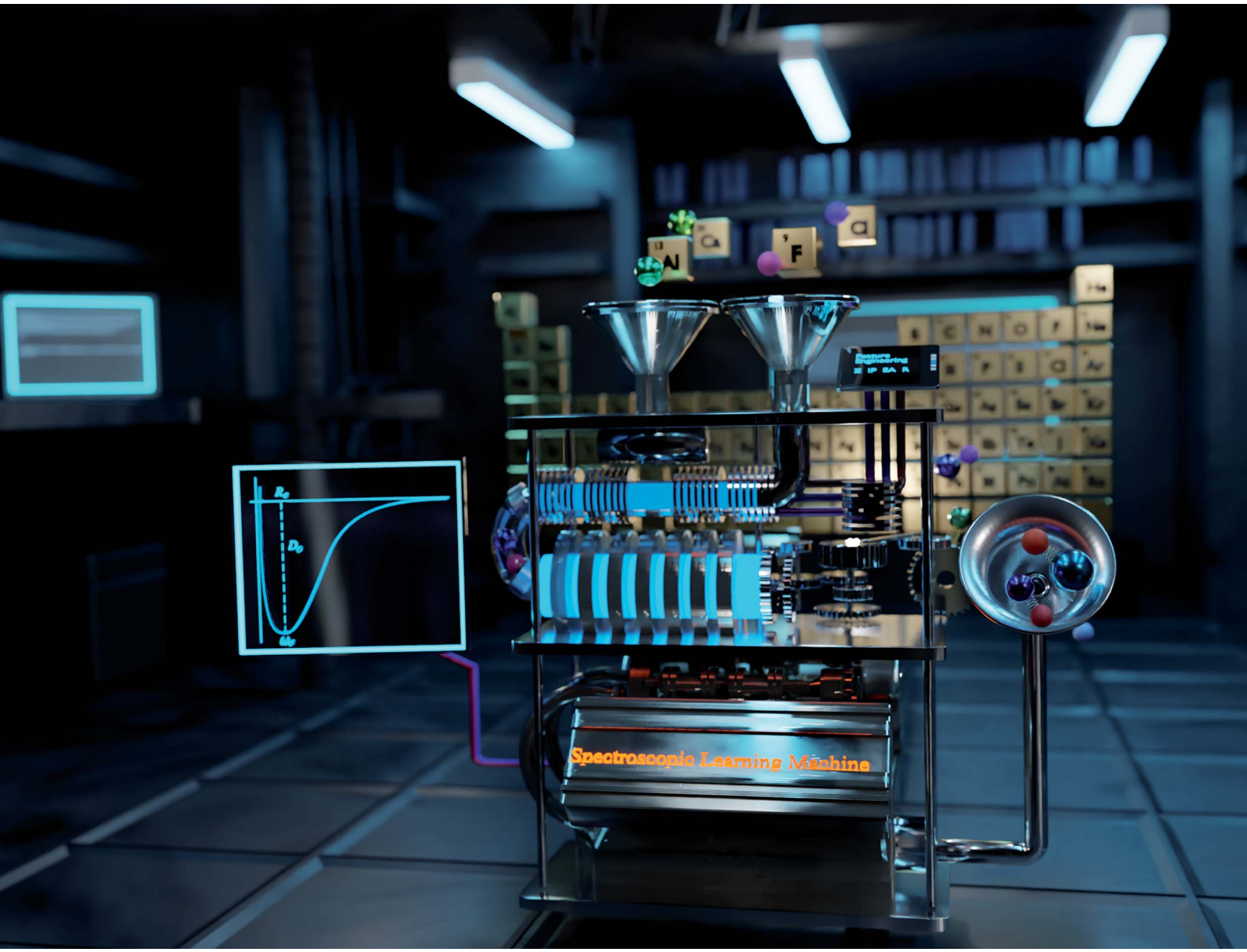
Volume 3
Number 1
January 2024
Pages 1-222

Digital Discovery

rsc.li/digitaldiscovery



ISSN 2635-098X

ROYAL SOCIETY OF CHEMISTRY

PAPER

J. Pérez-Ríos *et al.*
Spectroscopic constants from atomic properties: a machine learning approach

Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2024, **3**, 34

Spectroscopic constants from atomic properties: a machine learning approach

Mahmoud A. E. Ibrahim,[abc] X. Liu [d] and J. Pérez-Ríos *[ab]

We present a machine-learning approach toward predicting spectroscopic constants based on atomic properties. After collecting spectroscopic information on diatomics and generating an extensive database, we employ Gaussian process regression to identify the most efficient characterization of molecules to predict the equilibrium distance, vibrational harmonic frequency, and dissociation energy. As a result, we show that it is possible to predict the equilibrium distance with an absolute error of 0.04 Å and vibrational harmonic frequency with an absolute error of 36 cm^{-1}, including only atomic properties. These results can be improved by including prior information on molecular properties leading to an absolute error of 0.02 Å and 28 cm^{-1} for the equilibrium distance and vibrational harmonic frequency, respectively. In contrast, the dissociation energy is predicted with an absolute error $\lesssim$0.4 eV. Alongside these results, we prove that it is possible to predict spectroscopic constants of homonuclear molecules from the atomic and molecular properties of heteronuclears. Finally, based on our results, we present a new way to classify diatomic molecules beyond chemical bond properties.

Received 14th August 2023
Accepted 31st October 2023

DOI: 10.1039/d3dd00152k

rsc.li/digitaldiscovery

1 Introduction

Since the beginning of molecular spectroscopy in the 1920s, the relationship between spectroscopic constants of diatomic molecules has been an intriguing and captivating matter in chemical physics. Following early attempts by Kratzer, Birge and Mecke,[1–3] Morse proposed a relationship between the equilibrium distance, R_e, and the harmonic vibrational frequency, ω_e, as $R_e^3\omega_e = \gamma$, where γ is a constant, after analyzing the spectral properties of 16 diatomic molecules.[4] However, as more spectroscopic data became available, further examination of the Morse relation revealed its applicability to only a tiny number of diatomic molecules.[5] Next, in a series of papers, Clark *et al.* generalized Morse's idea *via* the concept of a periodic table of diatomic molecules. Eventually, Clark's efforts translated into several relations, each limited to specific classes of molecules.[5–8] Simultaneously, Badger proposed a more neat relationship, including atomic properties of the atoms constituting the molecule.[9] Following Badger's proposal, multiple authors have found new relations, which have seen some utility even for polyatomic molecules.[10–12] Nevertheless, Badger's relations are not generalizable to all diatomic molecules.[13–15] In general, several empirical relationships between R_e and ω_e were proposed in the 1930s and the 1940s.[7,8,16–25] In summary, from 1920 till now, the number of empirical relations published is around 70 collected by Kraka *et al.*[10] Most of these empirical relations were tested by several authors, finding some constraints on their applicability.[10,13–15,26] However, all of these relationships were based on empirical evidence rather than on a given physical or chemical principle.

On the other hand, in 1939 Newing proposed a theoretical justification for observed empirical relationships between spectroscopic constants given by

$$cf(R_e) = \mu\omega_e^2 \tag{1}$$

where c is a constant for similar molecules, $f(R_e)$ is some function of the equilibrium distance, and μ is the reduced mass of the molecule. In particular, Newing used Slater's application of the virial theorem, concluding that the empirical laws may be related to the existence of a universal repulsive field in diatomic molecules.[27,28] The theoretical justification given by Newing implies that several relations of the form given by (1) exist, each of which holds for a set of similar diatomic molecules; however, for any practical application of these empirical laws, the sets of similar diatomic molecules must be identified first. The approach was not viable because similarity needs to be defined precisely.

In the 1960s and 1970s, a number of authors devised the virial theorem, perturbation theory, and Helmann–Feynman theorem[29–32] to develop a better understanding of the nature of the relationship between R_e and ω_e *via* electron densities.[33–40] Most notably, Anderson and Parr were able to establish a relationship between R_e, ω_e, and atomic numbers Z_1 and Z_2, as

[a]*Department of Physics and Astronomy, Stony Brook University, Stony Brook, New York 11794, USA. E-mail: jesus.perezrios@stonybrook.edu*
[b]*Institute for Advanced Computational Science, Stony Brook University, Stony Brook, New York 11794, USA*
[c]*Department of Physics, Faculty of Science, Assiut University, Assiut, 71515, Egypt*
[d]*Fritz-Haber-Institut der Max-Planck-Gesellschaft, D-14195 Berlin, Germany*

$$\omega_e = \sqrt{\frac{4\pi A Z_1 Z_2 \exp(-\xi R_e)}{\mu}}, \quad (2)$$

where μ is the reduced mass of the molecule, and A and ξ (the electron density decay constant) are fitting parameters. Further, assuming that R_e is given by the sum of atomic radii of the constituent atoms and following simple arguments using the electron density function, it can be shown that

$$R_e = \frac{1}{\xi'} \ln\left(\frac{Z_1 Z_2}{B}\right), \quad (3)$$

where it is assumed that the electron density has a given decay constant ξ', and B is a fitting parameter. Using eqn (2) and (3), one finds

$$\omega_e = \sqrt{\frac{C(Z_1 Z_2)^{-\eta}}{\mu}}, \quad (4)$$

where $C = 4\pi A B^{(1+\eta)}$ and $\eta = (\xi' - \xi)/\xi'$. Anderson and Parr found that taking C, ξ and ξ' as functions of the groups and periods of the constituent atoms results in better fitting.[38,40] Anderson and Par tested their relationships against 186 molecules and agreed reasonably with experimental values. Recently Liu *et al.* tested eqn (2) and (3) against an extended data set of 256 molecules, finding that these relationships lead to errors $\gtrsim$10% upon adding more data.[26] Therefore, these relationships are not universal and further study is required to elucidate proper relationships. However, the pioneering work of Anderson, Parr, and others provided well-motivated relationships between spectroscopic constants theoretically for the first time. Most significantly, their work pointed towards a possible direct connection between a diatomic molecule's spectroscopic properties and its individual atoms' atomic properties and positions in the periodic table.

Alongside these developments, several authors attempted connecting the dissociation energy, D_0^0, with ω_e and R_e of diatomic molecules.[19,41–45] However, these received little attention due to the lack of reliable experimental data.[9,41,46–48] Most of the relationships are given by

$$D_0^0 = A' \mu {\omega_e}^2 {R_e}^l \quad (5)$$

where A' and l are constants depending on the form and parameterization of the potential energy functions that describe the molecule. For instance, Sutherland found that by taking A' as a function of groups and periods, better results can be obtained.[19,41]

Thanks to machine learning (ML) techniques and the development of extensive spectroscopic databases,[49] it has been possible to study the relationship between spectroscopic constants from a heuristic perspective, *i.e.*, from a data-driven approach,[26] find optimal potentials based on spectroscopy data[50] and to improve *ab initio* potentials to match experimental observations.[51] In particular, Gaussian process regression (GPR) models have been used on a large dataset of 256 heteronuclear diatomic molecules. As a result, it was possible to predict R_e from the atomic properties of the constituent atoms. Similarly, ω_e and the binding energy D_0^0 were predicted using combinations of atomic and molecular properties. However, the work of Liu *et al.* only studied heteronuclear molecules. Hence, the universality of the relationship between spectroscopic constants still needs to be revised. On the other hand, ML techniques can be used to boost density functional theory approaches to larger systems with low computational costs.[52–55] Hence, ML techniques are used to enlarge the capabilities of quantum chemistry methods. However, if sufficient data and information are available, could ML substitute quantum chemistry methods?

In this work, we present a novel study on the relationship between spectroscopic constants *via* ML models, including homonuclear molecules in a dataset of 339 molecules: the largest dataset of diatomics ever used. As a result, first, we show that it is possible to predict R_e and ω_e with mean absolute errors $\sim$0.026 Å and $\sim$26 cm^{-1}, leading to an improvement of factor 2 in predicted power and accuracy concerning previous ML models. Furthermore, the dissociation energy, D_0^0, is predicted with a mean absolute error $\sim$0.4 eV, in accordance with previous ML models. However, our model can benefit from having a more accurate and extensive database. Second, we show that it is possible to accurately predict the molecular properties of homonuclear molecules out of heteronuclear ones. Finally, since we use the same ML model in this work, we are in a unique situation to define similarity among molecules. Thus, we propose a data-driven classification of molecules. The paper is organized as follows: in Section 2, we introduce the database and analyze the main properties; in Section 3, we present the ML models with a particular emphasis on Gaussian process regression; in Section 4, we present our results and in Section 5, the conclusions.

2 The data set

In this work, we extend the data set by Liu *et al.*[26,49] by adding the ground state spectroscopic constants of 32 homonuclear and 54 extra heteronuclear diatomic molecules from ref. 56–127. The dataset counts 338 entries based on experimentally determined spectroscopic constants: R_e is available for 338, ω_e for 327, and D_0^0 is available for 250 molecules.

To assess the variation of the spectroscopic constants in the dataset, we display the histogram and box plots of R_e, ω_e, and D_0^0 in Fig. 1. This Figure shows that the spectroscopic constants' histogram is nearly unimodal. However, R_e and ω_e show a heavy-tailed distribution. In the case of R_e, the tail is due to the presence of van der Waals molecules. In contrast, light molecules are responsible for the tail in the histogram for ω_e. On the other hand, the box plot of D_0^0 shows almost no outliers and only an incipient peak for a molecule with binding energy smaller than 0.75 eV due to the presence of van der Waals molecules. On the other hand, we investigate the relationship between pairs of spectroscopic constants in panels (d)–(f) of Fig. 1. For example, panel (d) displays R_e *versus* ω_e, showing an exponential trend similar to the one suggested by eqn (2) or a power law (Morse relationship). On the contrary, by plotting R_e *versus* D_0^0 and D_0^0 *versus* ω_e in panels (e) and (f), respectively, we notice a large dispersion of the points with no observed trends in both panels.

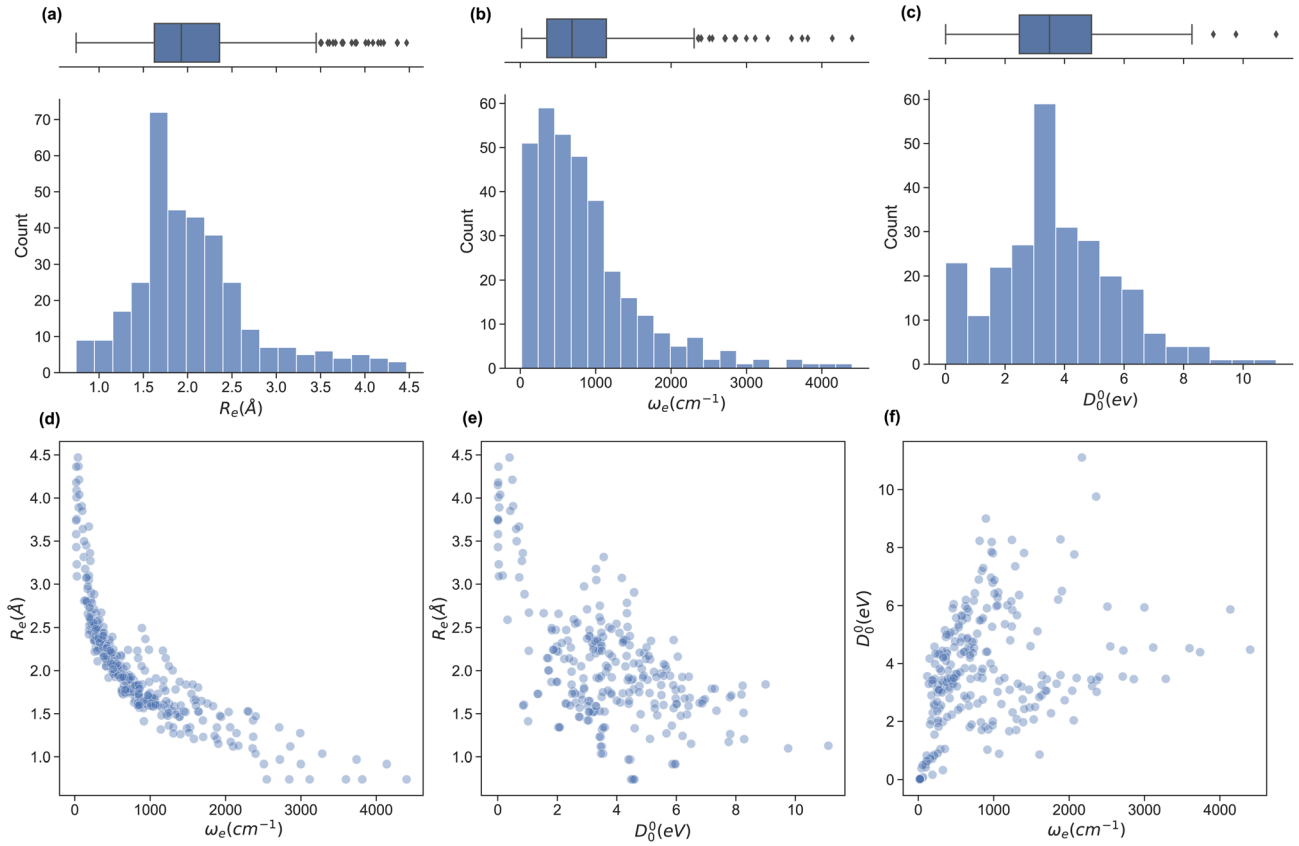


Fig. 1 The dataset of diatomic molecules' ground state spectroscopic constants. Panels (a–c) display the distribution of the main spectroscopic constants in the dataset-R_e, ω_e and D_0^0 – *via* a histogram representation and a box plot (at the top) for each. Panels (d–f) show the relationship between different spectroscopic constants of the molecules in the database.

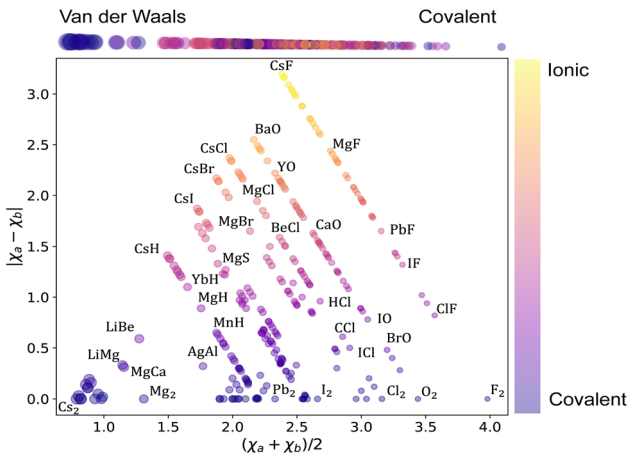


Fig. 2 Arkel–Ketelaar's triangle of the dataset. The average electronegativity of the constituent atoms on the *x*-axis, the difference in electronegativity of the constituent atoms correlates with the ionic character on the *y*-axis. The color of each circle demonstrates the ionic character of the corresponding molecule following the color bar on the right of the figure. The size of the circles differentiates between covalent (smaller circles) and van der Waals (larger circles) molecules, as illustrated at the top of the figure.

Next, we analyze the chemical properties of the molecules under consideration, employing the Arkel–Ketelaar triangle – also known as the Jensen triangle, which separates qualitatively covalent, ionic, and van der Waals molecules. The triangle plots the absolute value of the electronegativity difference between the constituent atoms $|\chi_a - \chi_b|$ *versus* their average electronegativity, as shown in Fig. 2, where χ_a and χ_b denote the electronegativities of the molecules' constituent atoms. The average electronegativity of the constituent atoms on the *x*-axis quantifies the van der Waals-covalent bonding. On the contrary, the difference in electronegativity of the constituent atoms quantifies the ionic character on the *y*-axis. The triangle shows that the data set comprises chemically diverse diatomic molecules with bonding characters ranging from covalent to ionic with many van der Waals. This chemical diversity strongly manifests itself in the range of the ground state spectroscopic constants depicted in Fig. 1.

3 The machine learning (ML) model

Machine learning (ML) is a vast discipline that utilizes data-driven algorithms to perform specific tasks (*e.g.*, classification, regression, clustering). Among the different ML techniques, in this work, we use Gaussian process regression (GPR), which is

especially suitable for small datasets. This section briefly describes GPR and our methods to generate and evaluate models.

3.1 Gaussian process regression

We define our data set $D = \{(\mathbf{x}_i, y_i)|i = 1,\cdots, n\}$, where $\mathbf{x}_i$ is a feature vector of some dimension D associated with the i-th element of the dataset, y_i is a scalar target label, and n is the number of observations, *i.e.*, the number of elements in the dataset. The set of all feature vectors and corresponding labels can be grouped in the random variables X and $\mathbf{y}$, respectively, where $X = (\mathbf{x}_1,\cdots, \mathbf{x}_n)$ and $\mathbf{y} = (y_1,\cdots, y_n)$. Here, $\mathbf{y}$ consists of values of molecular properties to be learned. y_i is R_e, ω_e, or D_0^0 of the i-th molecule, whereas $\mathbf{x}_i$ is a vector containing atomic or molecular properties of the same molecule.

We are interested in mapping features to target labels *via* a regression model $y_i = f(\mathbf{x}_i) + \varepsilon_i$, where $f(\mathbf{x}_i)$ is the regression function, and ε_i is an additive noise. We further assume that ε_i follows an independent, identically distributed (i.i.d.) Gaussian distribution with variance ${\sigma_n}^2$

$$\boldsymbol{\varepsilon} \sim \mathcal{N}\left(0, {\sigma_y}^2 I\right) \tag{6}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1,\cdots, \varepsilon_n)$ and I is the identity matrix.

One approach to tackle the regression problem is to specify a functional form of $f(\mathbf{x}_i)$. Then, set the free parameters of the regression model by fitting the data. Alternatively, one can disregard specifying a functional form of $f(\mathbf{x}_i)$ but instead place a prior distribution over a space of functions and infer the posterior predictive distribution following a Bayesian non-parametric approach. Within this group of methods, we find Gaussian process regression (GPR), assuming a Gaussian process prior $\mathcal{GP}$ over the space of functions.[128,129]

$$f(\mathbf{x}) \sim \mathcal{GP}\left(m(\mathbf{x}), k\left(\mathbf{x}, \mathbf{x}'\right)\right). \tag{7}$$

A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution. A Gaussian process is specified by a mean function $m(\mathbf{x})$ and a covariance function (kernel) $k(\mathbf{x}, \mathbf{x}')$, we will describe both shortly. A posterior distribution of the value of $f(\mathbf{x}^*)$ at some point of interest, $\mathbf{x}^*$, is determined through the Bayes theorem as

$$f(\mathbf{x}^*)|X, Y \sim \mathcal{N}(\mu^*, \Sigma^*) \tag{8}$$

where

$$\begin{aligned} \mu^* &= m(\mathbf{x}^*) + k(\mathbf{x}^*, X)\left[k(X, X) + {\sigma_n}^2 I\right]^{-1}(\mathbf{y} - m(X)). \\ \Sigma^* &= k(\mathbf{x}^*, X)\left[k(X, X) + {\sigma_n}^2 I\right]k(X, \mathbf{x}^*). \end{aligned} \tag{9}$$

The mean of the resulting predictive posterior distribution, μ^*, is used to obtain a point estimate of the value of $f(\mathbf{x}^*)$, and its covariance Σ^* provides a confidence interval.

In GPR, the regression model is completely specified by the kernel $k(\mathbf{x}, \mathbf{x}')$. The kernel is a similarity measure that specifies the correlation between a pair of values $f(\mathbf{x})$ and $f(\mathbf{x}')$ by only using the distance between a pair of feature vectors $\mathbf{x}$ and $\mathbf{x}'$ as its input variable. Specifying a kernel, we encode high-level structural assumptions (*e.g.*, smoothness, periodicity, *etc.*) about the regression function. Here, we focus on the Matérn class kernel given by given by

$$k\left(\mathbf{x}_p, \mathbf{x}_q\right) = {\sigma_f}^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \frac{\sqrt{2}d\left(\mathbf{x}_p, \mathbf{x}_q\right)^{\nu}}{l} K_{\nu}\left(\frac{\sqrt{2}d\left(\mathbf{x}_p, \mathbf{x}_q\right)^{\nu}}{l}\right) + {\sigma_n}^2 \delta_{pq}, \tag{10}$$

where $d(\mathbf{x}_p, \mathbf{x}_q)$ is the Euclidean distance between $\mathbf{x}$ and $\mathbf{x}'$, $K_\nu(z)$ is the modified Bessel function of the second kind of order ν and argument z, $\Gamma(x)$ is the Euler gamma function of argument x, l is the characteristic length scale, ${\sigma_f}^2$ is the signal variance, and δ_{pq} is the Kronecker delta. ν controls the smoothness of the process. For instance, for $\nu = 1/2$, the process is zero times differentiable. On the contrary, the process is infinitely differentiable at the limit $\nu \to \infty$: the so-called radial basis function (RBF) kernel. Values of ν that are suitable for regression applications are 1/2, 3/2, 5/2, and ∞.[128]

We can encode a physical model *via* the relationships between spectroscopic constants by specifying the Gaussian process prior mean function $m(\mathbf{x})$. A common choice of the prior mean function is $m(\mathbf{x}) = 0$. This choice is satisfactory in most cases, especially in interpolation tasks. However, selecting an appropriate prior mean function can simplify the learning process (delivering better results using fewer data). The mean function can also guide the model for better predictions as $k(\mathbf{x}_p, \mathbf{x}_q) \to 0$; this is necessary for extrapolation and interpolation among sparse data points. Further, a model with a specified mean function is more interpretable.

3.2 Model development and performance evaluation

Its parameters and hyperparameters characterize a GPR model. Parameters involve (σ_n, l, σ_f) of eqn (10) plus additional parameters depending on the form of the prior mean function. On the contrary, hyperparameters involve selecting features, the form of the prior mean function, and the order ν of the Matérn kernel. To determine the parameters and the hyperparameters, we divide the dataset D into two subsets: D_{tv} and D_{test}. First, D_{tv} is used for the training and validation stage, in which we determine the model's hyperparameters. Then, D_{test}, known as the test set, is left out for model final testing and evaluation and does not take any part in determining the parameters nor the hyperparameters of the model.

To design a model, we choose an X suitable to learn $\mathbf{y}$ through a GPR. We then choose a convenient prior mean function $m(X)$ based on physical intuition, alongside the last hyperparameter $\nu \in \{1/2, 3/2, 5/2, \infty\}$ is determined by running four models, each with a possible value of ν, and we chose the one that performs the best on the training data to be the final model. Precisely, a cross-validation (CV) scheme is used to evaluate the performance of each model iteratively: we split D_{tv} into a training set D_{train} (~90% of D_{tv}) and a validation set D_{valid}. We use D_{train} to fit the model and determine its parameters by maximizing the log-marginal likelihood. The fitted model is then used to predict the target labels of D_{valid}. We repeat the process with a different split in each iteration such that each element in D_{tv} has been sampled at least once in both D_{train} and

D_{valid}. After many iterations, we can determine the average performance of the model. We compare the average performance of the four models after the CV process. Finally, We determine the value of ν to be its value for the best-performing model.

We adopt a Monte Carlo (MC) splitting scheme to generate the CV splits. Using the MC splitting scheme, we expose the models to various data compositions, and thus, we can make more confident judgments about our models' performance and generality.[26] To generate a single split, we use stratified sampling.[129,130] First, we stratify the training set into smaller strata based on the target label. Stratification will be such that molecules in each stratum have values within some lower and upper bounds of the target label (spectroscopic constant) of interest. Then, we sample the validation set so that each stratum is represented. Stratified sampling minimizes the change in the proportions of the data set composition upon MC splitting, ensuring that the trained model can make predictions over the full range of the target variable. Using the Monte Carlo splitting scheme with cross-validation (MC-CV) allows our models to train on D_{tv} in full, as well as make predictions for each molecule in D_{tv}. In each iteration, D_{valid} simulates the testing set; thus, by the end of the MC-CV process, it provides an evaluation of the model performance against ∼90% of the molecules in the data set before the final testing stage.

We use 1000 MC-CV iterations to evaluate each model's performance. Two estimators evaluate the models' performance at each iteration, the mean absolute error (MAE) and the root mean squared error (RMSE), given by

$$\begin{aligned} \text{MAE} &= \frac{1}{N}\sum_{i=1}^{N}|y_i - y_i^*|, \\ \text{RMSE} &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - y_i^*\right)^2} \end{aligned} \tag{11}$$

where y_i^* and y_i are the true and predicted values, respectively, and N is the number of observations in consideration. We report the final training/validation $\overline{\text{MAE}}$ and $\overline{\text{RMSE}}$ with the sample standard deviation (STD) and the standard error of the means (SEM) given by

$$\begin{aligned} \overline{\text{MAE}} &= \frac{1}{M}\sum_{i=1}^{M}\text{MAE}_i, \\ \overline{\text{RMSE}} &= \frac{1}{M}\sum_{i=1}^{M}\text{RMSE}_i \end{aligned} \tag{12}$$

where M is the number of the MC-CV iterations, and MAE_i (RMSE_i) is the MAE (RMSE) of the ith MC-CV iteration.

$$\begin{aligned} \text{STD(MAE)} &= \sqrt{\frac{1}{M-1}\sum_{i=1}^{M}\left(\overline{\text{MAE}} - \text{MAE}_i\right)^2}, \\ \text{STD(RMSE)} &= \sqrt{\frac{1}{M-1}\sum_{i=1}^{M}\left(\overline{\text{RMSE}} - \text{RMSE}_i\right)^2} \end{aligned} \tag{13}$$

$$\begin{aligned} \text{SEM(MAE)} &= \frac{\text{STD(MAE)}}{\sqrt{M}}, \\ \text{SEM(RMSE)} &= \frac{\text{STD(RMSE)}}{\sqrt{M}} \end{aligned} \tag{14}$$

We use learning curves to evaluate the performance of the models as a function of the size of D_{train}. We use 500 CV splits at each training set size to generate the learning curves. The validation and training $\overline{\text{RMSE}} \pm 0.5\text{STD(RMSE)}$ are plotted as a function of the size of D_{train}.

Models that have the lowest validation $\overline{\text{MAE}}$, $\overline{\text{RMSE}}$, and SEM are elected for the testing stage. In the testing stage, we fit the models to D_{tv} and make predictions of the target labels of D_{test}. Finally, we report the validation $\overline{\text{MAE}} \pm \text{SEM(MAE)}$ and $\overline{\text{RMSE}} \pm \text{SEM(RMSE)}$ and test MAE and RMSE as our final evaluation of the model.

4 Results and discussion

We have developed seven new models to predict R_e, ω_e, and D_0^0: r2, r3, and r4 to predict R_e, models for predicting ω_e are denoted w2, w3, and w4, while only one model is used to predict D_0^0, labeled as d1. In addition, we implemented two of the best-performing models of Liu *et al.*[26] (denoted r1 and w1) using our updated dataset and compared them with our models. All models are divided into three categories: (i) r1, r2, and w2 only employ atomic properties as features in the kernel and as variables in the prior mean function, (ii) r3 and w3 employ atomic properties as features in the kernel but use spectroscopic data in the prior mean function, and (iii) r4, w4, and d1 include spectroscopic data both in the kernel and the prior mean function.

In all the seven newly developed models, we use the groups g_1 and g_2 and periods p_1 and p_2 of the molecules' constituent atoms and the square root of the reduced mass of the molecule $\mu^{1/2}$ as features. Therefore, the set of properties $\{p_1, g_1, p_2, g_2, \mu^{1/2}\}$ uniquely defines each molecule in the dataset. On the contrary, additional spectroscopic properties are added to these five features for models within the category (iii). Furthermore, we choose the models' features and prior mean functions using physical intuition based on the discussion in the introduction and observations from the data Fig. 1, and ν was set to 3/2 using the CV scheme discussed in the last section. The models' characteristics are given in Table 1.

For all the nine implemented models, we permute the groups and periods in D_{train} in the training and validation stage and in D_{tv} in the testing stage to impose permutational invariance.[26] That is, the models should not differentiate between $\mathbf{x} = (p_1, g_1, p_2, g_2, \ldots)$ and $\mathbf{x}' = (p_2, g_2, p_1, g_1, \ldots)$ upon exchanging the numbering of the two atoms in a molecule. Eight of the models use linear prior mean functions, the linear coefficients of which are determined by fitting the linear model to D_{train} in each CV iteration in the training and validation stage and by fitting to D_{tv} in the testing stage.

For the sake of comparison with baseline ML models we have implemented linear regression (LR) models based on eqn

Table 1 Machine learning models summary. The target column includes the molecular property to be predicted. The model column refers to the ML model used. The molecules column refers to the number of molecules in the training plus validation set D_{tv}. Features are the atomic and molecular properties employed to characterize every data point in the kernel. Prior mean stands for the prior mean function used for each model as indicated in the text, and ν represents the order of the Matérn kernel determined *via* the MC-CV scheme described in Section 3.2

Target	Model	Molecules	Features	Prior mean	ν
R_e (Å)	rlr1	314	$p_1 + p_2$, $g_1 + g_2$, $\ln(Z_1Z_2)$	—	—
	rlr2	308	$\ln(\omega_e)$, $p_1 + p_2$, $g_1 + g_2$, $\ln(Z_1Z_2)$, $\ln(\mu)$	—	—
	svmr1	314	p_1, g_1, p_2, g_2	—	3/2
	svmr2	314	p_1, g_1, p_2, g_2, $\mu^{1/2}$	—	3/2
	svmr3	308	$\ln(\omega_e)$, p_1, g_1, p_2, g_2, $\mu^{1/2}$	—	3/2
	r1	314	p_1, g_1, p_2, g_2	m_{r1}	1/2
	r2	314	p_1, g_1, p_2, g_2, $\mu^{1/2}$	m_{r2}	3/2
	r3	308	p_1, g_1, p_2, g_2, $\mu^{1/2}$	m_{r3}	3/2
	r4	308	$\ln(\omega_e)$, p_1, g_1, p_2, g_2, $\mu^{1/2}$	m_{r3}	3/2
$\ln(\omega_e)$	wlr1	308	$p_1 + p_2$, $g_1 + g_2$, $\ln(Z_1Z_2)$, $\ln(\mu)$	—	—
	wlr2	308	R_e, $p_1 + p_2$, $g_1 + g_2$, $\ln(Z_1Z_2)$, $\ln(\mu)$	—	—
	svmw1	308	p_1, g_1, p_2, g_2, $\mu^{1/2}$	—	3/2
	svmw2	308	R_e, p_1, g_1, p_2, g_2, $\mu^{1/2}$	—	3/2
	w1	308	R_e^{-1}, p_1, g_1^{iso}, p_2, g_2^{iso}, $\bar{g}$	0	5/2
	w2	308	p_1, g_1, p_2, g_2, $\mu^{1/2}$	m_{w2}	3/2
	w3	308	p_1, g_1, p_2, g_2, $\mu^{1/2}$	m_{w3}	3/2
	w4	308	R_e, p_1, g_1, p_2, g_2, $\mu^{1/2}$	m_{w4}	3/2
$\ln(D_0^0)$	dlr1	244 $\ln(R_e)$, $\ln(\omega_e)$, $p_1 + p_2$, $g_1 + g_2$, $\ln(\mu)$	—	—	
	svmd1	244	p_1, g_1, p_2, g_2	—	3/2
	d1	244	p_1, g_1, p_2, g_2, $\mu^{1/2}$	m_{d1}	3/2

(3)–(5). Specifically, models rlr1 and rlr2 to predict R_e, wlr1 and wlr2 to predict $\ln(\omega_e)$ and dlr1 to predict $\ln(D_0^0)$. The same MC-CV scheme used to train the GPR models was used to train the LR models. Further, we train support vector machines (SVM) models for regressions tasks to predict R_e, ω_e and D_0^0. The hyperparameters of the Matern 3/2 kernels for each SVM model are tuned *via* 1000 MC-CV steps using Bayesian optimization.[131] A description of these models is given in Table 1 and a statistical summary of their performance is given in Table 2.

4.1 R_e

The first spectroscopic constant under consideration is the equilibrium distance, R_e. We have implemented and developed two models for predicting R_e using only atomic properties: r1 and r2, as detailed in Table 1. Model r1 is the same as in Liu *et al.*[26] using groups and periods of the constituent atoms as features. The model r2 requires an extra feature related to the reduced mass of the molecule. For both models, we explicitly express the models' prior mean functions as linear functions in

Table 2 Statistical summary of the performance of the ML models using different features, kernels, and prior mean functions as listed in Table 1. The performance of each model is evaluated using both the validation and test scores. The values for MAE and RMSE with * show an SEM $\lesssim$0.001 Å

Target	Model	Validation $\overline{MAE} \pm$ SEM	Validation $\overline{RMSE} \pm$ SEM	Test MAE	Test RMSE
R_e (Å)	rlr1	0.33	0.54	—	—
	rlr2	0.112	0.146	—	—
	svmr1	0.043	0.069	0.044	0.068
	svmr2	0.039	0.059	0.046	0.068
	svmr3	0.025	0.038	0.025	0.037
	r1	0.060*	0.100*	0.047	0.070
	r2	0.041*	0.060*	0.046	0.066
	r3	0.027*	0.039*	0.027	0.038
	r4	0.026*	0.038*	0.027	0.040
ω_e (cm^{-1})	wlr1	218	316	—	—
	wlr2	118	197	—	—
	svmw1	39.4	65.2	36.4	53.7
	svmw2	25.8	42.3	24.7	31.8
	w1	33.2 ± 0.3	64.8 ± 1.0	33.5	61.2
	w2	40.3 ± 0.3	66.3 ± 0.6	37.9	53.4
	w3	27.7 ± 0.2	44.8 ± 0.4 31.3	39.35	
	w4	25.9 ± 0.2	41.6 ± 0.3	26.9	33.6
D_0^0 (eV)	dlr1	0.98	1.25	—	—
	svmd1	0.36	0.57	0.79	0.83
	d1	0.37 ± 0.002	0.52 ± 0.003	0.55	0.72

the groups and periods of the diatomic molecules' constituent atoms.

$$m_{\mathrm{r1-r2}} = \beta_0^{\mathrm{r1-r2}} + \beta_1^{\mathrm{r1-r2}}(\mathrm{p_1} + \mathrm{p_2}) + \beta_2 \mathrm{r1{-}r2}(\mathrm{g_1} + \mathrm{g_2}), \quad (15)$$

where $\beta_k^{\mathrm{r1-r2}}$, $k \in \{0, 1, 2\}$ are the linear coefficients of $m_{\mathrm{r1-r2}}$.

A comparison between models r1 and r2 is displayed in Fig. 3. The scatter plots show a more significant dispersion of the predictions for model r1 compared to model r2. Both models show the same outliers: homonuclear and van der Waals molecules. However, for model r2, the number of outliers is smaller than in model r1, and their dispersion from the true line is significantly suppressed. As a result, model r2 performs substantially better, especially in predicting molecules with $R_e \geq 3$ Å (mainly van der Waals molecules). The learning curves of models r1 and r2, displayed in panels (d) and (e) of Fig. 3, respectively, show a convergent validation curve towards the training set result as the size of the training set increases, indicative of the learning capability of the model, although, model r2 displays a faster convergence, indicating that the model learns more efficiently. Overall, model r2 shows an improvement in the prediction on $R_e \sim 20\%$ with respect r1 as shown in Table 2, leading to validation $\overline{\mathrm{MAE}}$ and $\overline{\mathrm{RMSE}}$ of 0.041 Å and 0.060 Å, respectively.

Motivated by previously proposed relationships between R_e and $\ln(\omega_e)$, we introduce models r3 and r4. Model r3 employs the same features as model r2 but incorporates spectroscopic information in the prior mean function. On the contrary, model r4 uses $\ln(\omega_e)$ as a feature. Both models have a prior mean given by

$$m_{\mathrm{r3-r4}} = \beta_0^{\mathrm{r3-r4}} + \beta_1^{\mathrm{r3-r4}}(\mathrm{p_1} + \mathrm{p_2}) + \beta_2^{\mathrm{r3-r4}}(\mathrm{g_1} + \mathrm{g_2}) + \beta_3^{\mathrm{r3-r4}} \ln(\mu^{1/2}) + \beta_4^{\mathrm{r3-r4}} \ln(\omega_e), \quad (16)$$

where $\beta_k^{\mathrm{r3-r4}}$, $k \in \{0, 1, 2, 3, 4\}$ are linear coefficients of $m_{\mathrm{r3-r4}}$. The two models have similar performance as shown in Table 2. The results of model r3 are presented in panels (c) and (f) of Fig. 3. Panel (c) shows a minimal scatter around the true line. The error bars are suppressed compared with panels (a) and (b) of the same figure, indicating a higher confidence level of the model's predictions. The validation curve in panel (f) shows that the learning rate of model r3 is significantly higher than the other two models. Using 50% of the available data is sufficient for model r3 to achieve a validation $\overline{\mathrm{RMSE}}$ comparable to model

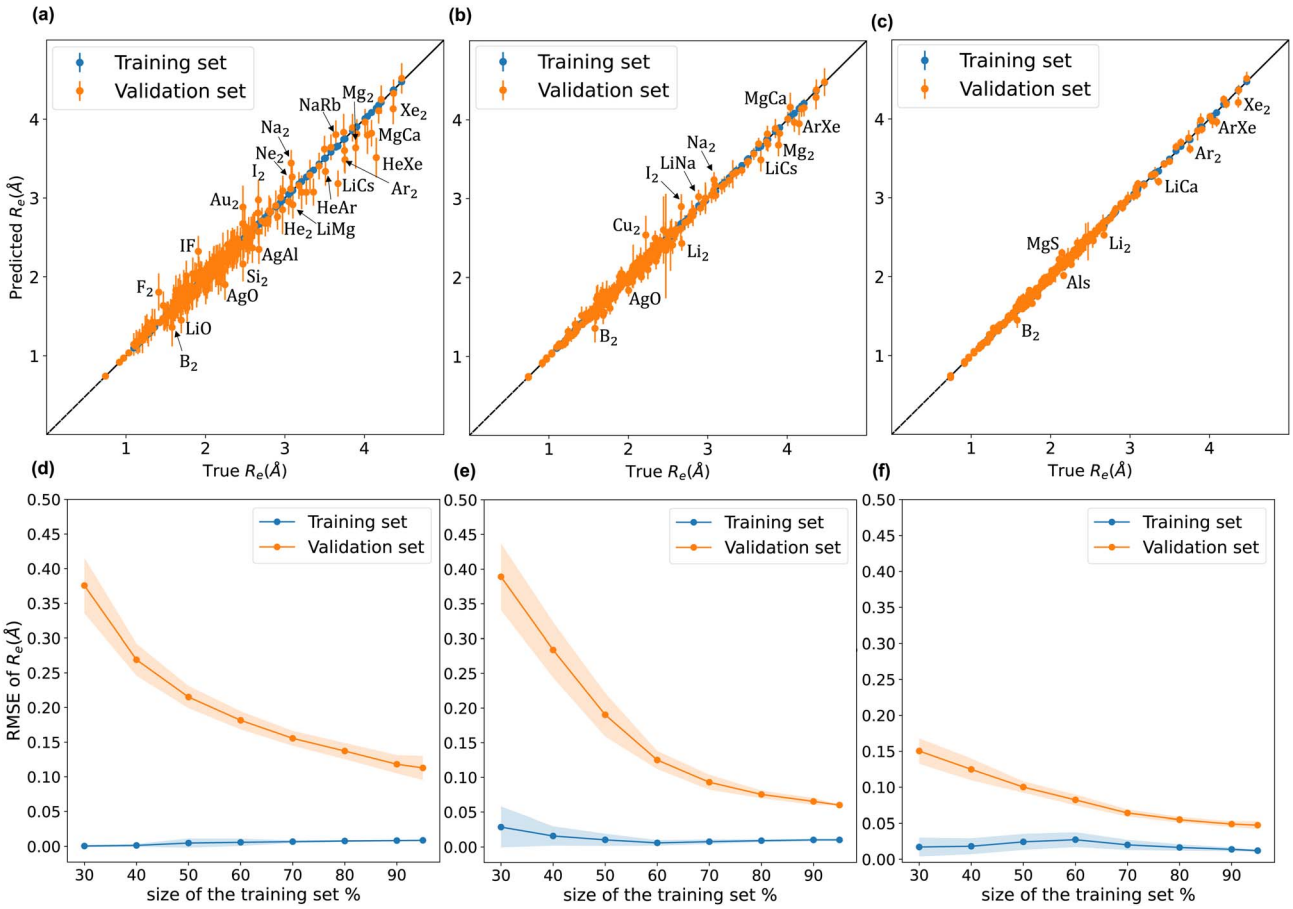


Fig. 3 Upper row shows scatter plots of experimental values of R_e on the x-axis and predicted R_e on the y-axis *via* models (a) r1 (b) r2 (c) r3, points, and error bars represent the predictive distribution means and standard deviations respectively after averaging over 1000 MC-CV steps. The lower row shows three learning curves of models (d) r1, (e) r2, and (f) r3. Points and shade around represent the $\overline{\mathrm{RMSE}}$ and $\pm 0.5\mathrm{STD}(\overline{\mathrm{RMSE}})$ over 500 MC-CV splits.

r1 using 90% of the data set. Overall, models r3 and r4 show an improvement in the prediction on $R_e \sim 40\%$ as shown in Table 2, leading to validation $\overline{MAE}$ of 0.027 Å and 0.026 Å and a validation $\overline{RMSE}$ of 0.039 Å and 0.038 Å, respectively. In other words, models r3 and r4 are almost two times more precise in predicting R_e than previously ML-based or empirically-derived predictions, and almost as accurate as the state-of-the-art *ab initio* calculations for diatomics.[132,133] Furthermore, the lower panes of Fig. 3 show converging learning curves characterized by relatively narrow gaps between validation and training curves. The decaying trend of the validation curves suggests that convergence toward lower levels of errors is possible upon further training on more extensive datasets. The training MAE of r2 is $\sim 6 \times 10^{-3}$ Å; this means that we might have the capacity to achieve an accuracy ~0.010 Å using only atomic properties. In other words, with more data our ML models could be as accurate as state-of-the-art *ab initio* quantum chemistry methods.

To highlight a few of the common outliers of the four models, we consider Li_2, B_2, LiCs, and LiCa. r1, r2, r3, r4 underestimate R_e for Li_2 by 6–10%. r1, r2, r3, r4 underestimate R_e for B_2 by 14%. 15%, 7%, and 8%, respectively, which could be connected to B_2 being the only homonuclear molecule from group 13 in the data set. For LiCs, $R_e = 3.67$ Å (ref. 87) and r2 predicts $R_e = 3.49 \pm 0.15$ Å; that is, the experimental value is 1.2 standard deviation away from the mean predictive posterior distribution of model r2 for LiCs, although most of the theoretical R_e values of LiCs are within one standard deviation.[86] For LiCa, the experimental value found by Krois *et al.* is $R_e = 3.34$ Å.[84] On the contrary, the r4 model predicts $R_e = 3.20 \pm 0.05$ Å, almost three standard deviations away from the experimental value. However, model r2 predicts $R_e = 3.33 \pm 0.09$ Å, with only 0.3% relative error. In addition, high-level *ab initio* calculations results are within one standard deviation from the mean predictive posterior distribution of model r2 for LiCa, namely CASPT2 predicts $R_e = 3.40$ Å,[134] QCISD(T) gives $R_e = 3.41$ Å,[135] MRCI leads to $R_e = 3.40$ Å,[135] and CIPI prediction is $R_e = 3.40$ Å.[136]

4.2 ω_e

We have implemented and developed four models to predict ω_e as listed in Table 1. Model w1 is the best-performing model of Liu *et al.*[26] It is characterized by six features, including atomic and molecular properties. Namely, the groups and periods of the constituent atoms, the average group, $\bar{g}=(g_1^{iso} + g_2^{iso})/2$, and R_e^{-1}. g^{iso} encodes isotopic information, such that $g_i^{iso} = 0$ for deuterium, $g_i^{iso} = -1$ for tritium, and $g_i^{iso} = g_i$ for every other

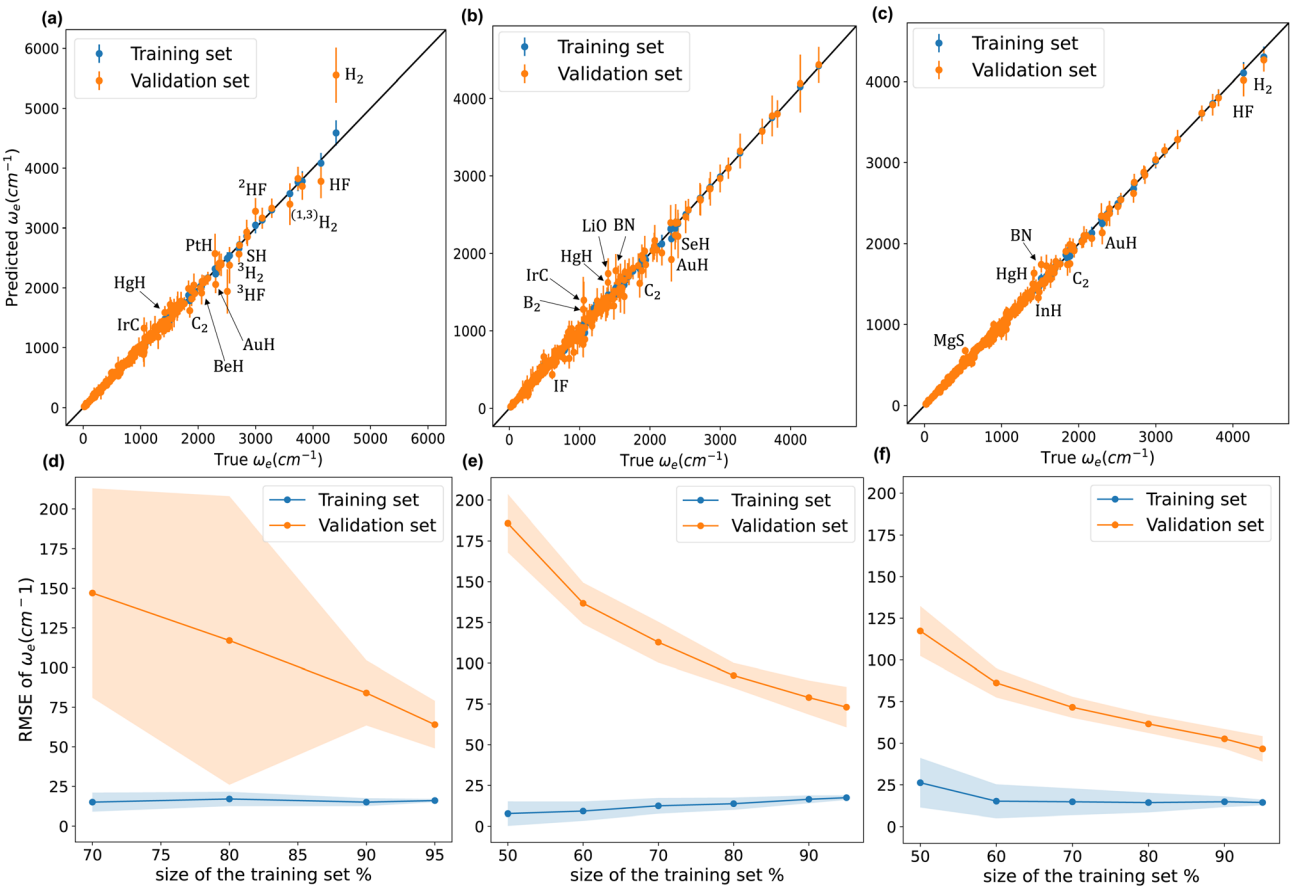


Fig. 4 Upper row show scatter plots of experimental values of ω_e on the *x*-axis and predicted ω_e on the *y*-axis *via* models (a) w1 (b) w2 (c) w4, points, and error bars represent the predictive distribution means and standard deviations respectively after averaging over 1000 MC-CV steps. The lower row shows three learning curves of models (d) w1, (e) w2, and (f) w4. Points and shade around represent the $\overline{RMSE}$ and $\pm 0.5 STD(\overline{RMSE})$ over 500 MC-CV splits.

element. The prior mean function is set to zero. On the other hand, model w2 only includes groups and periods of the constituent atoms and the reduced mass of the molecule. The prior mean of model w2 is given by

$$m_{w2} = \beta_0^{w2} + \beta_1^{w2}(p_1 + p_2) + \beta_2^{w2}(g_1 + g_2) + \beta_3^{w2}\ln(\mu^{1/2}), \quad (17)$$

where β_k^{w2-w3}, $k \in \{0, 1, 2, 3\}$ are the linear coefficients of m_{w2}. Model w3 uses the same features as model w2 but includes R_e in the prior mean function. Model w4 is characterized by six features as model w1 and uses R_e as a feature in both the kernel and in the prior mean function.

Motivated by the relationship between ω_e and R_e, both w3 and w4 use the same prior mean function

$$m_{w3-w4} = \beta_0^{w3-w4} + \beta_1^{w3-w4}(p_1 + p_2) + \beta_2^{w3-w4}(g_1 + g_2) + \beta_3^{w3-w4}R_e + \beta_4^{w3-w4}\ln(\mu^{1/2}), \quad (18)$$

where β_k^{w4}, $k \in \{0, 1, 2, 3, 4\}$ are the linear coefficients of m_{w3-w4}.The inclusion of the reduced mass in models w2, w3, and w4 eliminates the necessity of imposing isotopic information on the groups of constituent atoms.

Fig. 4 compares w1, w2, and w4 (plots of w3 are similar to those of w4). We notice from panel (a) that model w1 struggles against hydrides, and hydrogen and hydrogen fluoride isotopologues. Indeed, the model significantly overestimates ω_e for H_2. On the other hand, panel (b) shows that w2 performs much better against hydrides, and hydrogen and hydrogen fluoride isotopologues. w2 predictions for H_2 and HF are accurate and even better than those of models w3 and w4, as shown in panel (c). Panels (a) and (b) clearly show that model w2 outperforms model w1 when considering molecules with larger values of ω_e. Looking at the learning curves in panels (d) and (e), we see that model w2 is far more consistent than model w3, as indicated by the shade around the validation curves of both models. From Table 2, the validation SEM(RMSE) of models w2 and w1 show that model w2 is 40% more consistent in its performance than model w1 when both models are validated using the same 1000 MC-CV splits. Furthermore, the test RMSE of w2 is 20% lower than that of w1. Model w2 has lower dimensionality than model w1 and only implements atomic properties; nevertheless, it performs similarly to model w1.

From Table 2, we see that although model w3 has a test MAE almost equal to model w1, models w3 and w4 have validation $\overline{\text{MAE}}$s 15–21% lower than that of w1, indicating an overall better average performance of the newly developed models. Furthermore, w3 and w4 have validation $\overline{\text{RMSE}}$s and test RMSEs 28–36% lower than w1, showing the robustness of the two new models. Panel (c) of Fig. 4 shows minimal scatter around the true line. Few hydrides, along with BN and C_2, still challenge the model; however, their absolute errors are significantly suppressed compared to w1 and w2. The validation curve of model w4 in panel (f) shows a much higher learning rate than w1 and w2, with a much shallower gap between the validation and learning curves. Moreover, the shadow around the validation curve is minimal at all training sizes. From Table 2, we see that w3 and w4 are far more consistent than w1, with STD(RMSE) 60–70% lower than that of w1.

On the other hand, the lower three panels in Fig. 4 show that the validation and training curves can converge towards lower error values. Hence, all the models might benefit from training on a more extensive dataset. The training MAEs of w1, w2, w3, and w4 range between 8 to 7 cm^{-1}, so it might be possible to reach near spectroscopic accuracy ($\sim$10 cm^{-1}) by training these models on larger datasets. In the case of w2, if the validation curve's decaying trend persists upon further training, near spectroscopic accuracy might be achieved only through knowledge of atomic positions in the periodic table. Similarly, these models trained in larger database could outperform the state-of-the-art *ab initio* quantum chemistry methods.[132,133]

We highlight some of the outliers that are common to some of the models. All the models overestimate ω_e for HgH by at least 12%, while for IrC, w1 and w2 overestimate ω_e by 30% and 25%, while w3 and w4 only overestimate it by only 4% and 7%, respectively. The observed overestimation might be because HgH and IrC are the only molecules that consist of mercury or iridium in the dataset.

We have found two values of ω_e for AuF in the literature; Saenger *et al.* reported ω_e = 560 cm^{-1} in 1992 (ref. 63), while Andreev *et al.* reported ω_e = 448 cm^{-1} in 2000.[59] All our models predict values closer to 560 cm^{-1}: w2 predicts ω_e = 529 $\pm$ 87 cm^{-1}, while w3 and w4 are almost in exact agreement with Saenger's value with ω_e = 568 $\pm$ 54 cm^{-1} and ω_e = 565 $\pm$ 45 cm^{-1}, respectively.† Our predictions are in agreement with relativistic density functional and *ab initio* methods. Namely, first-order relativistic density functional calculation predicts ω_e = 491 cm^{-1} while Zeroth-order regular relativistic approximation within the Kohn–Sham density functional scheme ZORA(MP) predicts ω_e = 526 cm^{-1}.[61] In the same line, the relativistic MP2 approach predicts ω_e = 590 cm^{-1},[138] while relativistic MRCI predicts ω_e = 525 cm^{-1}.[139] A similar situation occurs in the case of ZnBr, as shown in Table 3. For 30 years, there was a discrepancy in the value of ω_e of ZnBr. Gosavi *et al.* reported ω_e $\approx$ 319 cm^{-1} in 1971.[140] Next, Givan *et al.* reported ω_e $\approx$ 198 cm^{-1} in 1982.[141] On the contrary, the MRCI calculations by Elmoussaoui and Korek predicted ω_e $\approx$ 267 cm^{-1} in 2015.[142] Finally, Burton *et al.* experimentally reported ω_e = 284 cm^{-1} in 2019.[118] Here, w2 predicts ω_e = 271.2 $\pm$ 21.7 cm^{-1}, w3 predicts ω_e = 289.5 $\pm$ 15.4 cm^{-1} and w4 predicts ω_e = 281.0 $\pm$ 12.0 cm^{-1}. Therefore, our predicted values are in great agreement with the most recent theoretical and experimental values.

4.3 D_0^0

Finally, we have developed model d1 to predict the dissociation energy, D_0^0, *via* $\ln(D_0^0)$ using (p_1, g_1, p_2, g_2, $\mu^{1/2}$) as features in

† The w2, w3, and w4 predictions for AuF in the main text were predicted, including HgCl, HgI, and HgBr in the training set. To test the robustness of the models, we removed those three molecules from the training set since their R_e values might be related to $HgCl_2$, HgI_2, and $HgBr_2$.[56,137] Indeed, those molecules could affect the model predictions because they are closely related to AuF since Au (group 11) and Hg (group 12) are members of the sixth period, and F, Cl, I, and Br are all halogens. However, in this case, w2, w3 and w4 predict $\omega_e \sim$ 530 cm^{-1}, $\omega_e \sim$ 600 cm^{-1}, and $\omega_e \sim$ 590 cm^{-1}, respectively, in good agreement with the predicted results in the main text, experimental results and *ab initio* methods.

Table 3 Predictions and experimental values of R_e and ω_e for 24 molecules in the testing set. References of experimental values are included. Ref. column includes references for experimental values

Molecule	Models for R_e, ω_e	Predicted R_e (Å)	Experimental R_e (Å)	Predicted ω_e (cm^{-1})	Experimental ω_e (cm^{-1})	Ref.
HCl	r4, w4	1.267 ± 0.029	1.274	2939 ± 114	2990	56
	r2, w2	1.275 ± 0.046		3020 ± 209		
^{2}HCl	r4, w4	1.286 ± 0.027	1.274	2172 ± 80.0	2145	56
	r2, w2	1.285 ± 0.0425		2123 ± 136		
RuC	r4, w4	1.614 ± 0.039	1.600	1106 ± 59.4	1100	96
	r2, w2	1.644 ± 0.074		1066 ± 119		
WO	r4, w4	1.667 ± 0.046	1.657	1049 ± 65.5	1067	143 and 144
	r2, w2	1.708 ± 0.088		994.9 ± 131		
MoC	r4, w4	1.652 ± 0.037	1.676	982.5 ± 49.2	1008	96 and 97
	r2, w2	1.714 ± 0.057		1011 ± 106		
WC	r4, w4	1.746 ± 0.0547	1.714	1065 ± 78.3	983.0	145
	r2, w2	1.645 ± 0.099		1097 ± 178		
NbC	r4, w4	1.739 ± 0.041	1.700	1019 ± 58.3	980.0	102
	r2, w2	1.664 ± 0.057		967.7 ± 115		
NiC	r4, w4	1.621 ± 0.048	1.627	857 ± 55.8	875.0	104
	r2, w2	1.668 ± 0.093		825.3 ± 114		
PdC	r4, w4	1.736 ± 0.032	1.712	872.0 ± 37.9	847.0	108
	r2, w2	1.720 ± 0.057		866.6 ± 74.0		
UO	r4, w4	1.863 ± 0.022	1.838	888.1 ± 27.2	846.0	121
	r2, w2	1.839 ± 0.033		893.7 ± 45.3		
NiO	r4, w4	1.585 ± 0.038	1.627	785.2 ± 40.2	839.0	105
	r2, w2	1.667 ± 0.055		796.9 ± 82.9		
YC	r4, w4	1.907 ± 0.076	2.050	649.2 ± 70.8	686.0 ± 20	122 and 123
	r2, w2	1.824 ± 0.094		834 ± 185		
ZnF	r4, w4	1.756 ± 0.029	1.768	603 ± 24.2	631.0	146
	r2, w2	1.801 ± 0.053		580.4 ± 45.8		
NiS	r4, w4	1.940 ± 0.044	1.962	482 ± 28.6	512.0	106
	r2, w2	1.999 ± 0.081		479.1 ± 58.6		
ZnCl	r4, w4	2.136 ± 0.028	2.130	384.8 ± 15.4	390.0	147
	r2, w2	2.164 ± 0.053		371.0 ± 29.0		
ZnBr	r4, w4	2.299 ± 0.029	2.268	284.9 ± 11.7	284.0	118
		—	—	—	319.0	140
		—	—	—	198.0	141
	r2, w2	2.321 ± 0.0542		271.1 ± 21.7		
ZnI	r4, w4	2.499 ± 0.030	2.460	235.5 ± 10.1	223.0	56
	r2, w2	2.484 ± 0.057		228.0 ± 19.2		
SnI	r4, w4	2.722 ± 0.035	2.732	193.3 ± 9.48	197.0	107
	r2, w2	2.725 ± 0.068		198.0 ± 19.9		
PbI	r4, w4	2.784 ± 0.030	2.798	156.8 ± 6.54	160.0	107
	r2, w2	2.814 ± 0.056		154.1 ± 13.0		
CoO	r2	1.543 ± 0.056	1.628	—	—	72–74
CrC	r2	1.517 ± 0.099	1.630	—	—	77
IrSi	r2	2.084 ± 0.171	2.09	—	—	78
UF	r2	2.002 ± 0.081	2.02	—	—	119
ZrC	r2	1.846 ± 0.058	1.740	—	—	124

a Matérn 3/2 kernel, and a prior mean function that employs both ω_e and R_e

$$m_{d1} = \beta_0{}^{d1} + \beta_1{}^{d1}(p_1 + p_2) + \beta_2{}^{d1}(g_1 + g_2) + \beta_3{}^{d1}R_e + \beta_4{}^{d1}\ln(\mu^{1/2}) + \beta_5{}^{d1}\ln(\omega_e), \quad (19)$$

where $\beta_k{}^{d1}$, $k \in \{0, 1, 2, 3, 4, 5\}$ are the linear coefficients of m_{d1}. The performance of the model is displayed in Fig. 5, where the scatter plot [left panel] shows some dispersion of the model predictions concerning the true values. From Table 2, the validation and test errors suggest that the model is consistent and generalizable to new data indicating that model d1 yields reasonable performance as far as $\ln(D_0^0)$ is concerned. However, converting $\ln(D_0^0)$ back to D_0^0, the errors are ~0.4 eV ≡ 10 kcal mol^{-1}, as shown in Table 4, which is a significant error considering the typical chemical accuracy (±1 kcal mol^{-1}). However, as shown in the right panel of Fig. 5, the model might benefit from training on more data, leading to a potential improvement of a factor of 3. On the other hand, it is possible to accurately predict bond energies, in complex molecules, by using intuitive chemical descriptors, as shown in ref. 148 and 149, which is something that we are planning on implementing in the future.

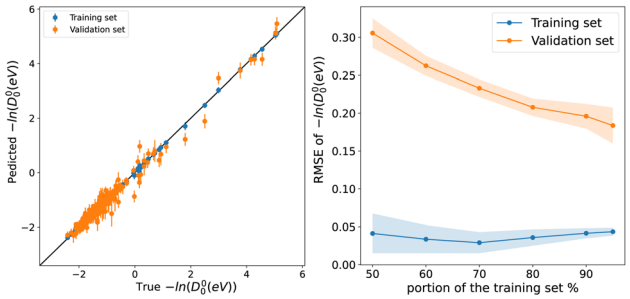


Fig. 5 Left panel, scatter plot of experimental values of $-\ln(D_0^0)$ on the x-axis and predicted $-\ln(D_0^0)$ on the y-axis *via* models d1, points, and error bars represent the predictive distribution means and standard deviations respectively after averaging over 1000 MC-CV steps. Right panel shows the learning curves of model d1. Points and shade around represent the $\overline{\text{RMSE}}$ and $\pm 0.5\text{STD}(\overline{\text{RMSE}})$ over 500 MC-CV splits.

Table 4 Predictions *via* model d1 and experimental values of D_0^0 for seven molecules in the testing set. References of experimental values are included. Ref. column includes references for experimental values

Molecule	True D_0^0 (eV)	Predicted D_0^0 (eV)	Ref.
RuC	6.34	6.2 ± 1.45	96
MoC	5.01	5.93 ± 1.41	96 and 97
NbC	5.85	5.50 ± 1.3	102 and 96
YC	4.29	4.40 ± 1.46	122 and 123
ZnBr	2.45	3.86 ± 0.71	118
SnI	2.89	3.28 ± 0.77	107

During the development of this work, we have realized that, historically, uncertainties about the dissociation energy experimental values had restrained the development of empirical relations connecting them to other atomic and molecular properties and have led several authors to focus their efforts on the $\omega_e - R_e$ relation.[9,41,47] More recently, Fu *et al.* used an ML model to predict dissociation energies for diatomic molecules, exploiting microscopic and macroscopic properties.[150] They tested their model against CO and highlighted that the reported experimental dissociation energy in the literature had increased by 100 kcal mol^{-1} over the course of 78 years from 1936 to 2014 (ref. 150–152) (in Table 1 of ref. 150). The data used to train model d1 is primarily collected from Huber and Herzberg's constants of diatomic molecules, first published in 1979.[56] Unlike experimental values of R_e and ω_e, since 1980, a significant number of D_0^0 values have been updated.[48] To name a few, MgD, MgBr, MgO, CaCl. CaO, SrI, SrO, TiS, NbO, AgF, AgBr, and BrF all have their experimental values updated with at least ±2.3 kcal mol^{-1} difference from their values in Huber and Herzberg.[57] Moreover, for some molecules, the uncertainties in D_0^0 experimental values are not within chemical accuracy. For instance, MgH, CaCl, CaO, CaS, SrH, BaO, BaS, ScF, Tif, NbO, and BrF have uncertainties ranging from ±1 kcal mol^{-1} up to ±8 kcal mol^{-1}.[48]

Based on the previous discussion, we can connect the unsatisfactory performance of model d1-in comparison to our developed R_e and ω_e models-to noise modeling. Unlike R_e and ω_e, it is most likely that uncertainties around D_0^0 experimental values drive from various systematic effects. Therefore, modeling the errors in D_0^0 experimental values to be identically distributed as in eqn (6) might not be a proper treatment. Thus, to develop better models for predicting D_0^0, more sophisticated techniques of error modeling might be required. To this endeavor, gathering more reliable data with experimental uncertainty within ±1 kcal mol^{-1} might be sufficient. Something that we are working on it, and it will be published elsewhere.

Table 5 R_e and ω_e ML predictions for molecules not contemplated in the database. The *ab initio* results are taken from ref. 133

Molecule	*Ab inito* R_e (Å)	r2 predicted R_e (Å)	*Ab inito* ω_e (cm^{-1})	w2 predicted ω_e (cm^{-1})
LiFr	3.691	3.709 ± 0.123	180.7	198.9 ± 35.9
KFr	4.284	4.483 ± 0.173	65.2	64.0 ± 16.2
RbFr	4.429	4.389 ± 0.145	46.0	48.8 ± 10.4
CsFr	4.646	4.403 ± 0.221	37.7	42.7 ± 13.7

4.4 Testing ML models *versus ab initio* results

To further assess the accuracy of our ML models regarding R_e and ω_e we have exposed our models to molecules containing Fr. Indeed, our dataset does not contain any Fr-containing molecule, defining the most complicated scenario for our ML models. The results in comparison with the state-of-the-art *ab initio* methods are shown in Table 5, where it is noticed that our ML predictions agree well with *ab initio* predictions. Furthermore, more data can quickly improve ML predictions, as presented in Fig. 3 and 4. Therefore, ML predictions can be competitive with *ab initio* quantum chemistry methods using a larger dataset.

4.5 Predicting homonuclear spectroscopic properties from heteronuclear data

To explore the capability of our models in predicting the spectroscopic properties of homonuclear molecules from spectroscopic and atomic information of heteronuclear molecules, we train our models for predicting R_e, ω_e, and D_0^0 (r3, w4, and d1) using a special split. We fit the three models to heteronuclear data in D_{tv} and then make predictions for the left-out homonuclear molecules. The performance of our models is displayed in Fig. 6, where we notice an outstanding performance. Only a few outliers are observed, showing a minimal deviation from the true line. In particular, we obtain MAEs of 0.08 Å, 74 cm^{-1}, and 0.149 for models r4, w4, and d1, respectively. Hence, it is possible to predict the accurate spectroscopic properties of homonuclear molecules from heteronuclear data. Furthermore, our results indicate that expanding the data set by including homonuclear molecules yields high-performing models able to predict spectroscopic properties for both heteronuclear and homonuclear molecules.

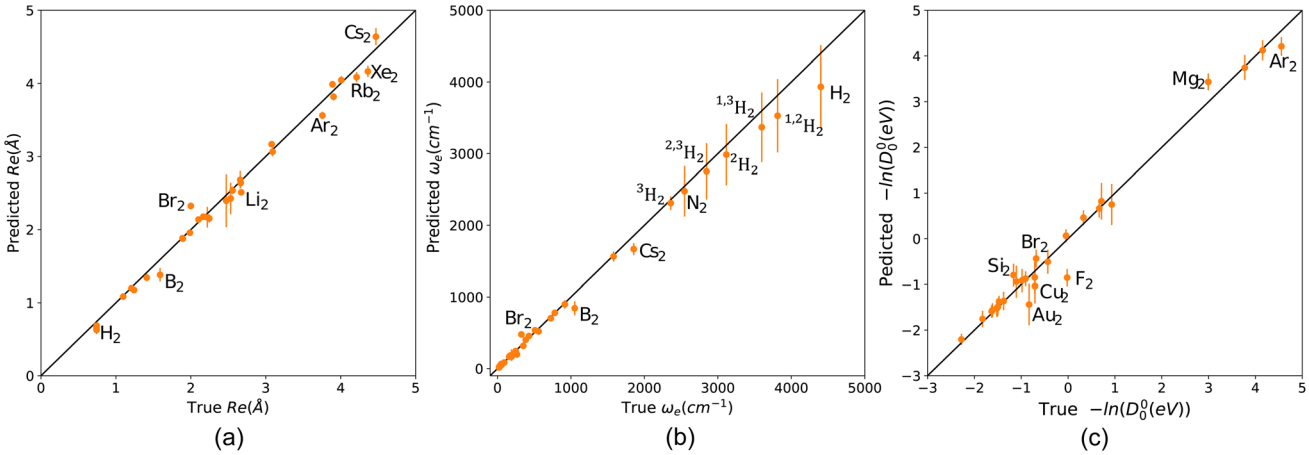


Fig. 6 Scatter plots of models predicting R_e, ω_e and $-\ln(D_0^0)$ for homonuclear molecules from heteronuclear molecules data (a) r3 (b) w4 (c) d1. Points and error bars represent the predictive distribution means and standard deviations, respectively.

4.6 Towards a classification of diatomic molecules

For models r2, r3, w2, w3, and d1, we have achieved good results using a kernel common to all five models. That Matérn kernel given by eqn (10) with $\nu = 3/2$, is a similarity measure. Therefore, it is possible to quantify the similarity between a pair of molecules denoted by $i = p$, q giving their feature vector $\mathbf{x}_i = (\mathrm{p}_1{}^i, \mathrm{g}_1{}^i, \mathrm{p}_2{}^i, \mathrm{g}_2{}^i, \sqrt{\mu^i})$. The models are fitted to the whole dataset to determine the parameters (σ_n, l, σ_f). The kernel given by eqn (10) with $\nu = 3/2$ and the determined parameters can be used to form a similarity matrix. Each element in the similarity matrix quantifies the similarity between a pair of molecules in the dataset. Off-diagonal elements are calculated *via* eqn (10) for $p \neq q$, with the diagonal representing the similarity of the molecules with themselves ($p = q$). A heat map representation of the similarity matrix is given in Fig. 7, while the degree of similarity from 0 to 1 is given over a greyscale as indicated by the color bar on the right side of the figure.

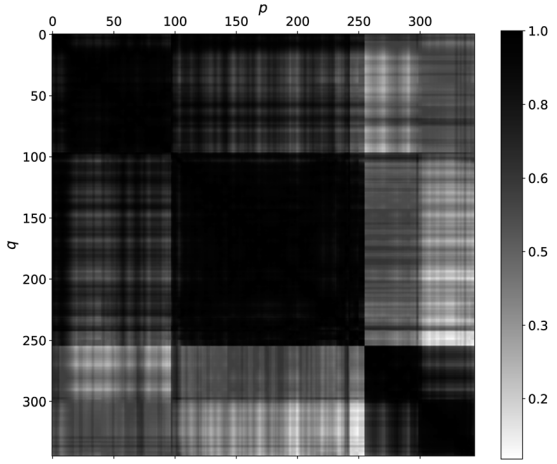


Fig. 7 A heat map quantifies the degree of similarity among molecules in the data set from 0 (white, not similar) to 1 (black, identical) on a grayscale. The heat map was generated by finding the matrix element of a similarity matrix. Each matrix element quantifies the similarity between a pair of molecules p (on the x-axis) and q (on the y-axis) *via* eqn (10) with $\nu = 3/2$ and parameters determined as described in the text.

To further explore the quantified similarity among molecules, we consider three subsets of molecules and show their heatmaps in the upper panels of Fig. 8. The lower panels of Fig. 8 show the corresponding network representation of the similarity among these subsets of molecules. Black squares in the heat map plots of Fig. 8 indicate that a pair of molecules is highly similar, whereas white squares indicate 0% similarity. The network representation represents each molecule as a node. The similarity between two molecules is diagrammatically shown with a line joining their corresponding nodes. The networks show similarities above the 80% level. A line joins two nodes only if they are at least 80% similar. The length of a joining line indicates the degree of similarity between a pair above the 80% level. A short line indicates a high degree of similarity, and a long line indicates a lower degree of similarity.

From panel (a) of Fig. 8, we see noble gas dimers clustering around Xe_2, and alkali metals-alkaline earth metals cluster around NaRb. Both clusters are isolated from each other and VF, indicating a lower degree of similarities between these clusters and VF. A similar scenario is observed in panel (b), where alkaline earth metal hydrides cluster upon themselves with tight interconnections indicating high similarity. On the other hand, ZnH is remotely connected to the cluster, indicating a lower degree of similarity. The upper right cluster shows an interconnection among diatomic reactive nonmetals, including halides and oxygen; notably, AgAl is connected to these molecules. Panel (c) displays a more involved clustering scheme involving transition metal hydrides (MnH and AgH), connected to a metalloid hydride (TlH and InH) and with a lower degree to alkaline earth metals hydrides (LiH and BeH). The right-hand side cluster consists of various transition metal diatomics, dihalides, and others, all closely related except for MgS. Note that all the molecules in the right-hand side cluster consist of atoms from the periodic table's right side. At the same time, MgS combines one atom from group 2 and one from group 16. Notably, homonuclear diatomic and heteronuclear molecules

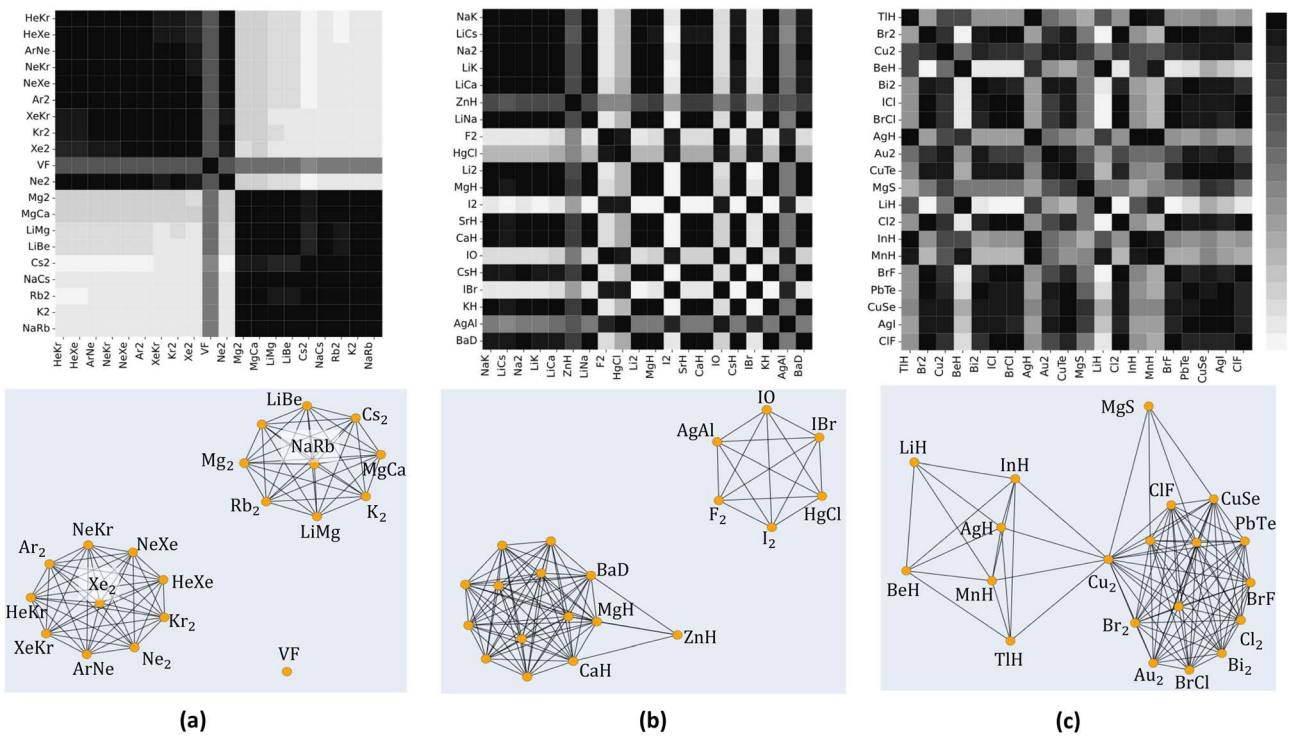


Fig. 8 Heat maps representing similarities among subsets of molecules (upper row) and their corresponding network representation (lower row). The color bar (top right) quantifies the similarity between a pair of molecules from 0 (white, not similar) to 1 (black, identical) on a greyscale. The network representations show similarities above the 80% (0.8) level. Each node represents a molecule. Short lines joining two nodes represent a high degree of similarity, while longer lines represent a lower degree of similarity above 80%. No line at all indicates a lower degree of similarity below 80%.

are firmly embedded within all the clusters, emphasizing the importance of including homonuclear data in our models.

Since only atomic properties are required to find elements of the matrix representation of the kernel, the similarity matrix can guide us in our data-gathering efforts. For example, we can determine which molecules can fill the gaps and connect clusters to build more robust models. More interestingly, we can systematically classify molecules based on the similarity matrix. Such classification would help develop potential energy surfaces (PES) for diatomic molecules. As pointed out by Newing, similar molecules will have similar potential energy surfaces, at least around R_e.[28]

5 Summary and conclusion

In this work, first, we have extended the previous database of Liu *et al.*,[26] gathering ground state spectroscopic data of 85 homonuclear and heteronuclear molecules leading to a data set of 338 molecules. Next, the database has been used to train 9 ML models to predict the main spectroscopic constants: R_e, ω_e, and D_0^0. These models can be categorized into three categories:

• Models in category (i) only employ information from the periodic table and thus can predict spectroscopic properties of any combination of two elements. These models can be used to systematically classify molecules made up of any two elements in the periodic table (Section 4.6). While spectroscopic data availability does not limit these models' ability to predict spectroscopic constants of any molecule, it affects the models' accuracy. These models are characterized by a relatively larger gap between validation and learning curves than models in categories (ii) and (iii). Thus, we would expect a better performance of category (i) models upon training on larger datasets.

• Models in category (ii) use spectroscopic information only in their mean function but not in the kernel, and are robust against noise in input variables. In this case, since the mean function is a linear function, we can apply standard errors-in-variables methods.[153] This might be advantageous if we would like to use uncertain experimental data or predictions from (i) models or *ab initio* methods to train our models.

• Models in category (iii) include our most flexible, accurate, and consistent models (r4, w4). These models benefit from a high learning rate and a narrow gap between validation and learning curves. Apart from their outstanding performance, we can train these models using ground and excited states simultaneously since each state will be labeled by its spectroscopic constant values R_e or ω_e along with other properties that define the molecule $\{p_1, g_1, p_2, g_2, \mu^{1/2}\}$.

In summary, the newly developed models in this work showed an outstanding performance in all metrics in comparison to the previous ML models and other empirical and semiempirical models, with mean absolute errors ranging between 0.02 Å and 0.04 Å for R_e, and 26 cm^{-1} to 40 cm^{-1} for ω_e. We have been able to predict homonuclear spectroscopic properties with good accuracy upon training our models on

heteronuclear molecules' data. Indeed, our models are almost as accurate as the state-of-the-art *ab inito* methods for diatomics.[132,133] In addition, our models only require data, whereas *ab initio* quantum chemistry methods require specific knowledge by the user.

On the other hand, since we use the same kernel for all models under consideration, we are uniquely positioned to study a way to classify diatomic molecules beyond the traditional one based on the nature of the chemical bond. We expect such classification to enhance the performance and facilitate the development of ML models predicting spectroscopic and molecular properties of diatomic molecules. Further, the classification of diatomic molecules should help develop potential energy surfaces (PES).

Finally, we have shown that for molecules with large ionic character and containing heavy atoms (*e.g.*, LiCs, LiCa, AuF, and ZnBr), our predictions are comparable to DFT and even the state-of-the-art *ab initio* methods. Moreover, two of our models (r2 and w2) offer a promising opportunity to predict spectroscopic properties from atomic positions in the periodic table with high accuracy. This is a stepping stone towards closing the gap between atomic and molecular information; more spectroscopy data is required to do so. More extensive, open, and user-friendly data will help the field of data-driven science to understand the chemical bonding and spectroscopy of small molecules. Indeed, that is something that we are currently working on in our group: we need more spectroscopic data in the big data era. Finally, it is worth mentioning that we are approaching a period in which machine learning techniques are as accurate as *ab initio* quantum chemistry methods for calculating spectroscopic constants of diatomics with almost no computational effort.

Data availability

The machine learning codes and the data employed in this work can be found on GitHub [**https://github.com/Mahmoud-Ibrahim-Mamrstein/Spectroscopic-constants-from-atomic-properties**]. In this repository, the user can download the folder called gpr, which contains all the codes and data employed in this paper. The data folder contains all the spectroscopic constants, including references, used as training and test sets in this work. In the same folder, the atomic properties can be found in periodictable.csv. On the contrary, the codes and performance analysis of the machine learning models can be found in each of the subfolders labeled with the model's name. For instance, folder r1 contains all the information relevant to model r1 of the paper. The folder entitled linear regression contains 5 subfolders accounting for each linear regression model employed as a baseline method. Folders labeled by svmr1, svmr2, and svmr3 correspond to support vector machine results for models r1, r2, and r3. The same holds for svmw1 and svmw2, regarding models w1 and w2, whereas the folder svmd1 is the support vector machine prediction for the d1 model. Finally, our study on the classification of molecules is found under the folder heat_maps_and_networks.

Author contributions

X. L. helped with the database and the first ML models. M. A. E. I. gathered the data and performed the new ML models, whereas J. P.-R. envisioned the idea and supervised the project. M. A. E. I. and J. P.-R. wrote the paper.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

J. P.-R. acknowledges the funding of the Simons Foundation and the Lorentz Center of the University of Leiden for organizing the workshop "New directions in cold and ultracold molecules", in which some part of this work was discussed. X. L. acknowledges the support or the the Deutsche Forschungsgemeinschaft (DFG – German Research Foundation) under the grant number PE 3477/2 – 493725479. M. A. E. I. acknowledges the funding and support of THE BINATIONAL FULBRIGHT COMMISSION and Assiut University in EGYPT. Finally, we would like to thank Rian Koots for useful suggestions to improve the GitHub repository.

References

1 A. Kratzer, *Z. Phys.*, 1920, **3**, 289–307.
2 R. Birge, *Phys. Rev.*, 1925, **25**, 240–254.
3 R. Mecke, *J. Phys.*, 1925, **32**, 823–834.
4 P. M. Morse, *Phys. Rev.*, 1929, **34**, 57.
5 C. Clark, *J. Sci.*, 1934, **18**, 459–470.
6 C. D. Clark, *Nature*, 1934, **133**, 873.
7 C. D. Clark, *Trans. Faraday Soc.*, 1941, **37**, 299–302.
8 C. D. Clark and K. Webb, *Trans. Faraday Soc.*, 1941, **37**, 293–298.
9 R. M. Badger, *J. Chem. Phys.*, 1934, **2**, 128–131.
10 E. Kraka, J. A. Larsson and D. Cremer, *Computational Spectroscopy*, ed. J. Grunenberg, Wiley, New York, NY, USA, 2010, pp. 105–149.
11 E. Kurita, H. Matsuura and K. Ohno, *Spectrochim. Acta, Part A*, 2004, **60**, 3013–3023.
12 R. M. Badger, *J. Chem. Phys.*, 1935, **3**, 710–714.
13 M. Kaupp, D. Danovich and S. Shaik, *Coord. Chem. Rev.*, 2017, **344**, 355–362.
14 J. Cioslowski, G. Liu and R. A. M. Castro, *Chem. Phys. Lett.*, 2000, **331**, 497–501.
15 W. G. Penney and G. B. B. M. Sutherland, *Proc. R. Soc. London, Ser. A*, 1936, **156**, 654–678.
16 H. Allen and A. Longair, *Nature*, 1935, **135**, 764.
17 M. L. Huggins, *J. Chem. Phys.*, 1935, **3**, 473–479.
18 M. L. Huggins, *J. Chem. Phys.*, 1936, **4**, 308–312.
19 G. B. Sutherland, *Proc. Indian Acad. Sci.*, 1938, 341–344.
20 J. Linnett, *Trans. Faraday Soc.*, 1940, **36**, 1123–1134.
21 C. Wu and C.-T. Yang, *J. Phys. Chem.*, 1944, **48**, 295–303.
22 C. Wu and S. Chao, *Phys. Rev.*, 1947, **71**, 118.
23 K. Guggenheimer, *Proc. Phys. Soc.*, 1946, **58**, 456.

24 J. Linnett, *Trans. Faraday Soc.*, 1945, **41**, 223–232.
25 W. Gordy, *J. Chem. Phys.*, 1946, **14**, 305–320.
26 X. Liu, G. Meijer and J. Pérez-Ríos, *RSC Adv.*, 2021, **11**, 14552–14561.
27 J. C. Slater, *J. Chem. Phys.*, 1933, **1**, 687–691.
28 R. Newing, *London, Edinburgh Dublin Phil. Mag. J. Sci.*, 1940, **29**, 298–301.
29 P. Güttinger, *Z. Phys.*, 1932, **73**, 169–184.
30 W. Pauli, *Handb. Phys.*, 1933, **24**, 43.
31 H. Hellmann, *Einführung in die Quantenchemie*, 1937.
32 R. P. Feynman, *Phys. Rev.*, 1939, **56**, 340.
33 L. Salem, *J. Chem. Phys.*, 1963, **38**, 1227–1236.
34 P. Empedocles, *J. Chem. Phys.*, 1967, **46**, 4474–4481.
35 P. Empedocles, *Theor. Chim. Acta*, 1968, **10**, 331–336.
36 A. B. Anderson, N. C. Handy and R. G. Parr, *J. Chem. Phys.*, 1969, **50**, 3634–3635.
37 A. B. Anderson and R. G. Parr, *J. Chem. Phys.*, 1970, **53**, 3375–3376.
38 A. Anderson and R. Parr, *Chem. Phys. Lett.*, 1971, **10**, 293–296.
39 G. Simons and R. G. Parr, *J. Chem. Phys.*, 1971, **55**, 4197–4202.
40 A. B. Anderson, *J. Mol. Spectrosc.*, 1972, **44**, 411–424.
41 G. B. Sutherland, *J. Chem. Phys.*, 1940, **8**, 161–164.
42 G. Somayajulu, *J. Chem. Phys.*, 1960, **33**, 1541–1553.
43 E. Lippincott, R. Schroeder and D. Steele, *J. Chem. Phys.*, 1961, **34**, 1448–1449.
44 J. Gazquez and R. G. Parr, *Chem. Phys. Lett.*, 1979, **66**, 419–422.
45 B. J. J. Wiener, J. S. Murray, M. E. Grice and P. Politzer, *Mol. Phys.*, 1997, **90**, 425–430.
46 H. Skinner, *Dissociation Energies of Diatomic Molecules*, 1954.
47 K. S. Jhung, I. H. Kim, K.-H. Oh, K. B. Hahn and K. H. C. Jhung, *Phys. Rev. A*, 1990, **42**, 6497.
48 Y.-R. Luo, *Comprehensive handbook of chemical bond energies*, CRC press, 2007.
49 X. Liu, S. Truppe, G. Meijer and J. Pérez-Ríos, *J. Cheminf.*, 2020, **12**, 31.
50 I. C. Stevenson and J. Pérez-Ríos, *J. Phys. B: At., Mol. Opt. Phys.*, 2019, **52**, 105002.
51 J. Fu, Z. Wan, Z. Yang, L. Liu, Q. Fan, F. Xie, Y. Zhang and J. Ma, *Int. J. Quantum Chem.*, 2022, **122**, e26953.
52 R. Pederson, B. Kalita and K. Burke, *Nat. Rev. Phys.*, 2022, **4**, 357–358.
53 R. Nagai, R. Akashi and O. Sugino, *npj Comput. Mater.*, 2020, **6**, 43.
54 M. Gao, B. Cai, G. Liu, L. Xu, S. Zhang and H. Zeng, *Phys. Chem. Chem. Phys.*, 2023, **25**, 9123–9130.
55 S. Dick and M. Fernandez-Serra, *Nat. Commun.*, 2020, **11**, 3509.
56 K.-P. Huber, *Molecular spectra and molecular structure: IV. Constants of diatomic molecules*, Springer Science & Business Media, 2013.
57 K. Huber and G. Herzberg, *NIST Chemistry WebBook*, NIST Standard Reference Database Number 69, 2021.
58 C. J. Evans and M. C. Gerry, *J. Am. Chem. Soc.*, 2000, **122**, 1560–1561.
59 S. Andreev and J. J. BelBruno, *Chem. Phys. Lett.*, 2000, **329**, 490–494.
60 D. Schröder, J. Hrušák, I. C. Tornieporth-Oetting, T. M. Klapötke and H. Schwarz, *Angew. Chem.*, 1994, **106**, 223–225.
61 C. van Wüllen, *J. Chem. Phys.*, 1998, **109**, 392–399.
62 D. Figgen, G. Rauhut, M. Dolg and H. Stoll, *Chem. Phys.*, 2005, **311**, 227–244.
63 K. Saenger and C. Sun, *Phys. Rev. A*, 1992, **46**, 670.
64 E. GharibNezhad, A. Shayesteh and P. F. Bernath, *J. Mol. Spectrosc.*, 2012, **281**, 47–50.
65 T. C. Steimle, T. Ma, A. G. Adam, W. D. Hamilton and A. J. Merer, *J. Chem. Phys.*, 2006, **125**, 064302.
66 Q. Nadhem, S. Behere and S. Behere, *Int. Lett. Chem., Phys. Astron.*, 2015, **58**, 91.
67 R. Ram, P. Bernath and S. Davis, *J. Chem. Phys.*, 1996, **104**, 6949–6955.
68 A. I. Boldyrev and J. Simons, *Periodic Tables of Diatomic Molecules*, John Wiley & Sons, 1997.
69 K. P. Jensen, B. O. Roos and U. Ryde, *J. Chem. Phys.*, 2007, **126**, 014103.
70 H. Wang, X. Zhuang and T. C. Steimle, *J. Chem. Phys.*, 2009, **131**, 114315.
71 M. M. F. de Moraes and Y. A. Aoto, *J. Mol. Spectrosc.*, 2022, **385**, 111611.
72 M. Barnes, D. Clouthier, P. Hajigeorgiou, G. Huang, C. Kingston, A. Merer, G. Metha, J. Peers and S. Rixon, *J. Mol. Spectrosc.*, 1997, **186**, 374–402.
73 F. Liu, F.-X. Li and P. Armentrout, *J. Chem. Phys.*, 2005, **123**, 064304.
74 S. McLamarrah, P. Sheridan and L. Ziurys, *Chem. Phys. Lett.*, 2005, **414**, 301–306.
75 O. Launila, *J. Mol. Spectrosc.*, 1995, **169**, 373–395.
76 S. Mishra, R. K. Yadav, V. Singh and S. Rai, *J. Phys. Chem. Ref. Data*, 2004, **33**, 453–470.
77 D. J. Brugh, M. D. Morse, A. Kalemos and A. Mavridis, *J. Chem. Phys.*, 2010, **133**, 034303.
78 M. A. Garcia, C. Vietz, F. Ruipérez, M. D. Morse and I. Infante, *J. Chem. Phys.*, 2013, **138**, 154306.
79 R. Schlachta, I. Fischer, P. Rosmus and V. Bondybey, *Chem. Phys. Lett.*, 1990, **170**, 485–491.
80 T. D. Persinger, J. Han and M. C. Heaven, *J. Phys. Chem. A*, 2021, **125**, 8274–8281.
81 A. Stein, M. Ivanova, A. Pashov, H. Knöckel and E. Tiemann, *J. Chem. Phys.*, 2013, **138**, 114306.
82 C. Wu, H. Ihle and K. Gingerich, *Int. J. Mass Spectrom. Ion Phys.*, 1983, **47**, 235–238.
83 J. V. Pototschnig, R. Meyer, A. W. Hauser and W. E. Ernst, *Phys. Rev. A*, 2017, **95**, 022501.
84 G. Krois, J. V. Pototschnig, F. Lackner and W. E. Ernst, *J. Phys. Chem. A*, 2013, **117**, 13719–13731.
85 A. Grochola, J. Szczepkowski, W. Jastrzebski and P. Kowalczyk, *J. Quant. Spectrosc. Radiat. Transfer*, 2014, **145**, 147–152.

86 N. Mabrouk, H. Berriche, H. B. Ouada and F. X. Gadéa, *J. Phys. Chem. A*, 2010, **114**, 6657–6668.
87 P. Staanum, A. Pashov, H. Knöckel and E. Tiemann, *Phys. Rev. A*, 2007, **75**, 042513.
88 W. Müller and W. Meyer, *J. Chem. Phys.*, 1984, **80**, 3311–3320.
89 E. J. Breford, F. Engelke, G. Ennen and K. H. Meiwes, *Faraday Discuss. Chem. Soc.*, 1981, **71**, 233–252.
90 K. F. Zmbov, C. Wu and H. Ihle, *J. Chem. Phys.*, 1977, **67**, 4603–4607.
91 T. D. Persinger, J. Han and M. C. Heaven, *J. Phys. Chem. A*, 2021, **125**, 3653–3663.
92 F. Engelke, G. Ennen and K. Meiwes, *Chem. Phys.*, 1982, **66**, 391–402.
93 H. Atmanspacher, H. Scheingraber and C. Vidal, *J. Chem. Phys.*, 1985, **82**, 3491–3501.
94 G. C. Rizkallah, A. A. Assaf and S. N. Tohme, *Chem. Phys.*, 2021, **550**, 111316.
95 L. B. Knight Jr and W. Weltner Jr, *J. Chem. Phys.*, 1971, **54**, 3875–3884.
96 R. S. DaBell, R. G. Meyer and M. D. Morse, *J. Chem. Phys.*, 2001, **114**, 2938–2954.
97 D. J. Brugh, T. J. Ronningen and M. D. Morse, *J. Chem. Phys.*, 1998, **109**, 7851–7862.
98 S. Leutwyler, M. Hofmann, H.-P. Harri and E. Schumacher, *Chem. Phys. Lett.*, 1981, **77**, 257–260.
99 M. Chaieb, H. Habli, L. Mejrissi, B. Oujia and F. X. Gadéa, *Int. J. Quantum Chem.*, 2014, **114**, 731–747.
100 O. Docenko, M. Tamanis, R. Ferber, A. Pashov, H. Knöckel and E. Tiemann, *Phys. Rev. A*, 2004, **69**, 042503.
101 N. Takahashi and H. Katô, *J. Chem. Phys.*, 1981, **75**, 4350–4356.
102 B. Simard, P. I. Presunka, H. P. Loock, A. Bérces and O. Launila, *J. Chem. Phys.*, 1997, **107**, 307–318.
103 J. Ogilvie and F. Y. Wang, *J. Mol. Struct.*, 1992, **273**, 277–290.
104 D. J. Brugh and M. D. Morse, *J. Chem. Phys.*, 2002, **117**, 10703–10714.
105 R. Ram and P. Bernath, *J. Mol. Spectrosc.*, 1992, **155**, 315–325.
106 R. Ram, S. Yu, I. Gordon and P. Bernath, *J. Mol. Spectrosc.*, 2009, **258**, 20–25.
107 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.
108 J. D. Langenberg, L. Shao and M. D. Morse, *J. Chem. Phys.*, 1999, **111**, 4077–4086.
109 J. O. Schroeder, C. Nitsch and W. E. Ernst, *J. Mol. Spectrosc.*, 1989, **132**, 166–177.
110 W. Ernst, J. Schröder and B. Zeller, *J. Mol. Spectrosc.*, 1989, **135**, 161–168.
111 S. Antrobus, D. Husain, J. Lei, F. Castaño and M. S. Rayo, *Z. Phys. Chem.*, 1995, **190**, 267–287.
112 A. Bernard, C. Effantin, J. d'Incan, A. Topouzkhanian and G. Wannous, *J. Mol. Spectrosc.*, 1999, **195**, 11–21.
113 M. W. Chase and N. I. S. O. (US), *NIST-JANAF thermochemical tables*, American Chemical Society Washington, DC, 1998, vol. 9.
114 V. Belyaev, I. Gotkis, N. Lebedeva and K. Krasnov, *Russ. J. Phys. Chem.*, 1990, **64**, 773.
115 T. Imajo, Y. Kobayashi, Y. Nakashima, K. Tanaka and T. Tanaka, *J. Mol. Spectrosc.*, 2005, **230**, 139–148.
116 R. Ram and P. Bernath, *J. Mol. Spectrosc.*, 2005, **231**, 165–170.
117 R. Ram, J. Peers, Y. Teng, A. Adam, A. Muntianu, P. Bernath and S. Davis, *J. Mol. Spectrosc.*, 1997, **184**, 186–201.
118 M. Burton and L. Ziurys, *J. Chem. Phys.*, 2019, **150**, 034303.
119 I. O. Antonov and M. C. Heaven, *J. Phys. Chem. A*, 2013, **117**, 9684–9694.
120 R. Ram, P. Bernath and S. Davis, *J. Chem. Phys.*, 2002, **116**, 7035–7039.
121 L. A. Kaledin, J. E. McCord and M. C. Heaven, *J. Mol. Spectrosc.*, 1994, **164**, 27–65.
122 B. Simard, P. A. Hackett and W. J. Balfour, *Chem. Phys. Lett.*, 1994, **230**, 103–109.
123 I. Shim, M. Pelino and K. A. Gingerich, *J. Chem. Phys.*, 1992, **97**, 9240–9248.
124 M. Sievers, Y.-M. Chen and P. Armentrout, *J. Chem. Phys.*, 1996, **105**, 6322–6333.
125 M. Bobetic and J. Barker, *J. Chem. Phys.*, 1976, **64**, 2367–2369.
126 K. Tang and J. Toennies, *J. Chem. Phys.*, 2003, **118**, 4976–4983.
127 L. Piticco, F. Merkt, A. A. Cholewinski, F. R. McCourt and R. J. Le Roy, *J. Mol. Spectrosc.*, 2010, **264**, 83–93.
128 C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, The MIT Press, 2005.
129 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.
130 Z. Botev and A. Ridder, *Wiley statsRef: Statistics reference online*, 2017, pp. 1–6.
131 T. Head, M. Kumar, H. Nahrstaedt, G. Louppe and I. Shcherbatyi, *scikit-optimize/scikit-optimize*, 2021, DOI: 10.5281/zenodo.5565057.
132 X. Liu, L. McKemmish and J. Pérez-Ríos, *Phys. Chem. Chem. Phys.*, 2023, **25**, 4093–4104.
133 H. Ladjimi and M. Tomza, *Diatomic molecules of alkali-metal and alkaline-earth-metal atoms: interaction potentials, dipole moments, and polarizabilities*, 2023.
134 A.-R. Allouche and M. Aubert-Frécon, *Chem. Phys. Lett.*, 1994, **222**, 524–528.
135 L. Russon, G. Rothschopf, M. Morse, A. Boldyrev and J. Simons, *J. Chem. Phys.*, 1998, **109**, 6655–6665.
136 G. Gopakumar, M. Abe, M. Hada and M. Kajita, *J. Chem. Phys.*, 2013, **138**, 194307.
137 N.-H. Cheung and T. A. Cool, *J. Quant. Spectrosc. Radiat. Transfer*, 1979, **21**, 397–432.
138 J. K. Laerdahl, T. Saue and K. Faegri Jr, *Theor. Chem. Acc.*, 1997, **97**, 177–184.
139 P. Schwerdtfeger, J. S. McFeaters, R. L. Stephens, M. J. Liddell, M. Dolg and B. A. Hess, *Chem. Phys. Lett.*, 1994, **218**, 362–366.

140 R. Gosavi, G. Greig, P. Young and O. Strausz, *J. Chem. Phys.*, 1971, **54**, 983–991.
141 A. Givan and A. Loewenschuss, *J. Mol. Struct.*, 1982, **78**, 299–301.
142 S. Elmoussaoui and M. Korek, *Comput. Theor. Chem.*, 2015, **1068**, 42–46.
143 C. Krumrey, S. A. Cooke, D. K. Russell and M. C. Gerry, *Can. J. Phys.*, 2009, **87**, 567–573.
144 R. S. Ram, J. Liévin, G. Li, T. Hirao and P. F. Bernath, *Chem. Phys. Lett.*, 2001, **343**, 437–445.
145 S. M. Sickafoose, A. W. Smith and M. D. Morse, *J. Chem. Phys.*, 2002, **116**, 993–1002.
146 M. Flory, S. McLamarrah and L. Ziurys, *J. Chem. Phys.*, 2006, **125**, 194304.
147 E. Tenenbaum, M. Flory, R. Pulliam and L. Ziurys, *J. Mol. Spectrosc.*, 2007, **244**, 153–159.
148 X. Qu, D. A. Latino and J. Aires-de Sousa, *J. Cheminf.*, 2013, **5**, 34.
149 A. Raza, S. Bardhan, L. Xu, S. S. R. K. C. Yamijala, C. Lian, H. Kwon and B. M. Wong, *Environ. Sci. Technol. Lett.*, 2019, **6**, 624–629.
150 J. Fu, S. Long, J. Jian, Z. Fan, Q. Fan, F. Xie, Y. Zhang and J. Ma, *Spectrochim. Acta, Part A*, 2020, **239**, 118363.
151 (*a*) M. W. Wolkenstein, *Molekular Optics*, Moscow-Leningrad, 1951; (*b*) *The Structure and Physical Properties of Molecules*, Moscow-Leningrad, 1955.
152 R. Kepa, M. Ostrowska-Kopeć, I. Piotrowska, M. Zachwieja, R. Hakalla, W. Szajna and P. Kolek, *J. Phys. B: At., Mol. Opt. Phys.*, 2014, **47**, 045101.
153 S. Van Huffel and P. Lemmerling, *Total least squares and errors-in-variables modeling: analysis, algorithms and applications*, Springer Science & Business Media, 2013.