RSC Advances

PAPER

View Article Online
View Journal | View Issue

Cite this: *RSC Adv.*, 2021, **11**, 14552

On the relationship between spectroscopic constants of diatomic molecules: a machine learning approach

Xiangyue Liu, Gerard Meijer and Jesús Pérez-Ríos*

Received 15th March 2021
Accepted 1st April 2021

DOI: 10.1039/d1ra02061g

rsc.li/rsc-advances

Through a machine learning approach, we show that the equilibrium distance, harmonic vibrational frequency and binding energy of diatomic molecules are related, independently of the nature of the bond of a molecule; they depend solely on the group and period of the constituent atoms. As a result, we show that by employing the group and period of the atoms that form a molecule, the spectroscopic constants are predicted with an accuracy of <5%, whereas for the A-excited electronic state it is needed to include other atomic properties leading to an accuracy of <11%.

1 Introduction

Early in the history of molecular spectroscopy, when it became a discipline within chemical physics in the 1920's,[1] some intriguing empirical relationships between different spectroscopic properties were observed.[2–4] In particular, it was found that the equilibrium distance, R_e, and the harmonic vibrational frequency, ω_e, were correlated in diatomic molecules. As the field evolved, the relationship between R_e and ω_e became more evident, and more empirical relations between spectroscopic constants were identified.[5–12] However, these empirical relationships were typically only valid for specific atomic numbers or groups of the constituent atoms. These results motivated the development of realistic diatomic molecular potentials[4,13–17] and triggered the physical chemistry community to think about the "periodicity" of diatomic molecules.[18]

The development of quantum chemistry helped to shed some light on the physics behind empirical relationships between spectroscopic constants. In particular, thanks to the application of the Hellmann–Feynman theorem, it was possible to connect ω_e directly with the electronic density at R_e.[19–22] As a result, a first principles-based explanation (containing a few free parameters), of the observed empirical relations between spectroscopic constants appeared.[23–30] Nevertheless, the obtained relations based on the electronic density were only valid for subsets of molecules. To date, it has not been possible to find general relations for spectroscopic constants of diatomic molecules in terms of the properties of their constituent atoms.

The accuracy of quantum chemistry methods relies on (finite) basis sets optimized for each element under certain bounds. At the same time, an accurate description of the system's electronic structure is required, which is achieved through a hierarchy of different treatments of the electron correlation.[31,32] On the other hand, the widely-used (Kohn–Sham) density functional theory (DFT) methods require accurate electron exchange–correlation density functionals. The non-empirical density functionals are derived under certain constraints, some with several free parameters,[33–36] while the semi-empirical density functionals employ more flexible functional forms with (sometimes even several tens of) coefficients fitted to various experimental or theoretical reference properties.[33,37] Machine learning (ML) methods, on the other hand, discover the underlying relationships from data (the so-called "training set") and build up models on top of them. These models can be quantitatively predictive for other systems that follow similar underlying physics. More importantly, they provide possibilities for discovering relationships between the different properties of the system under consideration.[38,39]

This work shows that the relationship between spectroscopic constants of heteronuclear diatomic molecules is general for most kinds of molecules at hand. Our findings rely upon applying state-of-the-art ML models to an orthodox dataset of experimental spectroscopic constants for diatomic molecules. In particular, we apply the Gaussian process (GP) regression model[40] to predict R_e, ω_e, and the binding energy, D_0, as a function of the group and period of the constituent atoms. Similarly, our model can predict R_e and ω_e for the A-excited electronic state of a given molecule. Our findings generalize the idea that some of the system's chemical properties depend on the atoms' group and period. Indeed, the periodicity of elements has long been used to predict chemical compounds' properties intuitively at a qualitative level. However, the correlations between the chemical properties and the constituent atoms' periodicity are not always straightforward, and such predictions can hardly be quantitative in most cases. On the contrary, our main result is quantitatively meaningful: it is possible to predict those spectroscopic constants with an

Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, 14195 Berlin, Germany. E-mail: jperezri@fhi-berlin.mpg.de

accuracy of <5% for ground electronic states and <11% for the A-excited electronic state. More interestingly, by analyzing our models' outliers, we show that molecules showing a non-chemical bond nature like bi-alkali molecules and molecules containing first-row elements, such as HF, are more difficult to predict. However, the spectroscopic constants of molecules containing transition metals challenging for quantum chemistry methods can be adequately described.

2 The quest of relationships between spectroscopic constants of diatomic molecules

As soon as molecular spectroscopy became an essential tool to analyze molecules' unique fingerprints and more spectra of molecules were taken, approximate relationships were found between spectroscopic constants. As a result, it was postulated that the molecules' spectroscopic constants might be correlated based on empirical grounds. In particular, it was observed that the equilibrium distance and the harmonic vibrational frequency are related as $R_e^{\,2}\omega_e^{\,2}m$ = const in hydrogen halides,[2,41–43] where m is the reduced mass of the molecule. This relationship was generalized as $R_e^{\,i}\omega_e^{\,2}m$ = const, the precursor of the well-known Badger's rule,[6] where i is a natural number. On the other hand, after studying the spectra of 16 molecules, including homonuclear molecules and molecular ions, Mecke and Birge found that the expression $R_e^{\,2}\omega_e$ = const described the observed spectra better.[3,44] In the same line, but using a given functional form for the interatomic interaction of a molecule, Morse proposed a relationship given as $R_e^{\,3}\omega_e$ = const.[4] Finally, more involved relationships between the equilibrium distance and the vibrational harmonic frequency were proposed[17] as $mR_e^{\,6}\omega_e^{\,2}n^a$, where n stands for the number of valence electrons, and a is a rational number. The results for a variety of the proposed empirical rules are shown in Fig. 1, where it is noticed that for a larger dataset, as the present one, none of the empirical relationships hold.

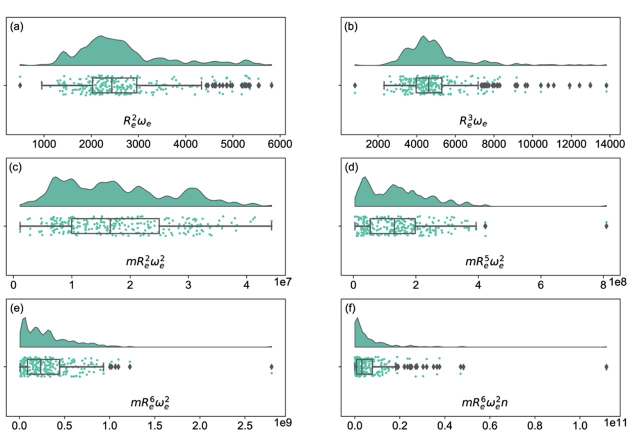

Fig. 1 Distribution and box plots of $R_e^{\,a}\omega_e^{\,b}$ with different powers combined with the reduced mass m and number of valence electrons n.

At the same time, more spectroscopic information of molecules became available, and more advanced and accurate quantum chemistry tools were developed. Therefore, it was possible to search for a first principle explanation of the empirically observed relationships between spectroscopic constants. In that endeavor, Parr and coworkers took the lead by looking at the electron density within a molecule as the source of the relationship between spectroscopic constants. The model assumes that the electron density mutually created by the one atom in the other atom is equal at the equilibrium distance, *i.e.*, at the sum of two atomic radii. In particular, the electron density of atom 1 at the position of atom 2, within a molecule, is given by[27]

$$\rho_1(2) = CZ_1\exp(-\xi R_1), \tag{1}$$

where C is a fitting parameter and ξ represents the decay constant of the electron density. Within this model, one finds a relationship between the atomic numbers of the two atoms, Z_1 and Z_2, and the equilibrium internuclear distance R_e of a diatomic molecule as[27,29,30,45]

$$Z_1Z_2 = A\exp(\xi R_e), \tag{2}$$

where A is a free parameter. According to this relationship, R_e depends linearly on $\log(Z_1Z_2)$ as

$$R_e = \xi^{-1}\log Z_1Z_2 - \xi^{-1}\log A. \tag{3}$$

However, the performance of this relationship has only been checked for molecules with atoms coming from the same group of the periodic table.[29]

Anderson, Parr and coworkers also suggested a relationship between ω_e and R_e[29] as

$$m\omega_e^{\,2} = 4\pi CZ_1Z_2e^{-2R_e}, \tag{4}$$

based on the Born–Oppenheimer approximation, the electron density of eqn (1) and the Hellmann–Feynman theorem. From eqn (4) it is possible to express the harmonic vibrational frequency in terms of the equilibrium distance and atomic properties as

$$\omega_e = \sqrt{\frac{C'Z_1Z_2e^{-2R_e}}{m}}. \tag{5}$$

In the same vein, following the relationship between the equilibrium distance and the harmonic vibrational frequency, it is possible to find a relationship between the atomic number Z_i, R_e, and the dissociation energy D_e, as[27,29,30,45]

$$\frac{D_e}{R_e^{\,l}} = 4\pi CZ_1Z_2\exp(-\xi' R_e), \tag{6}$$

which can be rewritten as

$$\log\frac{D_e}{R_e^{\,l}Z_1Z_2} = -\xi' R_e + \log(4\pi C). \tag{7}$$

For the derivation of eqn (6) it must be assumed that $D_e = Am\omega_e^{\,2}R_e^{\,l}$ without any further justification.[30] In eqn (7), $l = 3$ and

$\xi' = 0.97$. Eqn (7) has been tested in a dataset of 150 molecules leading to a good result, although no further characterization of the model performance was reported to objectively judge its quality. Finally, using the relation of the dissociation energy, D_e, and the binding energy, D_0,

$$D_e = D_0 + \frac{1}{2}\hbar\omega_e - \frac{1}{4}\hbar\omega_e x_e, \tag{8}$$

where $\omega_e x_e$ represents the first anharmonic correction to the harmonic vibrational frequency, it should be possible to find a linear regression model for $\log \frac{D_0}{{R_e}^l Z_1 Z_2}$.

3 The dataset

In this work, we focus on heteronuclear molecules due to their relevance on laser cooling of molecules with applications in ultracold chemistry.[47–49] The employed dataset contains the main spectroscopic constants: R_e, ω_e, and D_0 for the ground electronic state of heteronuclear diatomic molecules. In particular, it contains the experimental values of R_e, ω_e for 256 heteronuclear diatomic molecules taken from ref. 50–53, whereas the experimentally determined values of D_0 are only available for 197 of them.

As far as we know, this is the most extensive dataset for experimental ground state properties of heteronuclear diatomic molecules. Fig. 2 shows the equilibrium distance's distribution and its ratio to the sum of the atomic radii of the constituent atoms, $R_1 + R_2$, for molecules within the dataset. Most molecules show an equilibrium distance between 1.4 Å and 3.8 Å, with a most probable value of 1.7 Å. Looking at the values of $R_e/(R_1 + R_2)$, it is clear that the molecules within the dataset have different bonds: covalent, van der Waals, and ionic.

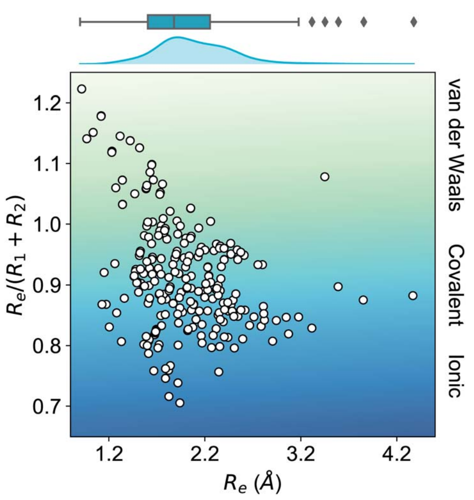

Fig. 2 Ratio of the equilibrium distance, R_e, to the sum of the atomic radii of the atoms forming a molecule, $R_1 + R_2$, *vs.* R_e. The background color indicates the nature of the molecular bond in each of the molecules. The density in the upper part of the figure shows the kernel density distribution of R_e. The box plot shows the minimum, the maximum, the sample median, and the first and third quarterlies of R_e. The empirical atomic radii of the atoms are taken from ref. 46.

We have classified the dataset based on the types of constituent atoms within a molecule, and the results are shown in Fig. 3. As a result, we notice that the dataset mainly consists of various metal and non-metal halides, hydrides, and metalloid compounds. It is worth noticing that more than 20% of the dataset contains transition metal compounds, including f-block elements. Therefore, the present dataset is general since it goes beyond the main-group diatomic molecules and deals with some of the more intriguing and complex atoms from a chemistry standpoint.

In addition to the dataset mentioned above of the ground state properties, we also study 131 molecules whose R_e, ω_e are available for the A-excited electronic state. The A-state dataset mainly consists of metal and non-metal compounds, including transition metal compounds and several f-block compounds.

4 Machine learning method

The quest for universal relationships between spectroscopic constants is related to the problem of how atomic and molecular properties describe a spectroscopic property of a molecule, $y = f(\mathbf{x})$. Here, $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, consists of different atomic properties of the constituent atoms or molecular properties, where n denotes the number of input features relevant for the problem at hand. Unlike traditional (non-)linear regression models, which assume a fixed form of function $f(\mathbf{x})$, GP embraces a Bayesian perspective and presumes a prior distribution over the space of functions

$$f(x_i) \sim GP\big(m(x_i),\ K(x_i,\ x_j)\big), \tag{9}$$

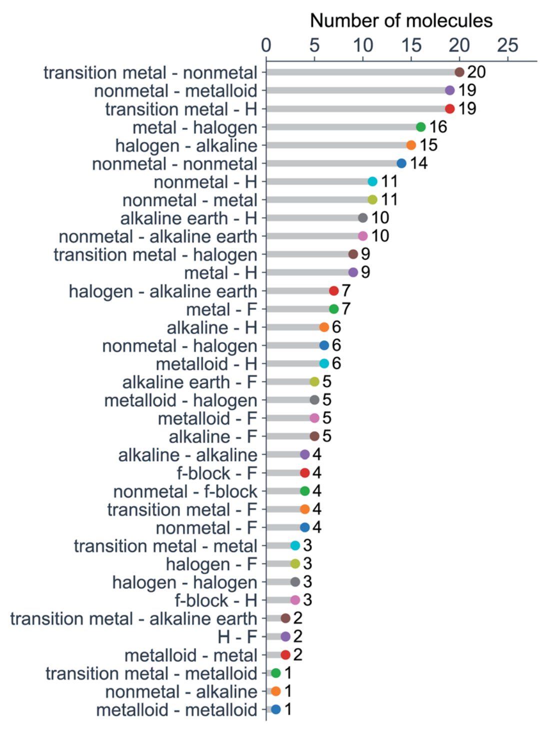

Fig. 3 Molecules in the dataset classified by the types of their constituent atoms.

with a joint multivariate-Gaussian distribution, centered at $m(\mathbf{x}_i)$ and characterized by the covariance function $K(\mathbf{x}_i, \mathbf{x}_j)$, which specifies the correlation (or "similarity") between data points.[40]

In this work, the spectroscopic properties y are modeled as

$$P(y_i|f(\mathbf{x}_i),\ \mathbf{x}_i) \sim N\left(y_i|\mathbf{h}(\mathbf{x}_i)\right)^{\mathrm{T}}\beta + f(\mathbf{x}_i),\ \sigma_y^{\,2}). \tag{10}$$

where the basis functions, $\mathbf{h}(\mathbf{x}_i)$, project $\{\mathbf{x}_i\}$ to a new (higher dimensional) feature space with coefficients β, and σ_y includes the noise in the observations.[40,54] The training set $D = \{(x_i,\ y_i)|i = 1,\ \ldots,\ N\}$ with N observations, constrains the available distribution of functions through Bayes theorem, and the mean of the posterior distribution is used for prediction. The functional form of $K(\mathbf{x}_i, \mathbf{x}_j)$ and $\mathbf{h}(\mathbf{x})$ can be selected according to the cross-validation performance of the models.

4.1 Model performance evaluation

In training and evaluating the regression models, as customary in ML, the ground state dataset is divided into training and test sets. The training set represents the set of molecules used for learning a given spectroscopic constant from the atomic properties of the constituents atoms. The test set is the set of molecules that have not been included in the learning procedure and hence are new to the regression algorithm. In learning the equilibrium internuclear distance, R_e, and the harmonic vibrational frequency ω_e, the training and test sets consist of 231 and 25 molecules, respectively. In learning $\log\frac{D_0}{R_e^{\,3}Z_1Z_2}$ the training/test splitting is 172/25. For learning R_e and ω_e for the A-excited electronic state, the training set consists of 106 molecules and the test set consists of 25 molecules.

The present dataset is relatively small from an ML perspective. When the dataset is split into training and test sets, the training set may not be representative. This may lead to a bias in the performance of the test set. To solve this problem, we have employed a Monte Carlo (MC) approach, in which the dataset is stratified into 25 strata based on the level of the true values of the labels (R_e, ω_e, and $\log\frac{D_0}{R_e^{\,3}Z_1Z_2}$ in the present work).

As shown in panel (a) of Fig. 4, we have two loops in the training and evaluation of the models. In the outer loop, we split the dataset into training set and test set. The training set is used to learn from the data and the test set is used for model evaluation. In the inner loop, we train the models with the training set, which is further split to perform a stratified 5-fold cross validation (CV) for the hyperparameter optimization. In particular, as shown in panel (b) of Fig. 4, in the outer loop, the training/test splittings are done by a Monte Carlo (MC) approach. Specifically, we randomly select 25 test molecules from the dataset, which is stratified into 25 strata based on the levels of the true values of the labels. The stratification helps to minimize the change of the proportions of the dataset compositions upon splitting.[55] In each MC step, a regression model is trained and gives the predictions to the training set and the test set. Therefore, in this work we report the mean and standard deviation of the predictions for each molecule when they are used in the training and test sets from all the MC steps. In total, we evaluate our models with 1000 MC steps for the training/test splittings for the model performance evaluation, and 500 MC steps for generating the learning curves.

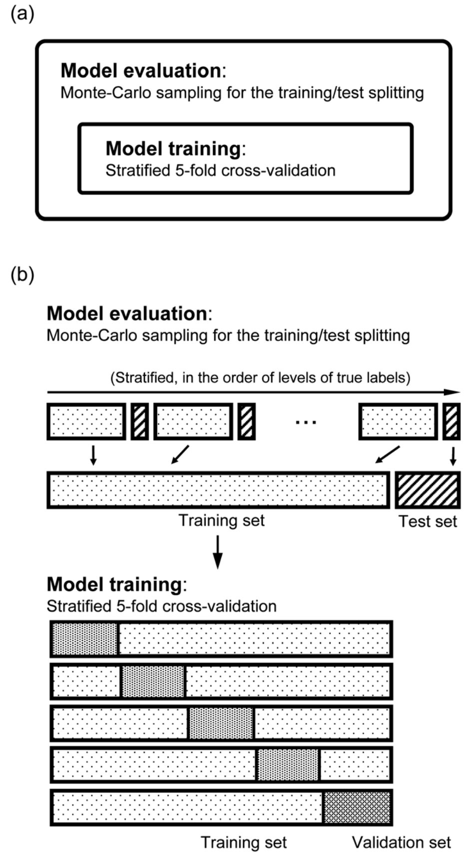

Fig. 4 Scheme of the training/test set splitting in the model evaluation. (a) There are two loops: the outer loop for the model performance evaluation, and the inner loop for the training of model and hyperparameter optimization. (b) In the outer loop, the data are stratified based on the true values of the labels, and each stratum is randomly split into training and test sets. In learning the properties, the training sets are further split into training and validation sets to perform a stratified 5-fold cross-validation.

The performance of the models is evaluated by three different estimators. The first estimator is the mean absolute error (MAE) defined as

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - y_i^*\right|, \tag{11}$$

where y_i^* are the true values, y_i are the predictions, and N is the number of observations. The second estimator is the root mean square error (RMSE), which is given by

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - y_i^*\right)^2}. \tag{12}$$

The last estimator is the normalized error r_E, defined as the ratio of the RMSE to the range of y,

$$r_E = \frac{\text{RMSE}}{y_{\max} - y_{\min}}. \tag{13}$$

4.2 The learning curves

The learning curves show the training and test performance of a model as a function of the training set size N. From the learning curves it is possible to infer the performance of a model by looking at its bias and variance. Similarly, it is possible to understand if the model performance improves with the training set size. For each of the points in the learning curve, the training is performed with 500 different training/test splittings by the MC approach.

5 Results and discussion

5.1 Learning ground state spectroscopic constants

Fueled by the idea of periodicity of molecules (see, *e.g.*, ref. 18 and references in it), we use the group, g_k, and period, p_k, of the atoms within a molecule, *i.e.*, $k = 1, 2$, as input features for a GP regression model to predict different combinations of spectroscopic constants: R_e, ω_e and $\log\left(\frac{D_0}{{R_e}^3 Z_1 Z_2}\right)$, as presented in Section 2. The training sets are permuted before feeding the learning algorithm to reproduce the permutational invariance of relevant properties upon exchanging two atoms in a molecule in the GP regression models.

The GP regression model performance of ground state R_e as a function of input features (g_1, g_2, p_1, p_2) is shown in Fig. 5, where the MAE associated with each of the distinct type of molecules is reported. As a result, most of the molecules are well described by our GP model, as confirmed in the inset of Fig. 5. In particular, it shows little dispersion of the predicted values concerning the true values except for a handful of molecules (transition metal–metal and bi-alkali molecules). To further quantify the GP regression model performance, we calculate the average RMSE of the predicted R_e on 1000 randomly selected test sets leading to 0.0968 ± 0.0070 Å (Table 1), and $r_E = 2.80 \pm 0.20\%$. Our results confirm that the model performance improves as the number of molecules in the training set, N, grows, as it is shown in the learning curve in panel (a) of Fig. 8. Indeed, it is not yet converged for $N = 231$, suggesting that the GP regression model can be further improved by learning more data in the training set.

In learning ω_e, we find $({R_e}^{*-1}, g_1^{\text{iso}}, g_2^{\text{iso}}, p_1, p_2, \bar{g})$ to be the best combination of features, where R_e^* is the predicted equilibrium distance from (g_1, g_2, p_1, p_2), g_k^{iso} encodes the information about the hydrogen isotopes of the k-th atom in the molecule, and $\bar{g}$ is the average of the groups of the two atoms. However, a much better performance is found when the true R_e value is employed. The GP model's performance is shown in the inset of Fig. 6, where it is noticed that the predicted values agree very well with the true values. Indeed, the test set MAE and RMSE are 46.7 ± 0.6 cm^{-1} and 73.4 ± 0.2 cm^{-1}, respectively, while $r_E = 1.80 \pm 0.005\%$, as shown in Table 1. Despite the outstanding performance of our GPR model some molecules are still not well described as shown in Fig. 6. These outliers include HF, DF, and HgH. The large errors predicting ω_e of HF and DF can be attributed to their unique bond mechanism compared to other halogen hydrides.

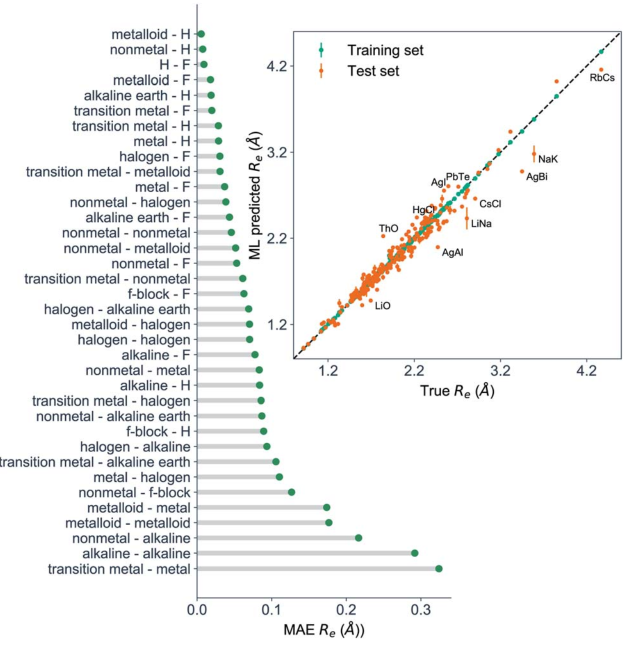

Fig. 5 GP regression performance on predicting R_e using (g_1, g_2, p_1, p_2) as input features classified by the types of the constituent atoms. In particular, the MAE of the test set is reported. The inset shows the test set predictions of R_e *versus* the true values. The values shown are the average of predictions from 1000 MC sampled training/test splittings. The GP regression model gives predictions of the test and training sets. Shown are the mean and standard derivation of each molecule's predictions when used as training data (green symbols) and test data (orange symbols).

Within the features $({R_e}^{-1}, g_1^{\text{iso}}, g_2^{\text{iso}}, p_1, p_2, \bar{g})$, it is interesting that the average of groups $\bar{g} = \frac{g_1 + g_2}{2}$ helps in learning ω_e. In particular, with $\bar{g}$, the MAE of the model reduces around 20% compared with the predictions using $({R_e}^{-1}, g_1^{\text{iso}}, g_2^{\text{iso}}, p_1, p_2)$ as the input feature, as summarized in Table 1. Analogously, the standard deviation of the MC training/test splittings predictions becomes much smaller, suggesting that the model is more robust for different kinds of molecules within the dataset. Actually, by introducing $\bar{g}$, the most significant improvement happens in the descriptions of bi-alkali molecules, where the MAE can be reduced by a factor of 3. The errors predicting HF and DF can also be reduced by a factor of 2, although they are still tricky cases for the model. On the contrary, introducing the average of periods $\bar{p} = \frac{p_1 + p_2}{2}$ does not help improve the model, suggesting that ω_e has a dependency on the total number of valence electrons of the two atoms rather than the number of electron shells.

Motivated by the pioneering work of Anderson, Parr, and coworkers,[27,29,30,45] we study the prediction of $\log\frac{D_0}{{R_e}^3 Z_1 Z_2}$ based on GP regression and the results are shown in Fig. 7. In

Table 1 Regression model predictions of R_e, ω_e, and D_0. g_i and p_i represent the group and period of the i-th atom, respectively. g_i^{iso} stand for the group encoding the information of isotopes of hydrogen, and $\bar{p}$, $\bar{g}$ are the average of groups and periods of the two atoms, respectively

Property	Model	Feature	Test MAE	Test RMSE	Test r_E (%)
R_e (Å)	GPR	(g_1, g_2, p_1, p_2)	0.0662 ± 0.0037	0.0968 ± 0.0070	2.80 ± 0.20
	LR	$\log(Z_1Z_2)$	0.2605 ± 0.0018	0.3591 ± 0.0006	10.41 ± 0.01
ω_e (cm^{-1})	GPR	$({R_e}^{-1}, g_1, g_2, p_1, p_2)$	126.7 ± 2.1	207.2 ± 2.6	5.07 ± 0.06
		$({R_e}^{*-1}, g_1, g_2, p_1, p_2)^a$	152.5 ± 3.6	227.5 ± 4.6	5.56 ± 0.11
		$({R_e}^{-1}, g_1^{\text{iso}}, g_2^{\text{iso}}, p_1, p_2)$	61.5 ± 2.9	142.8 ± 7.0	3.49 ± 0.17
		$({R_e}^{*-1}, g_1^{\text{iso}}, g_2^{\text{iso}}, p_1, p_2)$	96.9 ± 2.9	176.0 ± 13.1	4.30 ± 0.32
		$({R_e}^{-1}, g_1^{\text{iso}}, g_2^{\text{iso}}, p_1, p_2, \bar{p})$	67.5 ± 1.0	151.8 ± 9.5	3.71 ± 0.2
		$({R_e}^{*-1}, g_1^{\text{iso}}, g_2^{\text{iso}}, p_1, p_2, \bar{p})$	101.8 ± 5.4	188.7 ± 25.4	4.61 ± 0.62
		$({R_e}^{-1}, g_1^{\text{iso}}, g_2^{\text{iso}}, p_1, p_2, \bar{g})$	46.7 ± 0.6	73.4 ± 0.2	1.80 ± 0.005
		$({R_e}^{*-1}, g_1^{\text{iso}}, g_2^{\text{iso}}, p_1, p_2, \bar{g})$	81.0 ± 0.82	121.8 ± 0.8	2.98 ± 0.02
	LR	$\sqrt{Z_1Z_2e^{-2R_e}/m}$	376.5 ± 6.6	529.4 ± 1.2	12.95 ± 0.03
		${R_e}^{-2}$	209.6 ± 5.4	297.3 ± 1.4	7.27 ± 0.03
$\log\dfrac{D_0}{{R_e}^3Z_1Z_2}$	GPR	$(R_e, \bar{g}, \bar{p})$	0.249 ± 0.008	0.357 ± 0.007	3.52 ± 0.07
		$(R_e^*, \bar{g}, \bar{p})$	0.270 ± 0.006	0.451 ± 0.007	4.45 ± 0.07
	LR	R_e	0.833 ± 0.004	1.018 ± 0.014	10.03 ± 0.14

[a] R_e^* is the predicted value from (g_1, g_2, p_1, p_2).

particular, in the figure's inset, we show the GP regression model prediction of $\log\dfrac{D_0}{{R_e}^3Z_1Z_2}$ *versus* its true value, which shows a good performance with an RMSE $= 0.357 \pm 0.007$ and a r_E equal to $3.52 \pm 0.07\%$, as shown in Table 1. In this case, the GP is fed with $(R_e, \bar{g}, \bar{p})$ as input features and it shows a fast convergence with respect to the size of training set around $N = 150$ as shown in panel (c) of Fig. 8. The most significant outlier

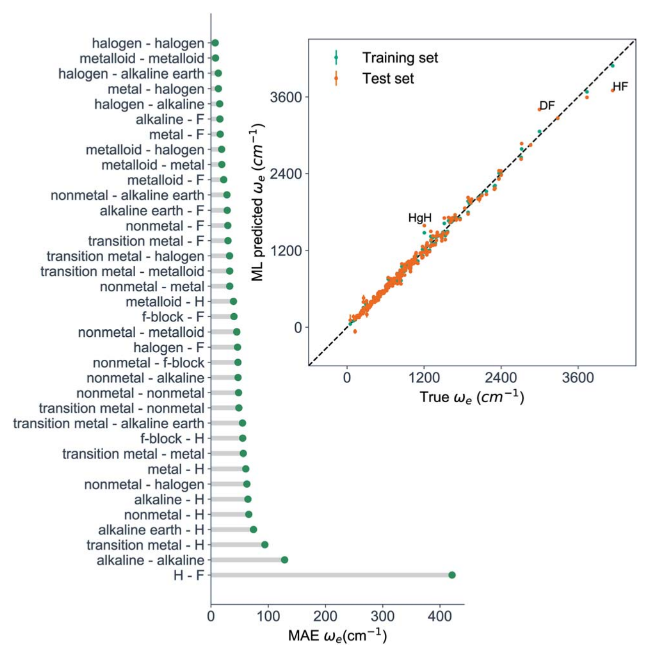

Fig. 6 GP regression performance based on the MAE predicting ω_e for molecules in the test set using $({R_e}^{-1}, g_1^{\text{iso}}, g_2^{\text{iso}}, p_1, p_2, \bar{g})$ as input features classified by the types of the constituent atoms. The inset shows the test set predictions of ω_e compared with respect to the true values. The values shown are the average of predictions from 1000 MC sampled training/test splittings. The GP regression model as learned from the training set gives predictions of the test and training set. Shown are the mean and standard derivation of each molecule's predictions when used as training data (green symbols) and test data (orange symbols).

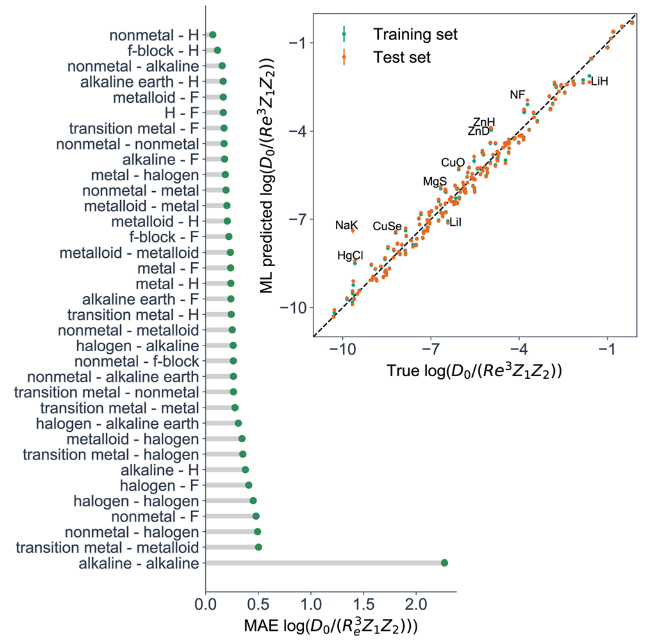

Fig. 7 GP regression performance on predicting $\log\dfrac{D_0}{{R_e}^3Z_1Z_2}$ using $(R_e, \bar{g}, \bar{p})$ as input features classified by the types of the constituent atoms. In particular, the MAE of the test set is reported. The inset shows the test set predictions of $\log\dfrac{D_0}{{R_e}^3Z_1Z_2}$ compared with respect to the true values. The values shown are the average of predictions from 1000 MC sampled training/test splittings. The GP regression model gives predictions of the test and training set. Shown are the mean and standard derivation of each molecule's predictions when used as training data (green symbols) and test data (orange symbols).

Table 2 Regression model predictions of the A excited electronic state R_e and ω_e. g_i and p_i are the groups and periods of the i-th atom, respectively whereas g_i^{iso} stand for the group encoding the information of isotopes of hydrogen. $\bar{p}$, $\bar{g}$ are the average of groups and periods of the two atoms, respectively. R_e(X) and R_e(A) refer to the ground state and A-state R_e, respectively. ω_e(X) refers to the ground state ω_e

Property	Model	Feature	Test MAE	Test RMSE	Test r_E (%)
R_e (Å)	GPR	$(R_e(X), g_1, g_2, p_1, p_2)$	0.0783 ± 0.0018	0.107 ± 0.0026	5.81 ± 0.14
		$(R_e(X), g_1, g_2, p_1, p_2, D(IP, EA))$	0.0691 ± 0.0062	0.098 ± 0.0097	5.32 ± 0.53
ω_e (cm^{-1})	GPR	$(\omega_e(X), {R_e}^{-1}(X), {R_e}^{-1}(A), g_1^{iso}, g_2^{iso}, p_1, p_2, \bar{g})$	71.8 ± 1.4	107.9 ± 4.4	11.3 ± 0.46
		$(\omega_e(X), {R_e}^{-1}(X), {R_e}^{-1}(A), g_1^{iso}, g_2^{iso}, p_1, p_2)$	70.4 ± 0.9	105.1 ± 1.5	11.0 ± 0.15
		$(\omega_e(X), {R_e}^{-1}(X), {R_e}^{-1}(A), g_1, g_2, p_1, p_2)$	70.6 ± 0.9	105.1 ± 1.1	11.0 ± 0.12

for $\log \frac{D_0}{{R_e}^3 Z_1 Z_2}$ is NaK, which is a van der Waals molecule. D_0 of NaK is overestimated and it may be attributed to the fact that NaK is the only bi-alkali molecule in the dataset having D_0. There are also some outliers having first-row elements and 3d transition metals.

A summary of our GP regression model performance for the different combinations of the ground state spectroscopic constants considered in this work is shown in Table 1, compared against the proposed models of Parr, Anderson *et al.*[27,29,30,45] As a result, the GP regression model shows a superior performance against the linear model (LR in the table) based on a particular functional form of the electron density within the molecule. Indeed, the GP performance is, in some cases, five times better than the linear model (in terms of the relative error). Therefore, the group and period (correlated to the number of valence electrons and the number of electrons shells, respectively) of constituent atoms within a molecule encapsulates more valuable information regarding spectroscopic constants than using simple, functional forms for the electron density within the framework of ref. 27, 29, 30 and 45. Indeed, it is interesting to notice that, when predicting R_e and ω_e, one needs groups and periods of each atom in the molecule, whereas $\log \frac{D_0}{{R_e}^3 Z_1 Z_2}$ can be well described only with the average of group and period of the two atoms. Therefore, $\log \frac{D_0}{{R_e}^3 Z_1 Z_2}$ is correlated to groups and periods' additive properties rather than the differences between the two atoms caused by their different groups.

To further examine if our ML approach is generalizable, we have selected 26 molecules out of the dataset and unseen by the ML algorithm including CoO,[56] CrC,[57] InBr,[58] IrSi,[59] MgD,[60] MoC,[61] NbC,[61] NiBr,[62] NiC,[63] NiO,[64] NiS,[65] PbI,[66] PdC,[61] RuC,[61] RuF,[67] ScBr,[62] SnI,[66] TiBr,[62] UF,[68] UO,[69] WC,[70] YC,[61] ZnBr,[62] ZrC,[61] ZrCl,[71] ZrF.[71] The MAE of the GP regression model predicting ground state R_e of the extra test set is 0.066 Å. The average relative error (defined as the absolute errors of each molecule divided by their true R_e) is 3.3%. Indeed, for CrC, InBr, MgD, ZnBr, ZrCl the relative errors are <1%. Within this extra test set, experimental ground state ω_e values are also available for 14 molecules: InBr, MoC, NbC, NiC, NiO, NiS, PbI, PdC, RuC, SnI, UO, WC, YC and ZnBr. The MAE of GPR model predictions is 30 cm^{-1} (4%). For RuC and ZnBr, the relative errors are below 1%, and for NiS and MoC, the relative errors are below 2%. For MoC, NbC, PbI, SnI, YC and ZrC, the experimental binding energy has been reported and the MAE of our GPR model to predict D_0 is 0.32 eV (7.6%). Therefore, our models perform fairly well in this extra test set.

5.2 Learning the first excited state spectroscopic constants

To learn the equilibrium internuclear distance R_e of the A excited electronic state for different molecules, we need to

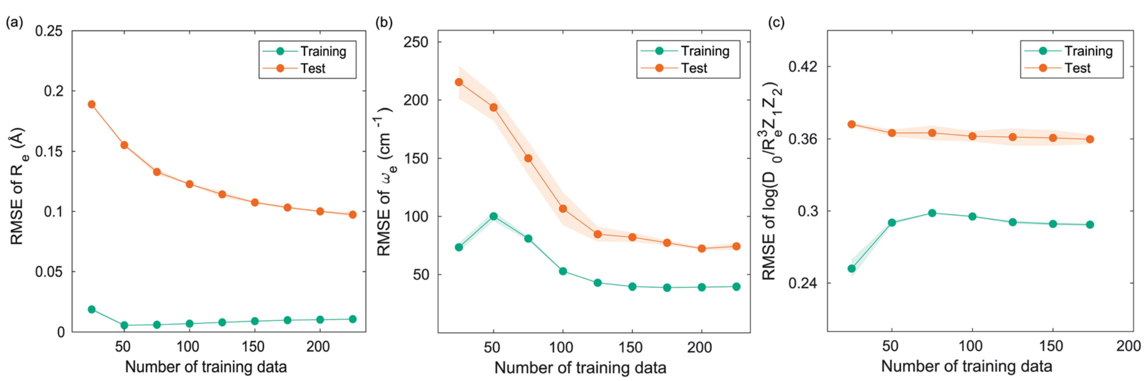

Fig. 8 Performance of the GP regression models as a function of the training set size N. (a) Learning curve of R_e as a function of the size of training set, predicted with the groups and periods of the two atoms, (g_1, g_2, p_1, p_2). (b) Learning curve of ω_e as a function of the size of training set, using the equilibrium internuclear distance R_e, as well as the groups and periods and the average of groups of the two atoms $({R_e}^{-1}, g_1^{iso}, g_2^{iso}, p_1, p_2, \bar{g})$ as the input feature. (c) Learning curve of $\log\left(\frac{D_0}{{R_e}^3 Z_1 Z_2}\right)$ as a function of the size of training set, using the equilibrium internuclear distance R_e, as well as the averages of groups and periods of the two atoms $(R_e, \bar{g}, \bar{p})$ as the input feature. The shade around the points denotes the variance of the errors regarding the MC method.

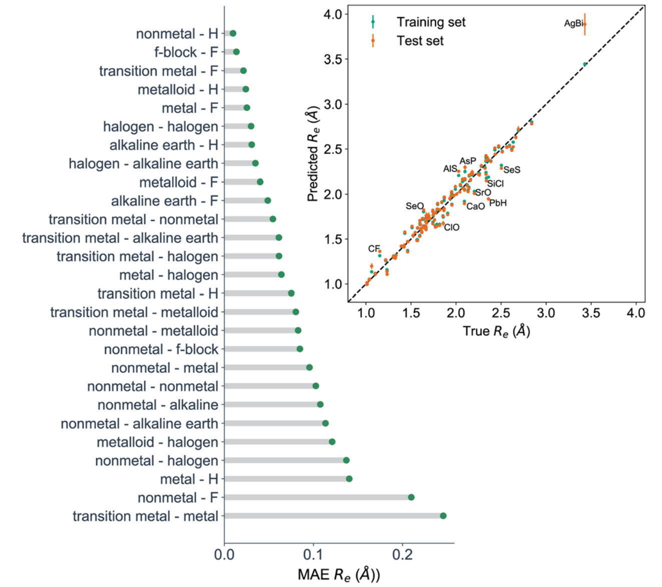

Fig. 9 The test set MAE predicting A excited electronic state R_e by GP regression, using (g_1, g_2, p_1, p_2, R_e(X), D(IP, EA)) as input features, classified by the types of the constituent atoms. The inset shows the test set predictions of the A-excited electronic state R_e compared with respect to the true values. The values shown are the average of predictions from 1000 MC sampled training/test splittings. The GP regression model as learned from the training set gives predictions of the test and training set. Shown are the mean and standard derivation of each molecule's predictions when used as training data (green symbols) and test data (orange symbols).

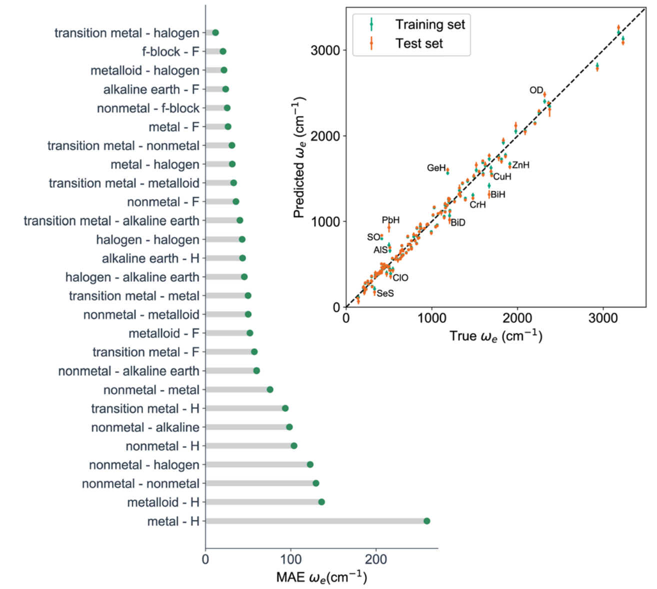

Fig. 10 The test set MAE predicting A excited electronic state ω_e by GP regression, using (ω_e(X), R_e^{-1}(X), R_e^{-1}(A), g_1, g_2, p_1, p_2) as input features, classified by the types of the constituent atoms. The inset shows the test set predictions of A-excited electronic state ω_e compared with respect to the true values. The values shown are the average of predictions from 1000 MC sampled training/test splittings. The GP regression model as learned from the training set gives predictions of the test and training set. Shown are the mean and standard derivation of each molecule's predictions when used as training data (green symbols) and test data (orange symbols).

employ atomic features of the two constituent atoms, including g_1, g_2, p_1, p_2, D(IP, EA), and the ground state R_e(X) when constructing the GP regression models. It is interesting that including D(IP, EA) can improve the predictions (Table 2), which is defined as

$$D(\mathrm{IP},\ \mathrm{EA}) = \begin{cases} \mathrm{EA}_2 - \mathrm{IP}_1, & \text{if } \chi_1 < \chi_2 \\ \mathrm{EA}_1 - \mathrm{IP}_2, & \text{otherwise} \end{cases}$$

where IP_i, EA_i and χ_i are the ionic potential, electron affinity and electronegativity of atom i, respectively. Therefore, D(EA, IP) qualitatively measures the electron transfer between the two constituent atoms. The resultant test set MAE, RMSE and r_E are 0.0691 ± 0.0062\AA, 0.098 ± 0.0097\AA, 5.32 ± 0.53, respectively. As shown in Fig. 9, similar to the results of ground state R_e, the transition metal–metal compounds are the most difficult ones to predict.

For learning ω_e of the A excited electronic state, in addition to the ground state R_e^{-1}(X), it is also necessary to include the A state R_e^{-1}(A). Furthermore, it is better to include the ground state ω_e(X) as the input feature. The results are shown in Fig. 10 in which (ω_e(X), R_e^{-1}(X), R_e^{-1}(A), g_1, g_2, p_1, p_2) leads to a RMSE of 105.1 ± 1.1 cm^{-1} and $r_E = 11.0 \pm 0.12\%$. We also find that including the average of groups $\bar{g}$ or the isotope information cannot further improve the model performance. This is expected, since this information have already been encoded in the ground state ω_e.

The performance of our models predicting the A excited electronic state R_e and ω_e are summarized in Table 2. Compared to the ground state predictions, the errors predicting the A excited electronic state spectroscopic constants are around two times larger, suggesting the difficulty predicting the excited state properties. However, we notice that ω_e is correlated with the inverse of R_e(A) as for ground state molecules. Our findings corroborate the hypothetical relationship between R_e and ω_e in the early times of molecular spectroscopy as it has been introduced in Section 2.

6 Conclusions

In summary, we have shown that using the GP regression model, the main spectroscopic constants of diatomic molecules are related. This result confirms the scenario that Kratzer and Mecke envisioned a century ago.[2,3] The relationships are mostly independent of the nature of the chemical bond of the diatomic molecule. In particular, we have demonstrated that merely using the atoms' group and the period within a molecule as input features can predict particular combinations of spectroscopic constants with an error $r_E < 5\%$. In other words, the spectroscopic constants of diatomic molecules can be efficiently learned from an appropriate dataset by GP regression models, and their values can be accurately predicted. Furthermore, we have shown that GP regression can efficiently learn spectroscopic relationships for excited electronic states of molecules with an error $r_E < 11\%$.

Despite the present GP models' outstanding performance, machine learning methods may be considered mere fitting techniques or as a black-box algorithm that one can hardly learn anything new from them. This statement is not accurate. As an example, here, we emphasize what we have learned from the present machine learning approach:

• It is generally assumed that some molecular properties can be predicted based on the forming atom's positions in the periodic table.[72] However, the predictions are only qualitative rather than quantitative. For instance, it is possible to anticipate the nature of a molecule's bond, but it cannot accurately guess its dissociation energy. However, thanks to ML, we know that it is possible to predict reasonably accurate spectroscopic constants using the constituent atoms' group and period.

• We have learned that ω_e and R_e depend strongly on the number of valence electrons and electrons shells of the atoms forming a molecule, whereas the average number of valence electrons also plays an important role in describing ω_e. $\log \frac{D_0}{R_e{}^3 Z_1 Z_2}$ depends on the average number of valence electrons and average number of electron shells of the molecule.

• The capability of learning excited electronic state properties of diatomic molecules may open the possibility of predicting Franck–Condon factors for interesting transitions regarding direct cooling of molecules.[47,73–76]

Finally, we would like to emphasize that there are around 7000 heteronuclear molecules, and we only utilize 256 of these for our GP regression model. The limited availability of spectroscopic data (only around 3% of possible heteronuclear diatomic molecules) shows the vast amount of spectroscopy that can be done within the realm of diatomic molecules. The more data we have, the more accurate will be the GP regression model predictions before reaching convergence of the learning curve, and the more knowledgeable the community will be about the fundamental properties of diatomic molecules. From our perspective, the present work may motivate data science-driven studies on the field of spectroscopy of diatomic molecules. In particular, it will help to evolve the field of spectroscopy towards the current information era and help to achieve a better understanding on the spectroscopic properties. Furthermore, our results may also bring some insight for the development of features and geometry representations in material science.

Appendix: details about the GP regression models

The choice of covariance functions defines the smoothness of the data points. In learning R_e, the covariance function employed is the exponential kernel defined as

$$k\left(\mathbf{x}_i,\ \mathbf{x}_j|\theta\right) = \sigma_f{}^2 \exp\left(-\frac{r}{l}\right), \tag{14}$$

where σ_f is the signal variance, l is the characteristic length scale, and r is the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$.

In learning ω_e, we use the Matérn class of covariance functions[40]

$$k_{\text{Matern}}(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \frac{\sqrt{2}r^{\nu}}{l} K_\nu \frac{\sqrt{2}r}{l}, \tag{15}$$

with $\nu = 5/2$. K_ν is modified Bessel function in D dimensions, r is the Euclidean distance between x and x', then the Matern 5/2 kernel function is

$$k_{\nu=5/2}(r) = \left(1 + \frac{\sqrt{5}r}{l} + \frac{5r^2}{3l^2}\right)\exp\left(-\frac{\sqrt{5}r}{l}\right). \tag{16}$$

The explicit basis functions in learning R_e are linear basis, while when learning ω_e and $\log\left(\frac{D_e}{R_e{}^3 Z_1 Z_2}\right)$ the basis functions are set to be constant.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

We thank Dr Matthias Rupp for his comments and suggestions and Drs Daniel Thomas and Uwe Hergenhahn for carefully reading the manuscript.

Notes and references

1 G. Herzberg, *Annu. Rev. Phys. Chem.*, 1985, **36**, 1.
2 A. Kratzer, *Z. Phys.*, 1920, **3**, 289.
3 R. Mecke, *Z. Phys.*, 1925, **32**, 823.
4 P. M. Morse, *Phys. Rev.*, 1929, **34**, 57.
5 C. D. Clark, *London, Edinburgh Dublin Philos. Mag. J. Sci.*, 1934, **18**, 459–470.
6 R. M. Badger, *J. Chem. Phys.*, 1933, **2**, 128.
7 C. D. Clark, *London, Edinburgh Dublin Philos. Mag. J. Sci.*, 1935, **19**, 476–485.
8 C. D. Clark and J. L. Stoves, *Nature*, 1934, **133**, 873.
9 W. Gordy, *J. Chem. Phys.*, 1946, **14**, 305–320.
10 K. M. Guggenheimer, *Proc. Phys. Soc.*, 1946, **58**, 456–468.
11 C. H. D. Clark, *Trans. Faraday Soc.*, 1941, **37**, 299–302.
12 C. H. D. Clark and K. R. Webb, *Trans. Faraday Soc.*, 1941, **37**, 293–298.
13 J. W. Linnett, *Trans. Faraday Soc.*, 1940, **36**, 1123–1134.
14 R. Newing, *London, Edinburgh Dublin Philos. Mag. J. Sci.*, 1940, **29**, 298–301.
15 J. W. Linnett, *Trans. Faraday Soc.*, 1942, **38**, 1–9.
16 Y. P. Varshni, *Rev. Mod. Phys.*, 1957, **29**, 664.
17 Y. P. Varshni, *J. Chem. Phys.*, 1958, **28**, 1081.
18 R. Hefferlin, *Periodic systems and their relation to the systematic analysis of molecular data*, The Edwin Mellen Press, Queenston, Canada, 1989.
19 L. Salem, *J. Chem. Phys.*, 1963, **38**, 1227–1236.
20 H. J. Kim and R. G. Parr, *J. Chem. Phys.*, 1964, **41**, 2892–2897.

21 R. F. Borkman and R. G. Parr, *J. Chem. Phys.*, 1968, **48**, 1116–1126.
22 W. T. King, *J. Chem. Phys.*, 1968, **49**, 2866–2867.
23 R. F. Borkman, G. Simons and R. G. Parr, *J. Chem. Phys.*, 1969, **50**, 58–65.
24 P. Politzer, *J. Chem. Phys.*, 1970, **52**, 2157–2158.
25 A. B. Anderson, N. C. Handy and R. G. Parr, *J. Chem. Phys.*, 1969, **50**, 3634–3635.
26 A. B. Anderson and R. G. Parr, *J. Chem. Phys.*, 1970, **53**, 3375–3376.
27 A. B. Anderson and R. G. Parr, *J. Chem. Phys.*, 1971, **55**, 5490–5493.
28 G. Simons and R. G. Parr, *J. Chem. Phys.*, 1971, **55**, 4197–4202.
29 A. B. Anderson, *J. Mol. Spectrosc.*, 1972, **44**, 411–424.
30 J. Gazquez and R. G. Parr, *Chem. Phys. Lett.*, 1979, **66**, 419–422.
31 K. Raghavachari, G. W. Trucks, J. A. Pople and M. Head-Gordon, *Chem. Phys. Lett.*, 1989, **157**, 479–483.
32 R. J. Bartlett, J. Watts, S. Kucharski and J. Noga, *Chem. Phys. Lett.*, 1990, **165**, 513–522.
33 N. Mardirossian and M. Head-Gordon, *Mol. Phys.*, 2017, **115**, 2315–2372.
34 J. P. Perdew, A. Ruzsinszky, J. Tao, V. N. Staroverov, G. E. Scuseria and G. I. Csonka, *J. Chem. Phys.*, 2005, **123**, 062201.
35 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865.
36 J. Tao, J. P. Perdew, V. N. Staroverov and G. E. Scuseria, *Phys. Rev. Lett.*, 2003, **91**, 146401.
37 Y. Zhao and D. G. Truhlar, *Theor. Chem. Acc.*, 2008, **120**, 215–241.
38 X. Liu, G. Meijer and J. Pérez-Ríos, *Phys. Chem. Chem. Phys.*, 2020, **22**, 24191–24200.
39 M. J. Willatt, F. Musil and M. Ceriotti, *Phys. Chem. Chem. Phys.*, 2018, **20**, 29661–29668.
40 C. K. Williams and C. E. Rasmussen, *Gaussian processes for machine learning*, MIT press Cambridge, MA, 2006, vol. 2.
41 S. Glasstone, *Recent advances in physical chemistry*, J. & A. Churchill, London, 2nd edn, 1933, p. 498.
42 M. Davies, *J. Chem. Phys.*, 1949, **17**, 374–379.
43 D. F. Heath, J. W. Linnett and P. J. Wheatley, *Trans. Faraday Soc.*, 1950, **46**, 137–146.
44 R. T. Birge, *Phys. Rev.*, 1925, **25**, 240.
45 A. Anderson and R. Parr, *Chem. Phys. Lett.*, 1971, **10**, 293–296.
46 J. C. Slater, *J. Chem. Phys.*, 1964, **41**, 3199–3204.
47 J. Pérez-Ríos, *An Introduction to Cold and Ultracold Chemistry*, Springer International Publishing, 2020.
48 N. Balakrishnan, *J. Chem. Phys.*, 2016, **145**, 150901.
49 L. D. Carr, D. DeMille, R. V. Krems and J. Ye, *New J. Phys.*, 2009, **11**, 055049.
50 X. Liu, S. Truppe, G. Meijer and J. Pérez-Ríos, *The Diatomic Molecular Spectroscopy Database*, https://rios.mp.fhi.mpg.de/index.php, accessed February 1, 2020.
51 K. P. Huber and G. Herzberg, *Molecular Spectra and Molecular Structure*, Springer-Verlag, Berlin, Germany, 1979.
52 B. M. Smirnov, *Reference Data on Atomic Physics and Atomic Processes*, Springer-Verlag, Berlin, Germany, 2008.
53 X. Liu, S. Truppe, G. Meijer and J. Pérez-Ríos, *J. Cheminf.*, 2020, **12**, 31.
54 *MATLAB, 9.7.0 (R2019b)*, The MathWorks Inc., Natick, Massachusetts, 2019.
55 S. Raschka, 2018, arXiv preprint arXiv:1811.12808.
56 S. McLamarrah, P. Sheridan and L. M. Ziurys, *Chem. Phys. Lett.*, 2005, **414**, 301–306.
57 D. J. Brugh, M. D. Morse, A. Kalemos and A. Mavridis, *J. Chem. Phys.*, 2010, **133**, 034303.
58 S. Mishra, R. K. Yadav, V. Singh and S. Rai, *J. Phys. Chem. Ref. Data*, 2004, **33**, 453–470.
59 M. A. Garcia, C. Vietz, F. Ruipérez, M. D. Morse and I. Infante, *J. Chem. Phys.*, 2013, **138**, 154306.
60 T. C. Steimle, R. Zhang and H. Wang, *J. Chem. Phys.*, 2014, **140**, 224308.
61 R. S. DaBell, R. G. Meyer and M. D. Morse, *J. Chem. Phys.*, 2001, **114**, 2938–2954.
62 M. Burton and L. M. Ziurys, *J. Chem. Phys.*, 2019, **150**, 034303.
63 D. J. Brugh and M. D. Morse, *J. Chem. Phys.*, 2002, **117**, 10703–10714.
64 R. Ram and P. Bernath, *J. Mol. Spectrosc.*, 1992, **155**, 315–325.
65 R. Ram, S. Yu, I. Gordon and P. Bernath, *J. Mol. Spectrosc.*, 2009, **258**, 20–25.
66 C. J. Evans, L.-M. E. Needham, N. R. Walker, H. Köckert, D. P. Zaleski and S. L. Stephens, *J. Chem. Phys.*, 2015, **143**, 244309.
67 T. C. Steimle, W. L. Virgo and T. Ma, *J. Chem. Phys.*, 2006, **124**, 024309.
68 I. O. Antonov and M. C. Heaven, *J. Phys. Chem. A*, 2013, **117**, 9684–9694.
69 L. A. Kaledin, J. E. McCord and M. C. Heaven, *J. Mol. Spectrosc.*, 1994, **164**, 27–65.
70 S. M. Sickafoose, A. W. Smith and M. D. Morse, *J. Chem. Phys.*, 2002, **116**, 993–1002.
71 A. Martinez and M. D. Morse, *J. Chem. Phys.*, 2011, **135**, 024308.
72 C. A. Coulson, *The shape and structure of molecules*, Clarendon Press, Oxford, 1973.
73 M. V. Ivanov, F. H. Bangerter and A. I. Krylov, *Phys. Chem. Chem. Phys.*, 2019, **21**, 19447–19457.
74 M. D. Di Rosa, *Eur. Phys. J. D*, 2004, **31**, 395–402.
75 B. L. Augenbraun, J. M. Doyle, T. Zelevinsky and I. Kozyryev, *Phys. Rev. X*, 2020, **10**, 031022.
76 S. Truppe, S. Marx, S. Kray, M. Doppelbauer, S. Hofsäss, H. C. Schewe, N. Walter, J. Pérez-Ríos, B. G. Sartakov and G. Meijer, *Phys. Rev. A*, 2019, **100**, 052513.