chemical calculations were used to estimate IP values, which were compared to the measurements. Quantum chemical calculations were also used to estimate the σ values for the group of species at the specified photon energy. The empirical IP values and the quantum-computed IP and σ values served as datasets to train multiple ML methods, including descriptor-based, sequence-based, graph-based, and quantum-based approaches. The ML predictions, the quantum chemical calculations, and the measurements across the entire group of molecules are intercompared and discussed in the analysis.

2. EXPERIMENTAL METHOD

2.1. Quantum Chemical Calculations. *2.1.1. Ionization Potential.* An ionization potential is the minimum incident photon energy required to expel an electron from a molecular orbital to vacuum, leaving behind a free ion. The lowest-lying ionization potential, which is the focus of this study, occurs from the highest occupied molecular orbital (HOMO). The physics of the interactions for the photoionization of gaseous matter is through the coupling of electromagnetic radiation with molecular electric dipoles.[12]

Photoionization is a multistage process: excitation from the ground state to an excited state (eq 1), release of a photoelectron (eq 2), and vibrational relaxation of the excited state (eq 3). The process can be described as follows

$$AB + h\nu \rightarrow AB^{\ddagger} \quad (1)$$

$$AB^{\ddagger} \rightarrow AB^{+,\ddagger} + e^- \quad (2)$$

$$AB^{+,\ddagger} \rightarrow AB^{+} \quad (3)$$

In these equations, AB indicates the molecule in its electronic and vibrational ground state, $h\nu$ refers to the incident photon, $AB^{\ddagger}$ denotes the molecular species in an excited state, $AB^{+,\ddagger}$ represents the molecular cation in an excited state, e^- is the ejected electron, and AB^+ corresponds to the ionized species. The kinetic energy of the ejected electron is equal to the incident photon energy minus the ionization potential.

Photoionization can occur adiabatically or vertically, as depicted in Figure 1. This distinction relates to the timescales

https://doi.org/10.1021/acsearthspacechem.3c00009
ACS Earth Space Chem. 2023, 7, 863−875

and net energy changes of the physical processes. Photon absorption (eq 1, $\tau \sim 10^{-15}$ s) and electron emission occur on a faster timescale (eq 2, $\tau \sim 10^{-17}$ s) than the vibrational relaxation of the molecular ion (eq 3, $\tau \sim 10^{-13}$ s). Adiabatic ionization corresponds to vibrational relaxation of the nuclear coordinates from the excited state. Thus, the adiabatic IP is the energy difference between zero-point levels of the ionized and neutral species, including the electronic ($E_{\text{electronic}}$) and zero-point vibrational energies (ZPVE). This energy corresponds to the energy of AB^+ of eq 3 minus the energy of AB of eq 1. By comparison, vertical ionization corresponds to the absence of vibrational relaxation in the excited state. This energy corresponds to the energy difference between $AB^{+,\ddagger}$ and AB of eqs 1 and 2.

The adiabatic IP is typically used in theory, whereas vertical IP is typically obtained by measurement. For example, photoelectron spectroscopy indirectly measures IP by considering the kinetic energy of the ejected electron (eq 2). Photoionization mass spectrometry determines IP by focusing on parent ion (eq 2 or eq 3). In this study, we focus on vertical IPs as they are the more relevant experimental quantity, but also compare prediction methods for computing adiabatic IP. The difference between adiabatic and vertical IPs is typically less than 0.4 eV. For both adiabatic and vertical IP, the following relationship holds

$$\text{IP}_{\text{molecule}} = (E_{\text{electronic}} + \text{ZPVE})_{\text{cation}} - (E_{\text{electronic}} + \text{ZPVE})_{\text{neutral}} \quad (4)$$

Herein, IPs measured for 1251 organic species were selected from the Chemistry WebBook of the USA National Institute of Science and Technology (NIST).[13] Measurements behind the dataset included photoionization mass spectrometry, photoelectron spectroscopy, and electron-impact techniques. Only molecules containing nitrogen, oxygen, carbon, and hydrogen were selected. Radicals, ions, and species having more than 20 nonhydrogen atoms were excluded. IPs in the dataset spanned 6.16−13.94 eV. Each molecule was encoded using the Simplified Molecular-Input Line-Entry System (SMILES)[14] and converted to a three-dimensional representation in *Z*-matrix form (Open Babel v3.3.1).[15] Geometries of the neutral molecules and the corresponding cations were optimized using density functional theory (DFT) at three levels: B3LYP/6-31G++(d,p),[16] ωB97XD/def2-TZVP,[17] and M06-2X/aug-cc-pVTZ.[18] Adiabatic and vertical IPs were computed at each level using eq 4. The electronic structure calculations were performed in Gaussian 16.[19]

2.1.2. Photoionization Cross Section. The photoionization cross section relates to the fraction of incoming photons that result in an ionization interaction between a photon and an ionizable molecule. Several theoretical procedures for calculating σ values near the ionization threshold include the frozen-core Hartree−Fock (FCHF) method,[20,21] time-dependent density functional theory,[22] variational methods,[23,24] and methods incorporating Dyson orbitals.[25] The FCHF and Dyson orbital methods provide estimates of σ values at a typical uncertainty of less than a factor of 2.[25,26] In the present study, the FCHF method is used because of its efficiency in calculating near-threshold cross sections for large molecules.[27]

Processes related to IP and σ are depicted in Figure 1. The photoionization cross section represents the transition probability between the initial and final wavefunctions of AB and AB^+. As such, σ values are proportional to the square of the transition dipole moment ($\vec{M}_{I,f}$), as follows

$$|\vec{M}_{\text{i,f}}|^2 = \langle \psi_{\text{i}} | \hat{\boldsymbol{\mu}} | \psi_{\text{f}} \rangle = \left(\int \psi_{\text{i}}(\boldsymbol{r}, \boldsymbol{R}) \hat{\mu} \psi_{\text{f}}(\boldsymbol{r}, \boldsymbol{R}) \, \text{d}\boldsymbol{r} \, \text{d}\boldsymbol{R} \right)^2 \quad (5)$$

In eq 5, the value of $\vec{M}_{\text{i,f}}$ is the integral across the product of the *N*-electron dipole operator $\hat{\boldsymbol{\mu}}$ with the initial and final wavefunctions (ψ_{i} and ψ_{i}, respectively). The wavefunctions depend on the electronic and nuclear coordinates, denoted by $\boldsymbol{r}$ and $\boldsymbol{R}$, respectively. Through the Born−Oppenheimer approximation, the total wavefunction can be separated into electronic (φ) and vibrational (χ) components, allowing separation of $\vec{M}_{\text{i,f}}$ into two terms. These terms are the electronic transition dipole moment, denoted by $D(E)$, and the vibrational Franck−Condon overlap envelope, denoted by the Supporting Information (SI), as follows

$$\sigma \propto |\vec{M}_{\text{i,f}}|^2 = \underbrace{\left(\int \varphi_{\text{i}}(\boldsymbol{r}; \boldsymbol{R}_{\text{e}}) \hat{\boldsymbol{\mu}} \varphi_{\text{f}}(\boldsymbol{r}; \boldsymbol{R}_{\text{e}}) \, \text{d}\boldsymbol{r} \right)^2}_{\text{electronic transition dipole moment}} \underbrace{\left(\int \chi_{\text{i}}(R) \chi_{\text{f}}(R) \, \text{d}\boldsymbol{R} \right)^2}_{\text{Franck–Condon overlap envelope}} = D(E)S(E) \quad (6)$$

where $\boldsymbol{R}_{\text{e}}$ represents the equilibrium nuclear coordinates for calculation oI(E) within the Condon approximation and E indicates the incident photon energy.

The electronic component $D(E)$ was computed using ePolyScat[20,21] within the single-channel FCHF approximation.[28] The single-center expansion technique for $L_{\text{max}} = 50$ was used. Molecular orbitals were computed in Gaussian from the restricted Hartree−Fock wavefunctions with the aug-cc-pVTZ basis set after geometry optimization at the M06-2X/aug-cc-pVTZ level of theory. Symmetry rules governing photoionization used in ePolyScat simulations are described in the Supporting Information. Ultrafine grid quadrature and very tight self-consistent field convergence were used.

For $S(E)$, Franck−Condon factors were computed in ezFCF[29] using normal modes and separable-mode harmonic oscillator assumptions at the M06-2X/aug-cc-pVTZ level of theory. The factors were cumulatively summed and normalized to provide $S(E)$. The value of $S(E)$ was taken as unity when the IP greatly exceeded the photon energy (i.e., ≥1 eV).[26] In this situation, the photon energy is assumed sufficiently high that all available vibrational modes within the molecule are active. In the envelope calculations, resonant states were omitted.

Photoionization cross sections were computed using eq 6 at 10.6 eV for a subset of 724 organic molecules in the NIST dataset for which IP < 10.6 eV. A photon energy of 10.6 eV is typical for a PID. An individual σ value was the average of the length gauge and the velocity gauge, which are two different approximations to describe the interaction of a photon with a molecule.[26] The length gauge assumes that the photon has a purely electric field, while the velocity gauge assumes that the photon has a mixed electric and magnetic field. The absolute difference between the length and velocity gauges was used as an approximate measure of the error in σ. Autoionization and other nonradiative processes, such as photodissociation and internal conversion, were not included in the analysis. Values for σ represented the parent ion only, meaning that the complexity of modeling photo-ion product channels was not incorporated. This simplification, however, did not result in a large error because ion fragmentation is typically not substantial for photon energies within 2 eV of IP.[30,31]

https://doi.org/10.1021/acsearthspacechem.3c00009
ACS Earth Space Chem. 2023, 7, 863−875

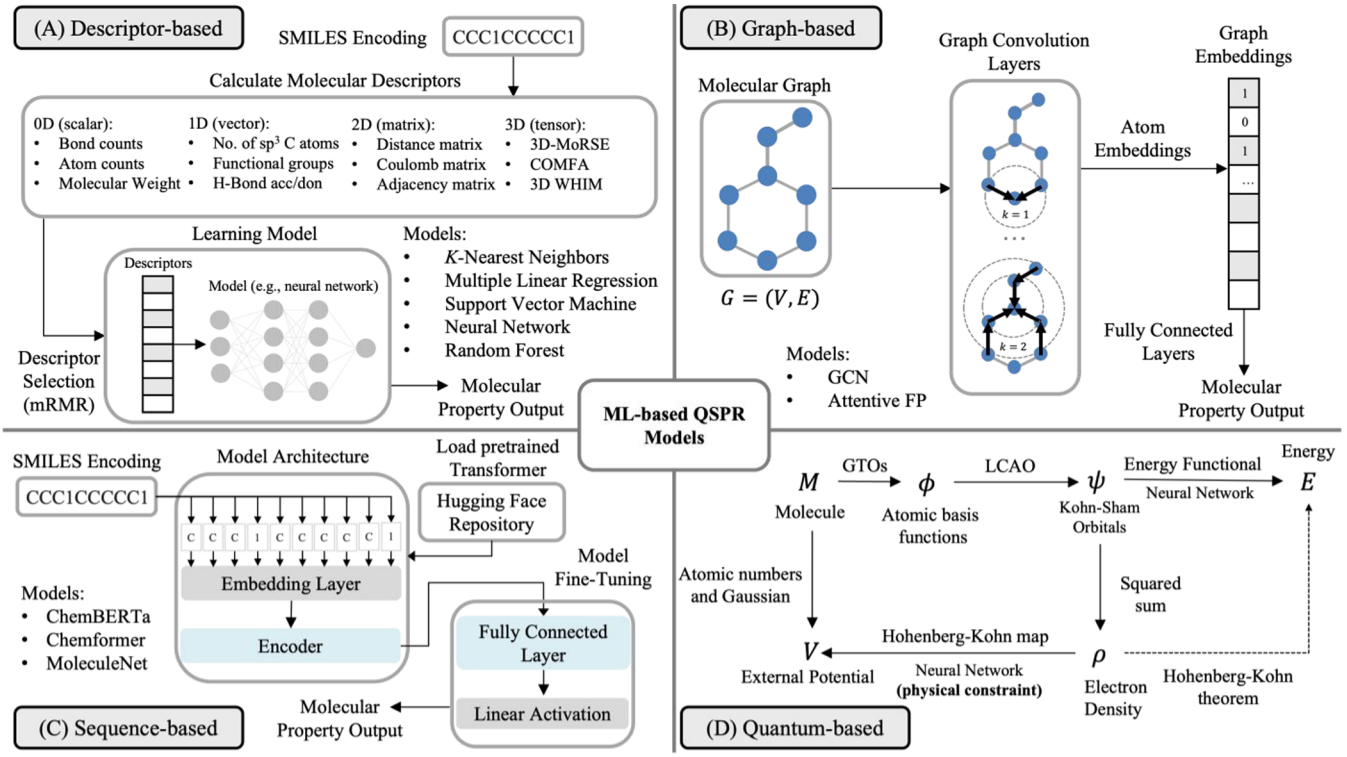


Figure 2. Overview of machine learning methods explored in this study for predicting IP and σ values as output for molecular SMILES strings as input. (A) Descriptor-based approach. (B) Graph-based approach (including AFP model). (C) Sequence-based approach. (D) Quantum-based approach. The main text presents further definitions of abbreviations and symbols as well as associated explanation and discussion.

2.2. Machine Learning Methods. Four different ML methods were used to construct estimative models of IP and σ values as output for chemical structures as input (Figure 2). The four methods employed varying explicitness in their treatment of electronic structure, which influenced the predictive accuracy of each. The measured values from the NIST database were used for training the ML models of IP values. The corresponding ePolyScat simulation results were used to train the ML models of σ values. Each ML model was trained using fivefold cross-validation, with onefold used as the validation set during each iteration, and a separate test set was set aside for the final evaluation of the model performance. Descriptor-based models were developed using RDKit,[32] sequence-based models using Simple Transformers, and graph-based models using DeepChem.[33] The quantum-based models were built using the QDF-NN code of Tsubaki and Mizoguchi.[34] The GNN and QDF-NN methods were augmented using pretraining on energies of the highest occupied molecular orbital (HOMO) from the Quantum Mechanics 9 (QM9) dataset.[35,36] The rationale for the pretraining was based on Koopman's theorem, which asserts that the negative value of the HOMO energy level is the first approximation of ionization potential.[37] This pretraining was applied for both IP and σ values because encodings of the electronic structure are relevant to both quantities. Further details on the training of each model are available in the Supporting Information.[38]

2.2.1. Descriptor-Based ML Method. Descriptor-based QSPR methods use molecular information, such as the number of valence electrons, heteroatoms, rings, stereocenters, molecular weights, and functional groups, as variables in predicting molecular properties (Figure 2). They have been widely used for predicting the physicochemical properties of materials, including melting point,[39] boiling point,[40] flash point,[41] flammability limits,[42] glass-transition temperature,[8] acid dissociation constant,[43] and thermal stability.[44] ML-based QSPR methods are increasingly supplanting traditional QSPRs because of superior accuracy. Descriptor-based methods, however, do not explicitly consider electronic structure.

A set of 2756 molecular descriptors (n = 1875) and fingerprints (n = 881) was used to train descriptor-based methods. The open-source PaDEL software[45] using the PaDELPy Python wrapper was used. The tested algorithms included (1) the random-forest algorithm, (2) deep neural networks, and (3) gradient boosting. These algorithms were chosen due to their strong performance on regression tasks and widespread use in the ML-based QSPR literature. Of the 2756 descriptors, the two most informative were selected for training using an algorithm of maximum relevance and minimum redundancy (mRMR).[46] Hyperparameter tuning was performed using Bayesian optimization for each model.[47] Each method was subsequently trained with optimized hyperparameters, and method performance was intercompared. After testing, the method of gradient boosting had the highest accuracy, and this descriptor-based ML method was therefore used as the baseline model of this study.

2.2.2. Sequence-Based ML Method. Sequence-based ML methods perform well for some molecular properties.[48−51] They couple transformer models[52] with SMILES notation. The architectures, such as Chemformer[48] and ChemBERTa,[49] are adapted from natural language tasks. Representations encoded by neuron weighting are obtained from self-supervised pretraining on large-scale datasets like QM9 (134,000 molecules). No knowledge of molecular properties is needed for this process, meaning that arbitrarily large datasets of known molecular structures can be used. In the iteration process, a representation of molecular bonding becomes encoded by

https://doi.org/10.1021/acsearthspacechem.3c00009
ACS Earth Space Chem. 2023, 7, 863−875

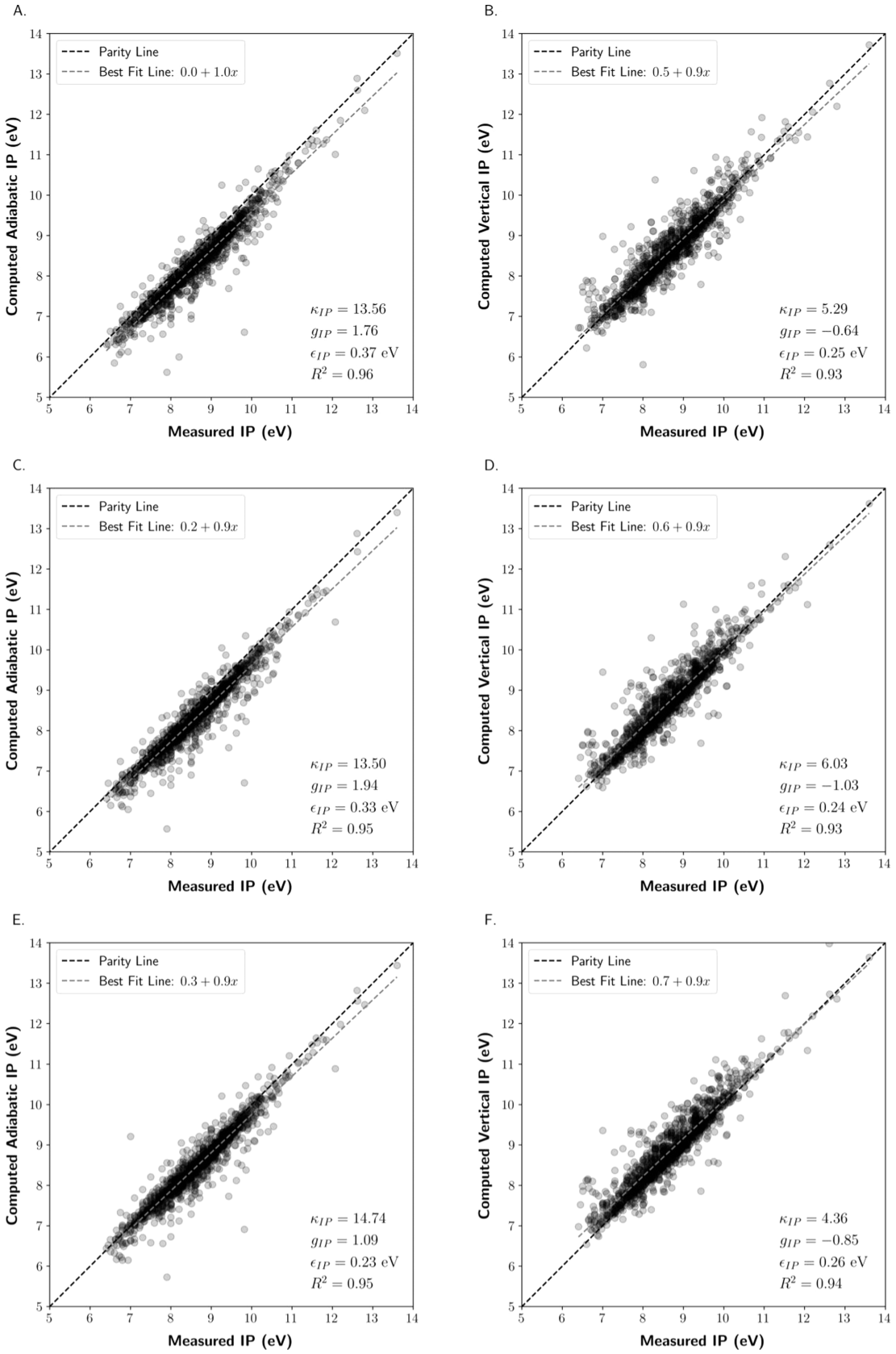


Figure 3. Parity plots between measured and computed adiabatic and vertical ionization potentials at different levels of quantum chemistry theory. (A, B) Adiabatic and vertical IPs for B3LYP/6-311++G(d,p) ($n_{\text{adiabatic}} = 1109$, $n_{\text{vertical}} = 1208$). (C, D) Adiabatic and vertical IPs for ωB97XD/def2-TZVP ($n_{\text{adiabatic}} = 1157$, $n_{\text{vertical}} = 1182$). (E, F) Adiabatic and vertical IP for M06-2X/aug-cc-pVTZ ($n_{\text{adiabatic}} = 1206$, $n_{\text{vertical}} = 1219$). Parity and best-fit lines are plotted. The mean absolute error (ϵ) and the coefficient of determination (R^2) are listed in each panel. Gray shading of data points is for visual illustration and has no other meaning. Values of n denote the number of molecules included in the panels. The values differ among panels because of convergence issues associated with the different levels of quantum chemistry theory.

https://doi.org/10.1021/acsearthspacechem.3c00009
ACS Earth Space Chem. 2023, 7, 863−875

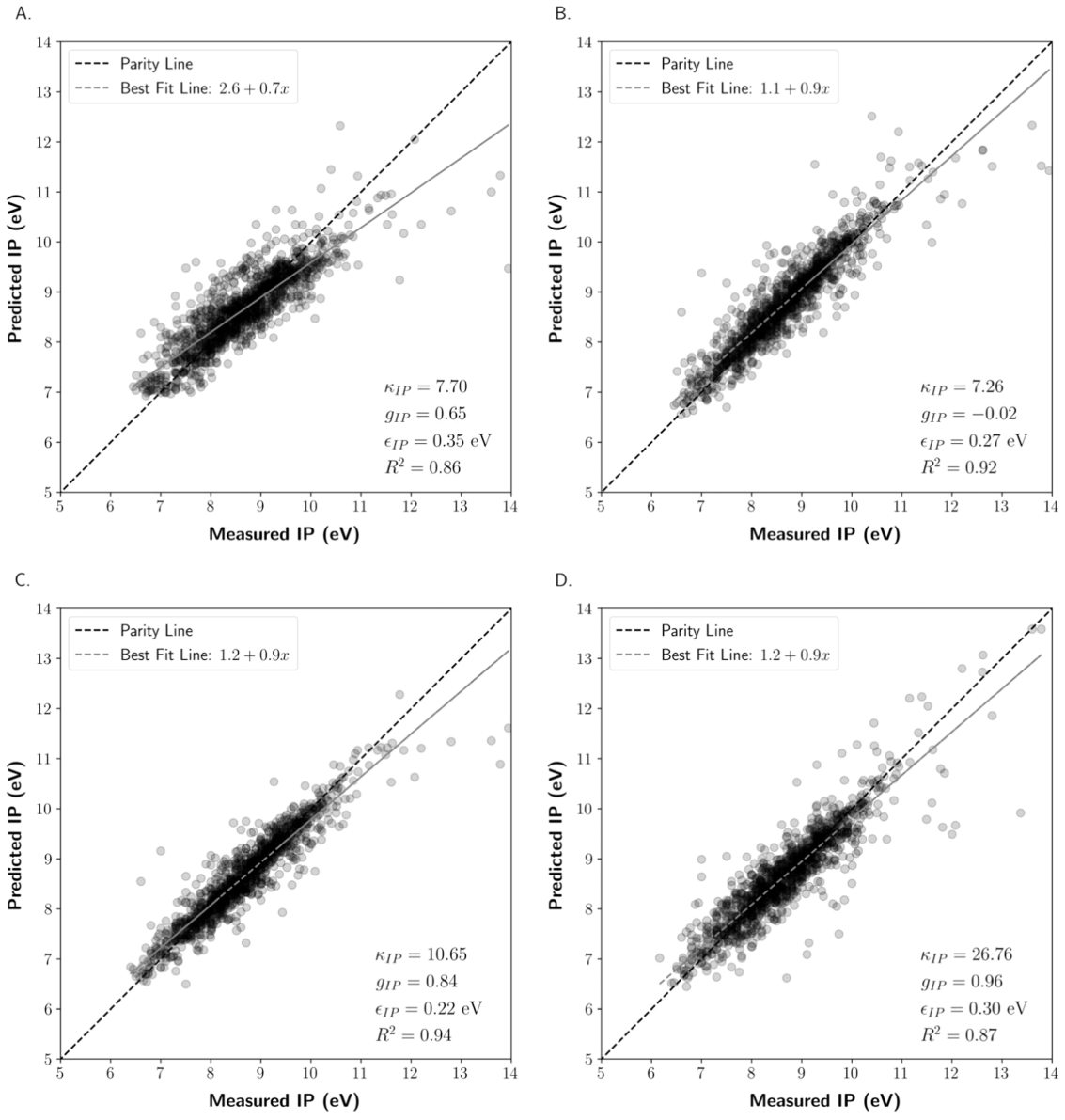


Figure 4. Parity plots between predicted and measured IP values (n = 1249) for the model of lowest mean absolute error for each ML method. (A) Descriptor-based approach. (B) Sequence-based approach. (C) Graph-based approach. (D) Quantum-based approach. Predicted IP values include the combined test set across all folds of cross-validation.

neuron weighting. In an approach similar to pretraining, the resulting encoded representation is then further adjusted to smaller datasets (e.g., typically 1000 molecules) for greater accuracy in a variety of downstream tasks.

For the present study, SMILES representations and encodings for neuron weighting were used from the ChemBERTa model,[49] as obtained from the HuggingFace model repository,[53] which had been trained on the PubChem 10M dataset.[54] The resulting setup was trained on IP and σ values for 50 iterations using a learning rate of 5×10^{-4}. Details regarding model execution and the hyperparameters are stored in a repository, for which information is provided at end of the main text.[38]

2.2.3. Graph-Based ML Method. Graph-based neural networks (GNNs) can predict molecular properties for non-Euclidean data.[55] Several studies demonstrated more accurate results over descriptor-based methods, although other studies reached the opposite conclusion.[56] GNNs generate a molecular graph that defines the connectivity between a set of nodes (i.e., atoms) and edges (i.e., bonds). GNNs aggregate information about neighboring atoms through recursion across the molecular graph. Two GNN methods were implemented in this study, including a graph convolutional network (GCN)[57] and attentive fingerprints (AFP).[58]

The GCN is the graph-based analogue of a standard convolutional neural network. It utilizes Duvenaud graph convolutions.[59] A baseline GCN model was created consisting of a single hidden layer of 512 neurons, a dense layer of 1024 neurons, and a single-neuron output layer representing a linear activation function. The baseline model was trained for 100 iterations at a learning rate of 10^{-4}. A second model utilized pretraining to improve model performance. The pretraining was based on the QM9 HOMO values on the same neural architecture at a learning rate of 10^{-4} for 50 iterations. The weights of the first two layers were subsequently fixed, and the network was trained on the IP and σ values for 50 iterations at a learning rate of 10^{-5}.

https://doi.org/10.1021/acsearthspacechem.3c00009
ACS Earth Space Chem. 2023, 7, 863−875

AFP models have superior performance over GCNs in some applications.[56,58] The AFP mechanism of graph attention allows nonlocal effects to be incorporated at the intramolecular level.[54] Herein, an AFP model was trained using the PyTorch framework[60] and Adam optimization[61] by gradient descent. Optimal hyperparameters were selected by a Bayesian approach. The model was trained by minimizing the mean squared error, and it was evaluated for the mean absolute error on the test set.

2.2.4. Quantum-Based ML Method. A quantum-based deep-field neural network (QDF-NN)[34] is a physics-informed approach to predict molecular properties based on Kohn–Sham DFT.[62] Neural networks simultaneously approximate both the energy functional map and the Hohenberg–Kohn (HK) map. The HK map is indirectly constructed in a self-supervised manner. The energy functional map is externally defined based on the IP and σ values. This two-pronged approach allows the network to approximate basis functions more closely than the previously discussed models.[63]

The QDF-NN models for IP and σ values were configured using a 6-31G basis set on Gaussian-type orbitals for a radius of 0.75 and a grid interval of 0.3. The neural architecture consisted of an input layer and four hidden layers, each having a dimensionality of 500. Each model was trained using a batch size of 4, a learning rate of 10^{-4}, a learning rate decay of 0.5 occurs at every 100 iterations, and a total of 1000 iterations. In the final layer, the output variables were averaged. The approach also included pretraining. HOMO values from the QM9 dataset were used for 1000 iterations. The weights for the IP and σ networks were then pretrained using the weights of the QM9 model. Subsequently, the first two network layers were fixed, and the training took place.

3. RESULTS AND DISCUSSION

3.1. Quantum Chemical Calculations Compared to IP Measurements. In Figure 3, the measured IPs from the NIST database are compared with the adiabatic and vertical IPs computed at three increasingly complex levels of quantum theory.[64,65] These levels included B3LYP/6-311++G(d,p), ωB97XD/def2-TZVP, and M06-2X/aug-cc-pVTZ. For adiabatic IPs, the most accurate results were obtained by using M06-2X/aug-cc-pVTZ. A coefficient of determination (R^2) of 0.95 and a mean absolute error (ε) of 0.23 eV were achieved. For vertical IPs, the most accurate results ($R^2 = 0.93$ and $\varepsilon = 0.24$ eV) were obtained with ωB97XD/def2-TZVP even though this basis set was smaller than that of M06-2X/aug-cc-pVTZ. For DFT, increasing the basis set does not necessarily improve the predictive performance.[66] While IP estimates obtained at a higher level of theory would be generally more accurate, it would be impractical for a dataset of the size used in the present study. Given the 0.4 eV uncertainty between adiabatic and vertical IP, methods with errors that fall within this range were deemed acceptable. Values of skew (g) and kurtosis (κ) are also listed in the figure. The positive skew for the adiabatic IPs indicated a bias toward larger errors. Conversely, the negative skew for the vertical IPs had a bias toward smaller errors. The skew for vertical IPs was less than that of adiabatic IPs for all three basis sets. Similarly, the kurtosis was larger for adiabatic IPs than for vertical IPs, meaning that the distribution of adiabatic IPs had more outliers far from the mean.

3.2. Machine Learning Methods Compared to IP Measurements. In Figure 4, the measured IPs from the NIST database are compared with the predicted IPs in a series of parity plots for descriptor-based, sequence-based, graph-based, and quantum-based ML methods. Results are summarized in Table 1. The mean absolute error approximately follows

Table 1. Calculated and Measured Ionization Potentials, Photoionization Cross Sections, and PID Response Factors for Various Organic Compounds

Molecule	Point Group	IP_{meas} (eV)[a]	IP_{calc} (eV)[b]	σ_{meas} (Mb)	σ_{calc} (Mb)[c]	Response Factor
isoprene	C_s	8.86	8.65	13.0[d]	14.7	1.67
acetaldehyde	C_s	10.23	10.21	7.5[d]	7.39	6
ethanol	C_s	10.47	10.32	2.1	2.45	11
benzaldehyde	C_s	9.49	9.65	18.2[e]	16.6	0.7
furan	C_{2v}	8.88	8.83	15[f]	14.5	0.4
dimethyl ether	C_{2v}	10.03	9.93	8.8[g]	8.12	1.3
isopropanol	C_1	10.17	10.09	5[d]	4.54	4
1,3-butadiene	C_2	9.07	8.84	13.2[d]	13.2	0.8
p-xylene	C_2	8.44	8.38	17.8[d]	14.6	0.55
2-propanol	C_1	10.17	10.09	6.5[g]	7.9	4
cyclopentene	C_s	9.01	8.80	13[g]	15.2	1.5

[a]Ionization potentials obtained from the NIST Chemical WebBook.[13] [b]IPs calculated at the M06-2X/aug-cc-pVTZ level of theory. [c]Calculated with ePolyScat using $L_{max} = 50$ at the M06-2X/aug-cc-pVTZ level of theory. [d]Adam and Zimmermann.[78] [e]Eschner and Zimmermann.[79] [f]Czekner et al.[80] [g]Cool et al.[31]

sequentially from worse to better along the model sequence from descriptor-based to quantum-based ML methods. The ε_{IP} of the ML models ranged from 0.22 to 0.35 eV relative to the NIST database, which can be compared to 0.23–0.37 eV for quantum chemistry calculations.

As the simplest and baseline method of this study, the descriptor-based random-forest mRMR model had $\varepsilon_{IP} = 0.33 \pm 0.02$ eV (Figure 4A). This higher error reflects the absence of molecular structure and connectivity in this treatment. Nevertheless, the descriptors allowed physical insight into the important molecular features that affected the predicted results. In this regard, the values for feature importance of 20 descriptors selected by the mRMR model are shown in Figure 5. Increasing the number of descriptors beyond 20 did not improve model accuracy. A full list of descriptors is provided in Table 3. For IP, the most informative descriptors were BCUTp-1h (0.26) and BCUTp-1l (0.15), which were linked to molecular polarizability.

The sequence-based model had $\varepsilon_{IP} = 0.26 \pm 0.01$ eV (Figure 4B), representing a 24% reduction in error compared to the baseline descriptor-based model. This result is consistent with the hypothesis that more explicit electronic modeling leads to improved performance. The graph-based GCN model initially performed poorly ($\varepsilon_{IP} = 0.39 \pm 0.02$ eV), but pretraining improved performance ($\varepsilon_{IP} = 0.32 \pm 0.02$ eV). Even so, this performance exceeded neither the sequence-based models nor the baseline descriptor-based model. The reduced performance of GCN graph-based networks over descriptor-based methods was previously reported.[56,67,68]

By comparison, the graph-based AFP model had the lowest error of $\varepsilon_{IP} = 0.23 \pm 0.01$ eV among all models (Figure 4C). The strong performance of the AFP model can be explained by its simultaneous consideration of descriptors, connectivity, and nonlocal effects. The quantum-based QDF-NN model had $\varepsilon_{IP} = 0.33 \pm 0.01$ eV without pretraining and $\varepsilon_{IP} = 0.30 \pm 0.02$ eV with pretraining. The AFP model performed significantly better than the QDF-NN model despite the more explicit treatment of electronic structure in the latter. Although pretraining improved the performance of the QDF-NN model, the AFP model still

https://doi.org/10.1021/acsearthspacechem.3c00009
ACS Earth Space Chem. 2023, 7, 863–875

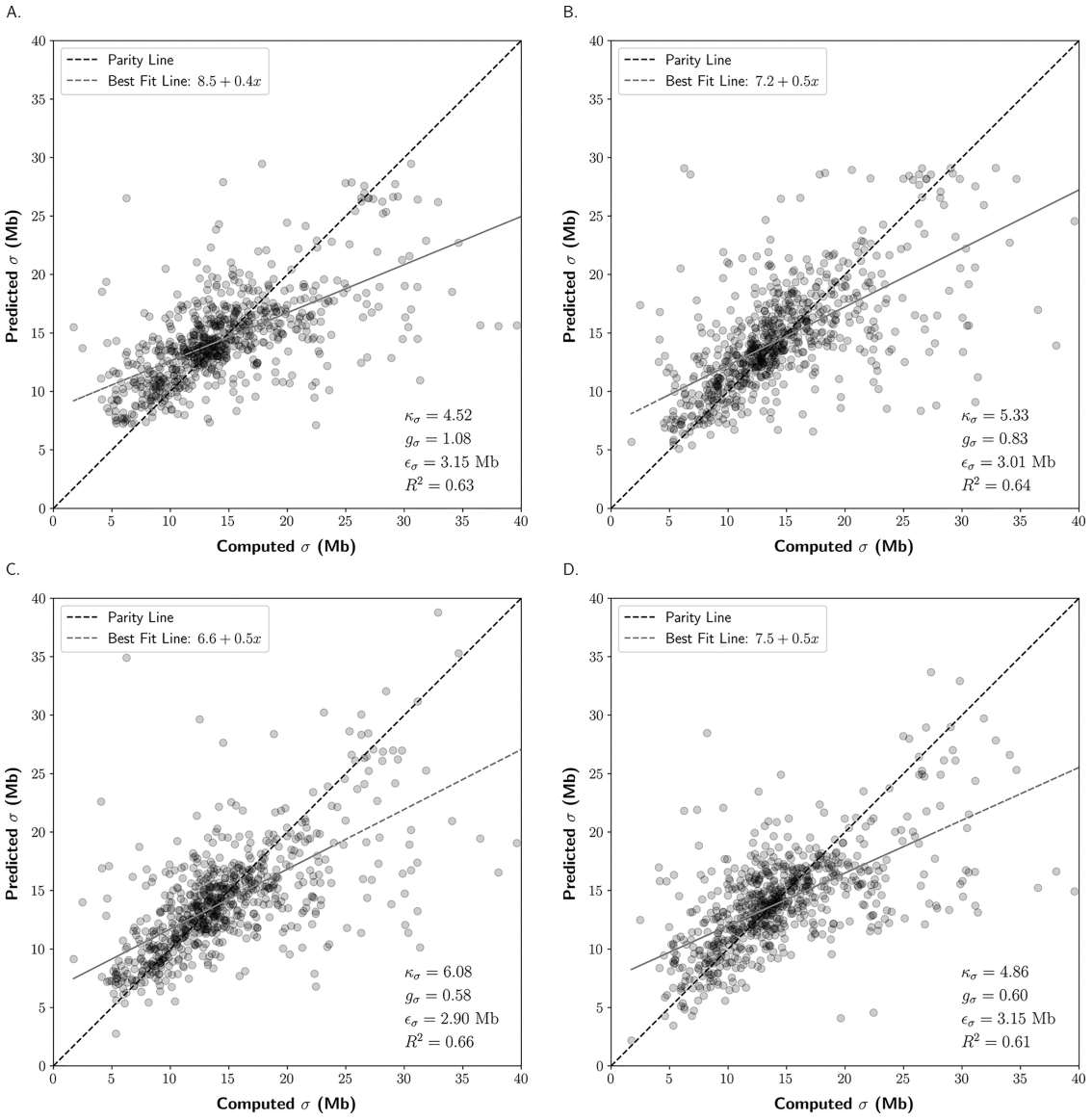


Figure 5. Parity plots between predicted and computed σ values (n = 724) for the model of lowest mean absolute error for each ML method. (A) Descriptor-based approach. (B) Sequence-based approach. (C) Graph-based approach. (D) Quantum-based approach. Predicted σ values include the test set across all folds of cross-validation. The σ values computed at 10.6 eV (117 nm) are based on the quantum chemistry at the level of M06-2X/aug-cc-pVTZ. The panels show σ values only in cases for which the measured IP values are 10.6 eV (117 nm) or lower. Vertical IP values were used in the ePolyScat computation of σ.

outperformed it. The QDF-NN also had larger kurtosis than the other models, indicating heavy tails in the distribution and more outliers. For predictive modeling, a smaller kurtosis is preferable to minimize the bound of predictive error.

3.3. Quantum Chemical Calculations Compared to Photoionization Cross Section Measurements. Table 1 lists calculated and measured σ values for a subset of molecules present in the NIST dataset. There is a high coefficient of determination (R^2 = 0.97). The absolute error in IP and σ for these molecules is 0.12 eV and 1.1 Mb. There is also high correlation with the PID response factors (Pearson's r = −0.84). These empirically determined values indicate the response of a PID sensor to a specific molecule. PID response factors and photoionization cross section are typically closely related.[69] Across all molecules in Table 1, the average uncertainty in the photoionization cross section was 1.0 Mb. For comparison, the σ value of stable species can typically be measured within a margin of error of <20%.[70−74] Error can arise from several factors, including (1) measurement uncertainty in IP (e.g., 0.1−0.2 eV for photoelectron spectroscopy, 0.05−0.1 eV for photoionization mass spectrometry), (2) measurement uncertainty associated in σ, and (3) computational limitations. With respect to the last, computations necessarily were not performed using an infinite basis set or with full configuration interactions.

3.4. Machine Learning Methods Compared to Calculations of the Photoionization Cross Sections. In a series of parity plots (Figure 5), the σ values computed by ePolyScat for 724 molecules are compared with those values predicted by the ML models. The subset included all molecules for which both IP and σ simulations completed successfully and ePolyScat simulations exhibited convergence of ExpOrb values. Results are summarized in Table 2. The AFP model yielded the best overall results ($\varepsilon_\sigma = 2.8 \pm 0.2$ Mb) for a correlation coefficient of R^2 = 0.66. This model again performed better than the quantum-

https://doi.org/10.1021/acsearthspacechem.3c00009
ACS Earth Space Chem. 2023, 7, 863−875

Table 2. Performance Comparison among the Different ML-Based Methods of This Study[a]

Method	Model	ε_{IP} (eV)	ε_{σ} (Mb)
Descriptor-based	Neural network	0.44 ± 0.02	3.5 ± 0.3
	Gradient boosting	0.33± 0.02	3.4 ± 0.2
	Random forest	0.33± 0.02	3.2 ± 0.3
Sequence-based	BPE-ChemBERTa	0.29 ± 0.01	3.1 ± 0.2
	SMILES-ChemBERTa	0.26 ± 0.01	3.0 ± 0.3
Graph-based	GCN base	0.39 ± 0.02	3.7 ± 0.3
	GCN with pretraining on GM9 HOMO	0.32 ± 0.02	3.9 ± 0.2
	AFP	**0.23 ± 0.01**	**2.8 ± 0.2**
Quantum-based	QDF-NN base	0.33 ± 0.01	3.4 ± 0.1
	QDF-NN with pretraining on GM9 HOMO	0.30 ± 0.02	3.1 ± 0.1

[a]In all cases, the IP and σ values were predicted as the ML output for molecular SMILES strings as the ML input. The ε value listed for each model represents the mean absolute error across a fivefold cross-validation of that model (see main text). The error for the IP values is referenced to measurements. The error for σ values is referenced to quantum chemistry computations at the level of M06-2X/aug-cc-pVTZ. The results of the best model are shown in bold.

based QDF-NN model, including pretraining (ε_{σ} = 3.1 ± 0.1 Mb). For comparison to the error, the computed σ values ranged from 0 to 40 Mb. The ML methods all had systematic underprediction for large σ values. The error and dispersion in the performance of the ML methods can in part be attributed to the uncertainties in the computed values constituting the training dataset itself. These values had an average error between the velocity and length gauge values of 5.0 Mb. This error could be reduced by using a larger value for L_{max} but at the cost of computational resources.

3.5. Implications. The ML-based approaches used herein have a significantly lower computational complexity than the quantum chemical calculations. The computational complexity of DFT scales with the number of electrons (n) as $O(n^3)$. When incorporating Hartree−Fock exchange, as used in the hybrid functionals B3LYP and M06-2X, this complexity increases to $O(n^4)$.[75,76] In contrast, the ML-based approaches are not significantly affected by the number of electrons in the system, particularly during the inference stage, making them significantly faster for making predictions. The downside of this approach is that the training process can be computationally intensive, particularly for QDF-NN. The training procedure is, however, significantly faster than IP predictions with M06-2X/aug-cc-pVTZ as well as for the σ simulations when a high L_{max} is used in the DFT calculations. The explicitness to which electronic correlation is considered did not significantly impact the computational complexity of the ML-based methods. This behavior can explain the superior performance of the AFP model compared to the QDF-NN model, despite the different approaches to modeling electronic correlation.

Although the ML-based approaches perform better in terms of scaling than the quantum chemical calculations, they offer limited chemical interpretability. By comparison, quantum chemical calculations determine the wavefunctions of a system, from which many properties can be derived. The corresponding molecular orbitals and other molecular properties can be examined to interpret the reliability of results. Neural-based approaches do not solve the Schrödinger equation and thus cannot be interpreted in the same way. Even so, other

Table 3. Label Explanation for Top Molecular Descriptors Selected by mRMR for Predicting IP and σ Values. Correlation Coefficients (R) Are Listed

Label	Correlation (R)	Description
IP Model		
BCUTp-1h	0.51	largest eigenvalue of Burden matrix weighted by polarizability
BCUTp-1l	0.50	lowest eigenvalue of Burden matrix weighted by polarizability
PubchemFP564	0.50	PubChem binary substructure fingerprint
AATS0s	0.49	averaged Moreau−Broto autocorrelation of lag zero weighted by intrinsic state
maxsssN	0.41	maximum atom-type *E*-state: >N−
minsssN	0.41	minimum atom-type *E*-state: >N−
TIC1	0.58	1-ordered neighborhood total information content
TDB2e	0.29	distance-based autocorrelation—lag 2/weighted by Sanderson electronegativities
GATS2c	0.26	Geary coefficient of lag 2 weighted by Gasteiger charge
SpMax5_Bhi	0.53	largest absolute eigenvalue of Burden modified matrix—n 5/weighted by relative first ionization
hmin	0.63	minimum H *E*-state
piPC3	0.54	3-ordered π-path count (log scale)
ntN	0.28	number of atom-type *E*-state: #N
piPC2	0.60	2-ordered π-path count (log scale)
MWC4	0.54	molecular walk count of order 4 ($\ln(1+x)$)
TWC	0.55	total walk count (up to order 10)
SRW4	0.58	self-returning walk count of order 4 ($\ln(1+x)$)
MWC3	0.57	molecular walk count of order 3 ($\ln(1+x)$)
SRW2	0.60	self-returning walk count of order 2 ($\ln(1+x)$)
MWC2	0.59	molecular walk count of order 2 ($\ln(1+x)$)
σ Model		
ATSC0e	0.32	centered Moreau−Broto autocorrelation—lag 0/weighted by Sanderson electronegativities
maxwHBa	0.37	maximum *E*-states for weak hydrogen bond acceptors
SM1_Dzs	0.38	spectral moment of order 1 from Barysz matrix weighted by *I*-state
MATS3c	0.07	moran autocorrelation—lag 3/weighted by charges
minwHBa	0.29	minimum *E*-states for weak hydrogen bond acceptors
StsC	0.37	sum of atom-type *E*-state: #C−
maxtsC	0.37	maximum atom-type *E*-state: #C−
mintsC	0.35	minimum atom-type *E*-state: =C=
SM1_Dzp	0.30	spectral moment of order 1 from Barysz matrix/weighted by polarizabilities
ntsC	0.41	count of atom-type *E*-state: #C−
PubchemFP460	0.40	PubChem binary substructure fingerprint
nBondsT	0.39	number of triple bonds
maxddC	0.07	maximum atom-type *E*-state: =C=
PubchemFP427	0.40	PubChem binary substructure fingerprint
PubchemFP495	0.14	PubChem binary substructure fingerprint
PubchemFP417	0.40	PubChem binary substructure fingerprint
PubchemFP797	0.13	PubChem binary substructure fingerprint
PubchemFP309	0.13	PubChem binary substructure fingerprint
nT10Ring	0.03	PubChem binary substructure fingerprint
MDEN-11	−0.04	molecular distance edge between all primary nitrogens

https://doi.org/10.1021/acsearthspacechem.3c00009
ACS Earth Space Chem. 2023, 7, 863−875

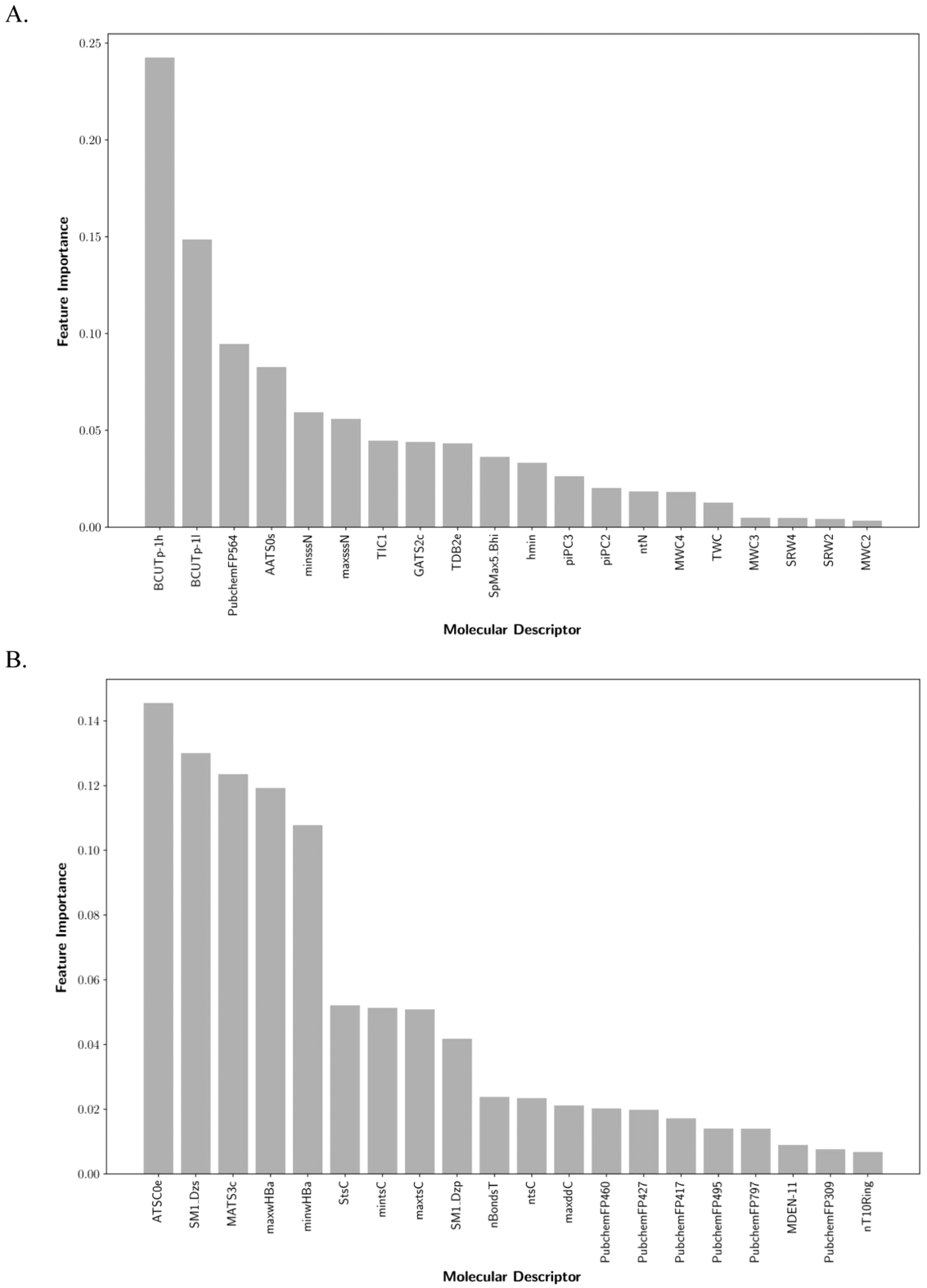


Figure 6. Bar charts of feature importance for the 30 most relevant molecular descriptors selected by the ML random-forest method of maximum relevance and minimum redundancy (mRMR). Abbreviations for the descriptors are listed in Table 3. Feature importance refers to the weighted decrease in the impurity criterion for a node (see the Supporting Information for more details). Higher scores indicate greater relevance to model performance. Values are computed for IP, and σ are plotted in panels (A) and (B), respectively.

approaches are available for interpreting neural-based methods. DeepChem allows the approach of Local Interpretable Model-Agnostic Explanations (LIME)[77] to determine which parts of a molecule contribute most to the prediction. Descriptor-based models can make use of LIME to check performance changes for successive removal of descriptors. In principle, the learned electron density can be extracted from the QDF-NN model, but implementation is still lacking. Non-neural descriptor-based

https://doi.org/10.1021/acsearthspacechem.3c00009
ACS Earth Space Chem. 2023, 7, 863−875

models can use feature importance or regression coefficients (Figure 6). A description of each of these molecular descriptors is provided in Table 3. The procedure for calculating feature importance is available in the Supporting Information. Examples of the chemical structures and SMILES strings for the least accurate IP and σ values from the AFP model also appear in the Supporting Information.

In conclusion, obtaining accurate estimates for the ionization potentials and photoionization cross sections of VOCs is important for predicting the species-specific response of the photoionization detectors and other nonselective optical chemical sensors. Computing these values for thousands of different molecules involved in the oxidation mechanisms of atmospheric VOCs is infeasible because of computational cost. Developing ML-based alternatives to accurately predict these quantities is an alternative approach. In this study, multiple ML methods were used to predict IP and σ values at 10.6 eV (117 nm). Improvements in model performance correlated with progressively more explicit consideration of electronic structure. For IP values, pretraining boosted performance substantially for the GNN and QDF-NN models, reducing ε by 24% for both cases. Similar but smaller improvements in accuracy were achieved for σ values. Overall, the AFP models provided the most accurate estimates, achieving $\varepsilon = 0.23 \pm 0.01$ eV for IP and $\varepsilon = 2.8 \pm 0.2$ Mb for σ. Average relative errors for these two methods were 2.5% for IP and 14% for σ across all molecules. The results show that accurate parameter estimates of IP and σ can be obtained for VOCs by ML methods to comparable accuracy as that of quantum chemical methods. The ML methods, however, have a lower computational cost by several orders of magnitude. The ML methods can rapidly parse species-specific responses of nonselective optical chemical sensors, such as photoionization detectors. Future studies are planned to incorporate this predictive model into experimental studies utilizing PID sensors.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsearthspacechem.3c00009.

Details on calculating feature importance for the descriptor-based model, as well as details and hyperparameters for each of the four ML-based model approaches, are provided in the Supplementary Information (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Scot T. Martin − *School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, United States; Department of Earth and Planetary Sciences, Harvard University, Cambridge, Massachusetts 02138, United States;* orcid.org/0000-0002-8996-7554; Email: scot_martin@harvard.edu

Author

Matthew P. Stewart − *School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, United States;* orcid.org/0000-0002-4851-1315

Complete contact information is available at: https://pubs.acs.org/10.1021/acsearthspacechem.3c00009

Notes

The authors declare no competing financial interest.

The dataset and code used in this work is stored in Harvard's Dataverse (https://doi.org/10.7910/DVN/5VBRLS). They can be used to reproduce the results of the machine learning methods.[38] Stored information includes the output data from Gaussian and ePolyScat as well as the training dataset of IP and σ values. The code for each of the models outlined in Table [1] is included.

■ ACKNOWLEDGMENTS

Financial support for this project provided by the USA National Science Foundation is gratefully acknowledged (Environmental Chemical Sciences of the Division of Chemistry, award 2003368, and Atmospheric Chemistry of the Division of Atmospheric and Geospace Sciences, award 1829025). The computations in this paper were supported by the FASRC Cannon cluster of the Harvard FAS Science Research Computing Group.

■ REFERENCES

(1) Kesselmeier, J.; Kuhn, U.; Rottenberger, S.; Biesenthal, T.; Wolf, A.; Schebeske, G.; Andreae, M. O.; Ciccioli, P.; Brancaleoni, E.; Frattoni, M.; Oliva, S. T.; Botelho, M. L.; Silva, C. M. A.; Tavares, T. M. Concentrations and species composition of atmospheric volatile organic compounds (VOCs) as observed during the wet and dry season in Rondônia (Amazonia). *J. Geophys. Res.: Atmos.* **2002**, *107*, LBA 20-1−LBA 20-13.

(2) Chandrasekaran, A.; Kamal, D.; Batra, R.; Kim, C.; Chen, L.; Ramprasad, R. Solving the electronic structure problem with machine learning. *npj Comput. Mater.* **2019**, *5*, No. 22.

(3) Chinta, S.; Rengaswamy, R. Machine Learning Derived Quantitative Structure Property Relationship (QSPR) to Predict Drug Solubility in Binary Solvent Systems. *Ind. Eng. Chem. Res.* **2019**, *58*, 3082−3092.

(4) Rácz, A.; Bajusz, D.; Miranda-Quintana, R. A.; Héberger, K. Machine learning models for classification tasks related to drug safety. *Mol. Diversity* **2021**, *25*, 1409−1424.

(5) Li, R.; Herreros, J. M.; Tsolakis, A.; Yang, W. Machine learning-quantitative structure property relationship (ML-QSPR) method for fuel physicochemical properties prediction of multiple fuel types. *Fuel* **2021**, *304*, No. 121437.

(6) Chaudhari, P.; Ade, N.; Pérez, L. M.; Kolis, S.; Mashuga, C. V. Quantitative Structure-Property Relationship (QSPR) models for Minimum Ignition Energy (MIE) prediction of combustible dusts using machine learning. *Powder Technol.* **2020**, *372*, 227−234.

(7) Schustik, S. A.; Cravero, F.; Ponzoni, I.; Díaz, M. F. Polymer informatics: Expert-in-the-loop in QSPR modeling of refractive index. *Comput. Mater. Sci.* **2021**, *194*, No. 110460.

(8) Yu, X. Support vector machine-based QSPR for the prediction of glass transition temperatures of polymers. *Fibers Polym.* **2010**, *11*, 757−766.

(9) Tsubaki, M.; Mizoguchi, T.On the Equivalence of Molecular Graph Convolution and Molecular Wave Function with Poor Basis Set. *Advances in Neural Information Processing Systems*; Curran Associates Inc., 2020; pp 1982−1993.

(10) Goh, G. B.; Siegel, C.; Vishnu, A.; Hodas, N. O.; Baker, N. Chemception: a deep neural network with minimal chemistry knowledge matches the performance of expert-developed QSAR/QSPR models. 2017, arXiv:1706.06689. arXiv.org e-Print archive. https://arxiv.org/abs/1706.06689.

(11) Wieder, O.; Kohlbacher, S.; Kuenemann, M.; Garon, A.; Ducrot, P.; Seidel, T.; Langer, T. A compact review of molecular property prediction with graph neural networks. *Drug Discovery Today: Technol.* **2020**, *37*, 1−12.

(12) Berkowitz, J. *Photoabsorption, Photoionization, and Photoelectron Spectroscopy*; Academic Press, 2012.

(13) Linstrom, P. *NIST Chemistry WebBook*, NIST Standard Reference Database 69 [Data Set]; National Institute of Standards and Technology, 1997.
(14) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(15) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, No. 33.
(16) Tirado-Rives, J.; Jorgensen, W. L. Performance of B3LYP Density Functional Methods for a Large Set of Organic Molecules. *J. Chem. Theory Comput.* **2008**, *4*, 297−306.
(17) Chai, J.-D.; Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom−atom dispersion corrections. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615−6620.
(18) Zhao, Y.; Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* **2008**, *120*, 215−241.
(19) Ogliaro, F.; Bearpark, M.; Heyd, J.; Brothers, E.; Kudin, K.; Staroverov, V.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. *Gaussian 09*, revision A.02; Gaussian, Inc.: Wallingford, CT, 2009.
(20) Gianturco, F. A.; Lucchese, R. R.; Sanna, N. Calculation of low-energy elastic cross sections for electron-CF4 scattering. *J. Chem. Phys.* **1994**, *100*, 6464−6471.
(21) Natalense, A. P. P.; Lucchese, R. R. Cross section and asymmetry parameter calculation for sulfur 1s photoionization of SF6. *J. Chem. Phys.* **1999**, *111*, 5344−5348.
(22) Stener, M.; Decleva, P. Time-dependent density functional calculations of molecular photoionization cross sections: N 2 and PH 3. *J. Chem. Phys.* **2000**, *112*, 10871−10879.
(23) Le Rouzo, H.; Raseev, G. Finite-volume variational method: First application to direct molecular photoionization. *Phys. Rev. A* **1984**, *29*, 1214.
(24) Stephens, J. A.; McKoy, V. Photoionization of the valence orbitals of OH. *J. Chem. Phys.* **1988**, *88*, 1737−1742.
(25) Gozem, S.; Gunina, A. O.; Ichino, T.; Osborn, D. L.; Stanton, J. F.; Krylov, A. I. Photoelectron wave function in photoionization: Plane wave or Coulomb wave? *J. Phys. Chem. Lett.* **2015**, *6*, 4532−4540.
(26) Moshammer, K.; Jasper, A. W.; Popolan-Vaida, D. M.; Wang, Z.; Bhavani Shankar, V. S.; Ruwe, L.; Taatjes, C. A.; Dagaut, P.; Hansen, N. Quantification of the Keto-Hydroperoxide (HOOCH2OCHO) and Other Elusive Intermediates during Low-Temperature Oxidation of Dimethyl Ether. *J. Phys. Chem. A* **2016**, *120*, 7890−7901.
(27) Huang, C.; Yang, B.; Zhang, F. Calculation of the absolute photoionization cross-sections for C1−C4 Criegee intermediates and vinyl hydroperoxides. *J. Chem. Phys.* **2019**, *150*, No. 164305.
(28) Lucchese, R. R.; Raseev, G.; McKoy, V. Studies of differential and total photoionization cross sections of molecular nitrogen. *Phys. Rev. A* **1982**, *25*, 2572.
(29) Gozem, S.; Krylov, A. I. The ezSpectra suite: An easy-to-use toolkit for spectroscopy modeling. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2022**, *12*, No. e1546.
(30) Shaw, D.; Holland, D.; MacDonald, M.; Hopkirk, A.; Hayes, M.; McSweeney, S. A study of the absolute photoabsorption cross section and the photoionization quantum efficiency of nitrous oxide from the ionization threshold to 480 Å. *Chem. Phys.* **1992**, *163*, 387−404.
(31) Cool, T. A.; Wang, J.; Nakajima, K.; Taatjes, C. A.; Mcllroy, A. Photoionization cross sections for reaction intermediates in hydrocarbon combustion. *Int. J. Mass Spectrom.* **2005**, *247*, 18−27.
(32) Landrum, G. *RDKit: A Software Suite for Cheminformatics, Computational Chemistry, and Predictive Modeling*, 2013.
(33) Ramsundar, B.; Eastman, P.; Walters, P.; Pande, V. *Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More*; O'Reilly Media, 2019.
(34) Tsubaki, M.; Mizoguchi, T. Quantum Deep Field: Data-Driven Wave Function, Electron Density Generation, and Atomization Energy Prediction and Extrapolation with Machine Learning. *Phys. Rev. Lett.* **2020**, *125*, No. 206401.
(35) Ruddigkeit, L.; Van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864−2875.
(36) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, *1*, No. 140022.
(37) Koopmans, T. Über die Zuordnung von Wellenfunktionen und Eigenwerten zu den einzelnen Elektronen eines Atoms. *Physica* **1934**, *1*, 104−113.
(38) Stewart, M.*Replication Data for: Machine Learning for Ionization Potentials and Photoionization Cross Sections of Volatile Organic Compounds*; Harvard Dataverse, 2022.
(39) Katritzky, A. R.; Maran, U.; Karelson, M.; Lobanov, V. S. Prediction of Melting Points for the Substituted Benzenes: A QSPR Approach. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 913−919.
(40) Goll, E. S.; Jurs, P. C. Prediction of the Normal Boiling Points of Organic Compounds from Molecular Structures with a Computational Neural Network Model. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 974−983.
(41) Mirshahvalad, H.; Ghasemiasl, R.; Raoufi, N.; Malekzadeh Dirin, M. A neural network QSPR model for accurate prediction of flash point of pure hydrocarbons. *Mol. Inf.* **2019**, *38*, No. 1800094.
(42) Yuan, S.; Jiao, Z.; Quddus, N.; Kwon, J. S.-I.; Mashuga, C. V. Developing quantitative structure−property relationship models to predict the upper flammability limit using machine learning. *Ind. Eng. Chem. Res.* **2019**, *58*, 3531−3537.
(43) Goudarzi, N.; Goodarzi, M. Prediction of the acidic dissociation constant (pKa) of some organic compounds using linear and nonlinear QSPR methods. *Mol. Phys.* **2009**, *107*, 1495−1503.
(44) Fayet, G.; Rotureau, P.; Joubert, L.; Adamo, C. QSPR modeling of thermal stability of nitroaromatic compounds: DFT vs. AM1 calculated descriptors. *J. Mol. Model.* **2010**, *16*, 805−812.
(45) Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2011**, *32*, 1466−1474.
(46) Ding, C.; Peng, H. Minimum Redundancy Feature Selection from Microarray Gene Expression Data. *J. Bioinf. Comput. Biol.* **2003**, *3*, 185−205.
(47) Wu, J.; Chen, X.-Y.; Zhang, H.; Xiong, L.-D.; Lei, H.; Deng, S.-H. Hyperparameter optimization for machine learning models based on Bayesian optimization. *J. Electron. Sci. Technol.* **2019**, *17*, 26−40.
(48) Irwin, R.; Dimitriadis, S.; He, J.; Bjerrum, E. J. Chemformer: a pre-trained transformer for computational chemistry. *Mach. Learn.: Sci. Technol.* **2022**, *3*, No. 015022.
(49) Chithrananda, S.; Grand, G.; Ramsundar, B. ChemBERTa: Large-Scale Self-supervised Pretraining for Molecular Property Prediction. 2020, arXiv:2010.09885. arXiv.org e-Print archive. https://arxiv.org/abs/2010.09885.
(50) Wang, S.; Guo, Y.; Wang, Y.; Sun, H.; Huang, J. In *SMILES-BERT: Large Scale Unsupervised Pre-training for Molecular Property Prediction*, Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, 2019; pp 429−436.
(51) Honda, S.; Shi, S.; Ueda, H. R. Smiles Transformer: Pre-trained Molecular Fingerprint for Low Data Drug Discovery. 2019, arXiv:1911.04738. arXiv.org e-Print archive. https://arxiv.org/abs/1911.04738.
(52) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Advances in Neural Information Processing Systems*; Curran Associates, Inc., 2017; Vol. *30*.
(53) Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M. Huggingface's Transformers: State-of-the-Art Natural Language Processing. 2019, arXiv:1910.03771. arXiv.org e-Print archive. https://arxiv.org/abs/1910.03771.
(54) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; et al. PubChem 2019 update:

improved access to chemical data. *Nucleic Acids Res.* **2019**, *47*, D1102−D1109.
(55) Bronstein, M. M.; Bruna, J.; LeCun, Y.; Szlam, A.; Vandergheynst, P. Geometric deep learning: going beyond euclidean data. *IEEE Signal Process. Mag.* **2017**, *34*, 18−42.
(56) Jiang, D.; Wu, Z.; Hsieh, C.-Y.; Chen, G.; Liao, B.; Wang, Z.; Shen, C.; Cao, D.; Wu, J.; Hou, T. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *J. Cheminf.* **2021**, *13*, No. 12.
(57) Zhang, S.; Tong, H.; Xu, J.; Maciejewski, R. Graph convolutional networks: a comprehensive review. *Comput. Soc. Networks* **2019**, *6*, No. 11.
(58) Xiong, Z.; Wang, D.; Liu, X.; Zhong, F.; Wan, X.; Li, X.; Li, Z.; Luo, X.; Chen, K.; Jiang, H.; Zheng, M. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *J. Med. Chem.* **2020**, *63*, 8749−8760.
(59) Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P.Convolutional Networks on Graphs for Learning Molecular Fingerprints. *Advances in Neural Information Processing Systems*; Curran Associates, Inc., 2015; p 28.
(60) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L. Pytorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*; Curran Associates Inc., 2019; Vol. 32.
(61) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. 2014, arXiv:1412.6980. arXiv.org e-Print archive. https://arxiv.org/abs/1412.6980.
(62) Raissi, M.; Perdikaris, P.; Karniadakis, G. E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **2019**, *378*, 686−707.
(63) Tsubaki, M.; Mizoguchi, T. On the Equivalence of Molecular Graph Convolution and Molecular Wave Function with Poor Basis Set. In *Advances in Neural Information Processing Systems*; Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; Lin, H., Eds.; Curran Associates, Inc., 2020; pp 1982−1993.
(64) Cooper, J. K.; Grant, C. D.; Zhang, J. Z. Ab initio calculation of ionization potential and electron affinity of six common explosive compounds. *Rep. Theor. Chem.* **2012**, *1*, 11−19.
(65) DiLabio, G.; Pratt, D.; Wright, J. S. Theoretical calculation of gas-phase ionization potentials for mono-and polysubstituted benzenes. *Chem. Phys. Lett.* **1999**, *311*, 215−220.
(66) Orio, M.; Pantazis, D. A.; Neese, F. Density functional theory. *Photosynth. Res.* **2009**, *102*, 443−453.
(67) Alon, U.; Yahav, E. On the Bottleneck of Graph Neural Networks and Its Practical Implications. 2020, arXiv:2006.05205. arXiv.org e-Print archive. https://arxiv.org/abs/2006.05205.
(68) Fung, V.; Zhang, J.; Juarez, E.; Sumpter, B. G. Benchmarking graph neural networks for materials chemistry. *npj Comput. Mater.* **2021**, *7*, No. 84.
(69) Freedman, A. N. The photoionization detector: Theory, performance and application as a low-level monitor of oil vapour. *J. Chromatogr. A* **1980**, *190*, 263−273.
(70) Yang, B.; Wang, J.; Cool, T. A.; Hansen, N.; Skeen, S.; Osborn, D. L. Absolute photoionization cross-sections of some combustion intermediates. *Int. J. Mass Spectrom.* **2012**, *309*, 118−128.
(71) Ho, G. H.; Lin, M. S.; Wang, Y. L.; Chang, T. W. Photoabsorption and photoionization of propyne. *J. Chem. Phys.* **1998**, *109*, 5868−5879.
(72) Samson, J. A. R.; Haddad, G.; Masuoka, T.; Pareek, P.; Kilcoyne, D. Ionization yields, total absorption, and dissociative photoionization cross sections of CH4 from 110 to 950 Å. *J. Chem. Phys.* **1989**, *90*, 6925−6932.
(73) Person, J. C.; Nicole, P. P. Isotope effects in the photoionization yields and the absorption cross sections for acetylene, propyne, and propene. *J. Chem. Phys.* **1970**, *53*, 1767−1774.
(74) Haddad, G. N.; Samson, J. A. Total absorption and photoionization cross sections of water vapor between 100 and 1000 Å. *J. Chem. Phys.* **1986**, *84*, 6623−6626.
(75) Martin, R. M. *Electronic Structure: Basic Theory and Practical Methods*; Cambridge University Press, 2020.
(76) Ciarlet, P. G.; Cucker, F.; Lions, J.-L.; Du, Q. *Handbook of Numerical Analysis*; Gulf Professional Publishing, 1990; Vol. *11*.
(77) Ribeiro, M. T.; Singh, S.; Guestrin, C. In *"Why Should I Trust You?" Explaining the Predictions of Any Classifier*, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016; pp 1135−1144.
(78) Adam, T. W.; Zimmermann, R. Determination of single photon ionization cross sections for quantitative analysis of complex organic mixtures. *Anal. Bioanal. Chem.* **2007**, *389*, 1941−1951.
(79) Eschner, M. S.; Zimmermann, R. Determination of photoionization cross-sections of different organic molecules using gas chromatography coupled to single-photon ionization (SPI) time-of-flight mass spectrometry (TOF-MS) with an electron-beam-pumped rare gas excimer light source (EBEL): influence of molecular structure and analytical implications. *Appl. Spectrosc.* **2011**, *65*, 806−816.
(80) Czekner, J.; Taatjes, C. A.; Osborn, D. L.; Meloni, G. Absolute photoionization cross-sections of selected furanic and lactonic potential biofuels. *Int. J. Mass Spectrom.* **2013**, *348*, 39−46.

https://doi.org/10.1021/acsearthspacechem.3c00009
ACS Earth Space Chem. 2023, 7, 863−875