Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 239 (2020) 118363

Contents lists available at ScienceDirect

Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy

journal homepage: www.elsevier.com/locate/saa

A joint data and model driven method for study diatomic vibrational spectra including dissociation behavior

Jia Fu [a], ShanShan Long [a], Jun Jian [a], Zhixiang Fan [a,*], Qunchao Fan [a,*], Feng Xie [b], Yi Zhang [c], Jie Ma [d]

[a] College of science, Xihua University, Chengdu 610039, China
[b] Institute of Nuclear and New Energy Technology, Collaborative Innovation Center of Advanced Nuclear Energy Technology, Key Laboratory of Advanced Reactor Engineering and Safety of Ministry of Education, University, Beijing 100084, China
[c] College of Advanced Interdisciplinary Studies, National University of Defense Technology, Changsha, Hunan 410073, China
[d] State Key Laboratory of Quantum Optics and Quantum Optics Devices, Laser Spectroscopy Laboratory, College of Physics and Electronics Engineering, Shanxi University, Taiyuan 030006, China

ARTICLE INFO

Article history:
Received 15 January 2020
Received in revised form 5 April 2020
Accepted 10 April 2020
Available online 07 May 2020

Keywords:
Vibrational spectroscopy
Dissociation energy
Machine learning

ABSTRACT

The details of quantum multi-body interactions are so rich and subtle which make it difficult to accurately model for some situations such as the behavior of diatomic long-range vibrations. In recent years, data-driven machine learning has made remarkable achievements in capturing complex relationships that are subtle. Combining the characteristics of these two fields, we propose a joint machine learning method to obtain reliable diatomic vibrational spectra including dissociation energy by using accessible heterogeneous micro/macro information such as low lying vibrational energy levels and heat capacity. Applications of this method to *CO* and Br_2 in the ground state yield their state of the art of vibrational spectra including dissociation limit. The strategy introduced here is an exploration of combining the model-driven and data-driven method to cover subtle physical details that are difficult to study in a single way.

PACS:
33.20.Sn; 33.20.Vq; 33.15.Mt

PACS:
33.20.Sn
33.20.Vq
33.15.Mt

1. Introduction

The spectroscopy of molecules plays an important role in studying physical and chemical phenomena in atomic level. For a long time, people have made great progress in studying molecular structures, chemical reactions, astrophysics, interstellar matter, nuclear energy and fundamental constants by taking advantage of molecular spectra (and corresponding energy levels) [1–6].

In recent years, considerable progress has been made in the experimental study of vibration-rotational energy spectra of small molecular systems especially for diatomic molecules. Many spectrometric techniques (microwave, molecular beam, sub-doppler laser spectroscopy, infrared spectroscopy, photoassociation spectroscopy, etc.) have been developed for “hot molecules” [7,8] and “ultra-cold molecules” [9,10]. By these techniques, in the “hot molecule” part, the spectral lines in the short-range region can be located with high quality. However, it is still difficult to obtain the same achievements in the medium-long range region due to the reasons of doppler broadening, laser continuous modulation limit and too small transition probability. In the “ultra-cold molecules” part, thanks to the extremely low temperature of the sample, the doppler effect is almost negligible, a few near-dissociation energy levels can be measured [11]. At the limit of the vibrational energy sequence lies the dissociation energy, which is even more difficult for direct spectroscopy measurement. A few molecules can be measured directly [12] and many others are obtained by indirect methods like near disassociation expansion [13]. In general, part of the vibration-rotational spectrum can been obtained experimentally in a high quality but the remain is still hard to acquire.

On the theoretical side, it is mainly dependent on the ab initio methods (such as Hartree-Fock and Post-HF Extensions [14], Density Functional Theory [15]) based on quantum mechanics to calculate the spectra. For some simple one electron diatomic molecules (such as HD^+), one can rely on more precise QED [5] method. However, even in diatomic molecule, there may be hundreds of

* Corresponding authors.
E-mail addresses: fanzhixiang235@126.com (Z. Fan), fanqunchao@sina.com (Q. Fan).

https://doi.org/10.1016/j.saa.2020.118363
1386-1425/

electrons, which can lead to very complicated multi-body interactions, letting alone incomplete basis set, relativistic effect [16], etc. One has to model these things approximately according to the nature of particular system, and make a delicate balance between the computational cost and model complexity in real calculation (neglecting a part of correlation effects and using truncated basis sequence). Consequently, many ab initio approximation methods are introduced under the framework of Post-HF (CI, MCSCF, CASSCF, CASPT2, MRCI, etc.) [14] and DFT (LDA, GGA, mGGA, Hybrid, Double Hybrid, etc.) [15]. It turns out that these approximations still lag behind the experimental results in the position far away from the equilibrium especially for exited electron states. The diatomic calculation error for dissociation energy may usually be in hundreds cm^{-1} [17,18] and in tens [17] for some vibrational energy levels.

While there are computational problems in ab initio and QED methods, they confirm us that each quantum configuration (electron state, vibration and rotation) does correspond to a certain energy level (bound state) and provide computed levels, however, with uncertainty. It's just too arduous to eliminate these uncertainties by considering more detailed interactions with acceptable computing power. From an information theory [19] point of view, data (information) can serve to reduce uncertainty. In our previous study [20,21] dissociation energy and low energy level data are successfully used to predict the ro-vibrational spectra. However, on the one hand, the dissociation energy of some molecular systems may be difficult to obtain or have a large uncertainty, it is very worthy of study in itself. On the other hand, there may be over-fitting problems for all the data is used for building up the projection function, which makes it difficult to assess its reliability. In recent years, machine learning has found ways to build reliable higher-dimensional functions from data [22], which performs well in solving quantum and statistical mechanics problems [23–25]. Their main strategy is to transform modeling into optimizing (uncertainty is covered by parameters), and separate the fitting and testing procedure with different data to avoid over-fitting. [22].

The main purpose of this work are to combine the model-driven quantum result with data-driven method to overcome the weakness of their own to cover subtle physical effects in a reliable machine learning way. The findings show that the new method can obtain spectra including the unknown high-lying vibrational levels and dissociation energy under experimental accuracy at a low computational expense.

2. Theoretical method

2.1. Quantum models that are flexible enough to cover subtle physical effects in long-range vibrations

The relationship between quantum number ν and energy levels of bound state is definite, however, the exact form may be very complicated and change with physical effects detail such multi-body interaction. The components (nuclei and electronics) and their interactions (coulombic electron-electron, electron-nucleus and nucleus-nucleus) of diatomic molecule are clear. Consequently, the Hamiltonian of the system is as follows:

$$\begin{aligned}\hat{H} &= \hat{T} + \hat{V} \\ &= -\frac{\hbar^2}{2}\sum_{\alpha}\frac{1}{m_\alpha}\nabla_\alpha^2 - \frac{\hbar^2}{2m_e}\sum_i \nabla_i^2 + \sum_\alpha\sum_{\alpha>\beta}\frac{Z_\alpha Z_\beta e^2}{r_{\alpha\beta}} \\ &\quad - \sum_\alpha\sum_i \frac{Z_\alpha e^2}{r_{i\alpha}} + \sum_j\sum_{i>j}\frac{e^2}{r_{ij}}\end{aligned} \tag{1}$$

where Greek alphabets stand for nuclei and digital labels are on behalf of electrons. Then, for a stable system, the time-independent Schrödinger equation is

$$\begin{aligned}\hat{H}\ \ &\psi(r_1, r_2, \ldots, r_n, R_\alpha, R_\beta, \ldots, R_m) \\ &= E\psi(r_1, r_2, \ldots, r_n, R_\alpha, R_\beta, \ldots, R_m)\end{aligned} \tag{2}$$

After introducing the potential energy hyper-surface (PES) according to Born-Oppenheimer approximation (BOA) and considering the geometrical symmetry of two atoms, the state of nuclei (vibration and rotation) is contained in the radial equation:

$$\left\{-\frac{\hbar^2}{2\mu}\frac{d^2}{dr^2} + V(r) + \frac{\left[J(J+1) - \Lambda^2\right]\hbar^2}{2\mu r^2}\right\}\varphi(r) = E_{\nu J}\varphi(r) \tag{3}$$

J, Λ and ν represent the total angular momentum quantum number, the absolute value of the projection of electron orbital angular momentum onto the nucleus line (correspond to electronic state) and the vibrational quantum number. The diatomic PES is just a function of nuclear distance. One can expand it to its eighth order at the equilibrium

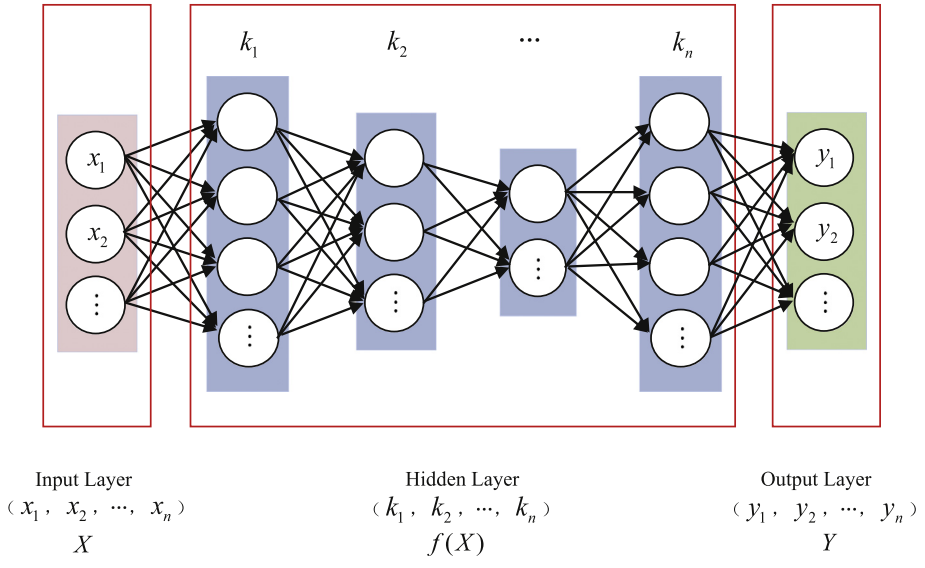

Fig. 1. A typical Artificial neural network to build relationship between *X* and *Y*.

position [26]:

$$V(r) = \sum_{n=2}^{n_{max}=8} \frac{1}{n!} f_n (r - r_e)^n = \sum_{n=2}^{n_{max}=8} \frac{1}{n!} f_n x^n \tag{4}$$

$f_n = V^{(n)}(r_e) = (\frac{d^n V}{dr^n})_{r=r_e}$ is the so called n-rank force constant. Using the second order perturbation method to regroup the Hamiltonian in Eq. (3):

$$\hat{H} = \hat{H}_0 + \hat{H}' \tag{5}$$

where $\hat{H}_0 = -\frac{\hbar^2}{2\mu}\frac{d^2}{dx^2} + \frac{1}{2} f_2 x^2$,

$$\hat{H}' = \frac{1}{6} f_3 x^3 + \ldots + \frac{1}{40320} f_8 x^8 + \frac{\{J(J+1) - \Lambda^2\}\hbar^2}{2\mu r^2}$$

and the final energy levels are

$$E_{vJ} = E_v^{(0)} + H'_{vv} + \sum_{m \neq v} \frac{\left|H'_{vm}\right|^2}{E_v^{(0)} - E_m^{(0)}} \tag{6}$$

When we focus on the vibrational part, the corresponding energy levels are:

$$\begin{aligned} E_v = \; & \omega_0 + (\omega_e + \omega_{e0})\left(v + \frac{1}{2}\right) - \omega_e x_e \left(v + \frac{1}{2}\right)^2 + \\ & \omega_e y_e \left(v + \frac{1}{2}\right)^3 + \omega_e z_e \left(v + \frac{1}{2}\right)^4 + \cdots \end{aligned} \tag{7}$$

and dissociation energy is a function of last three vibrational energy [20]:

$$D_e^{cal} \cong E_{v_{max}} + \frac{\Delta E_{v_{max}, v_{max-1}}{}^2}{\Delta E_{v_{max}, v_{max-2}} - \Delta E_{v_{max}, v_{max-1}}} \tag{8}$$

The perturbation result has the similar Herzberg [27] like polynomial form as well as Dunham [28] under WKB approximation. Comparing it to Taylor expansion of normal function

$$f(x) = \frac{f(x_0)}{0!} + \frac{f'(x_0)}{1!}(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \ldots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n + R_n(x) \tag{9}$$

one can find that, $E(v)$ are actually be expanded as series at $v = -1/2$. Any complicated physical effects can be reflected in the expansion coefficient that is usually called spectroscopic constants [8]. In previous quantum study the function $E(v)$ is expanded at low v, so the long-range vibrations according to high v are far from the expansion point, which means long-range vibrations may require a lot of expansion terms. In fact, spectroscopic constants are usually fitted by the low-lying experimental levels through least square method. However, with the increasing of polynomial items number, the fitting ability increases rapidly. Complex details like nonphysical experimental errors and subtle physical effects can be covered as an indistinguishable whole (over-fitting). Therefore, in practice, only few expansion items (usually less than 4 [7,8]) are kept in order to gain statistical significant. Higher-order constants that related to long-range interactions but mixed with uncontrollable errors inside the experimental results have to be abandoned (under-fitting). Therefore, the least square strategy leads to poor performance in the predictive extrapolation for high-lying vibrational and disassociation energy [20]. We have to find a new way to unlock the power of higher-order items in Eq. (7) to achieve long-range vibrations.

Table 1
Ground state dissociation energy of *CO* [30].

D_e[a]	Year	Method
55,821.120	1936	Spectrum
70,976.136	1939	Electron impact
75,815.428	1947	Theoretical calculation
81,461.247	1943	Spectrum
89,615.437	1945	Spectrum
90,679.1[b]	2014	Spectrum

[a] Dissociation energy in cm^{-1}.
[b] Newly added from [31].

2.2. Data driven machine learning method

To unlock the power of functions with massive parameters, machine learning is the master [22]. It can approximate higher-dimensional functions to forecast globally rather than a local fitting by making good use of data. First, it introduces a function that may contain billions of parameters such Deep-neural-network (DNN), recurrent-neural-network (RNN), convolutional-neural-network (CNN) to cover the relationship it want to build such as $Y = f(X)$ in Fig. 1 (to deal with under-fitting problem). Then, the data (sample of $X \to Y$) are used to determine the parameters which are called learning. In order to deal with the over-fitting problem, the data set is divided into training and testing part. In the learning procedure, only training data is used to determine the parameter by optimizing the difference between prediction and exist data in training set. The learning results are final assessed by the data in test set. It is worth to note that, cross-validation, normalization and so on are introduced to settle over-fitting issues (The details can be found in [22]). The main idea of them can be summarized as introducing big enough but restricted parameter space for valid model searching by training and testing with different data.

Back to spectral part, Eq. (7) is a better choice than artificial neural networks for the parameter are much fewer and the flexibility is confirmed by bound state quantum theory and Taylor's series Eq. (9). The challenge remains solving the under and over fitting problem according to its own characteristic, which are managed by spectroscopy learning that is inspired by general machine learning.

2.3. Spectroscopy learning

2.3.1. A formal description of spectroscopy learning

Under the framework of general machine learning, the task of spectroscopy learning is defined to find a reliable functional approximation

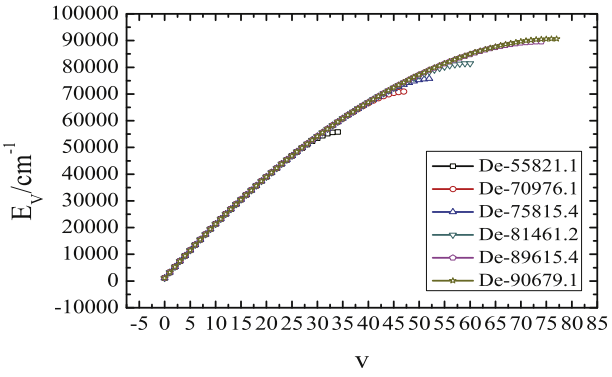

Fig. 2. The full vibrational spectrum corresponding to different dissociation energies for ground state of *CO*.

Table 2
Full vibrational spectrum prediction of *CO* molecule in ground state.

ν	E_ν^{exi} [32]	E_ν^{cal}	ν	E_ν^{cal}
0	1081.701	1081.756	42	69,159.054
1	3225.042	3225.036	43	70,251.348
2	5341.833	5341.831	44	71,319.269
3	7432.210	7432.210	45	72,362.715
4	9496.241	9496.242	46	73,381.565
5	11,533.994	11,533.995	47	74,375.680
6	13,545.540	13,545.541	48	75,344.898
7	15,530.954	15,530.954	49	76,289.034
8	17,490.307	17,490.307	50	77,207.878
9	19,423.677	19,423.677	51	78,101.196
10	21,331.141	21,331.141	52	78,968.723
11	23,212.778	23,212.778	53	79,810.166
12	25,068.668	25,068.668	54	80,625.202
13	26,898.893	26,898.893	55	81,413.472
14	28,703.535	28,703.535	56	82,174.582
15	30,482.679	30,482.679	57	82,908.102
16	32,236.407	32,236.407	58	83,613.561
17	33,964.805	33,964.805	59	84,290.446
18	35,667.957	35,667.957	60	84,938.200
19	37,345.949	37,345.949	61	85,556.217
20	38,998.865	38,998.865	62	86,143.843
21	40,626.788	40,626.788	63	86,700.371
22	42,229.802	42,229.802	64	87,225.037
23	43,807.989	43,807.989	65	87,717.022
24	45,361.428	45,361.428	66	88,175.441
25	46,890.196	46,890.196	67	88,599.345
26	48,394.370	48,394.370	68	88,987.720
27	49,874.020	49,874.020	69	89,339.474
28	51,329.216	51,329.216	70	89,653.443
29	52,760.022	52,760.022	71	89,928.381
30	54,166.498	54,166.498	72	90,162.961
31	55,548.698	55,548.698	73	90,355.764
32	56,906.672	56,906.672	74	90,505.279
33	58,240.461	58,240.460	75	90,609.901
34	59,550.101	59,550.099	76	90,667.917
35	60,835.619	60,835.616	77	90,677.513
36	62,097.034	62,097.029		
37	63,334.355	63,334.347		
38	64,547.581	64,547.568		
39	65,736.698	65,736.681		
40	66,901.681	66,901.660		
41	68,042.490	68,042.469		
D_e^{exi}	90,679.1 [31]		D_e^{cal}	90,679.099

to construct the whole vibrational spectrum by drawing information from accessible experimental or theoretical data. For diatomic case, take Eq. (7) into matrix form:

$$AX = E \tag{10}$$

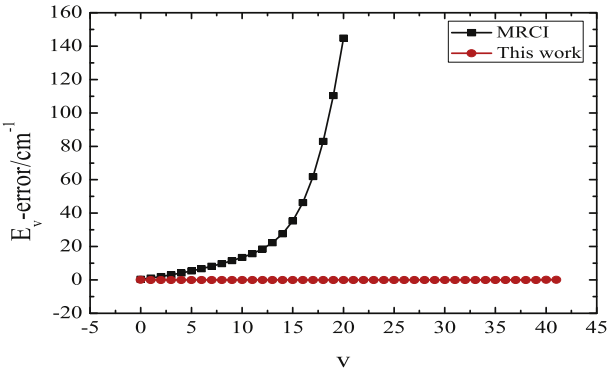

Fig. 4. Error comparison of different vibration levels for the ground state of *CO* [32,34].

where

$$A_{\nu k} = \left(\nu + \frac{1}{2}\right)^k, X = \begin{pmatrix} \omega_0 \\ \omega_e' \\ -\omega_e x_e \\ \vdots \end{pmatrix}, E = \begin{pmatrix} E_{\nu_1} \\ E_{\nu_2} \\ E_{\nu_3} \\ \vdots \end{pmatrix} \tag{11}$$

Eq. (10) provides alternative functions of any dimension through arguments X to cover the valid spectrum. However, due to the unavoidable uncertainty in existing samples (theoretical or experimental) $\{E_n^{exi}\} \coloneqq \{(\nu_i, E_{\nu_i})\}$, the equation cannot be solved directly and the high-lying energy levels are always not accessible. Thus, the descriptive statistical approaches like least square method are introduced but generally can make good use of the previous few parts of X for low-lying vibrational levels. In order to improve the prediction of high-lying part, one has to find a way to unlock the power of much higher order coefficients in X while avoiding over-fitting.

Big enough restricted parameter space.

For getting a moderately flexible model to avoid under and over-fitting, reasonable restrictions should be applied to the range of higher order parameters. Eq. (10) can play an important role here, by which, one can obtain the one-to-one mapping from $\{E_n^{exi}\}$ to constants X.

$$X = EA^{-1} \tag{12}$$

The $\{E_n^{exi}\}$ given by experiments/calculations could own high quality (generally within $1cm^{-1}$) in low-lying levels. Therefore, through

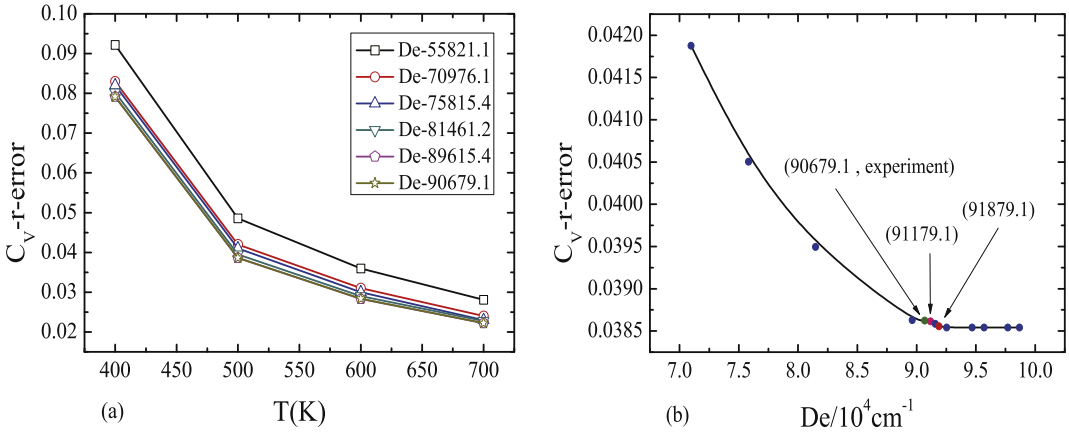

Fig. 3. (a) The vibrational molar heat capacity relative errors based on different dissociation energy comparing to experiment for the ground state of *CO*. (b) The relative error correspondence between dissociation energy and heat capacity (500 K) in detail [33].

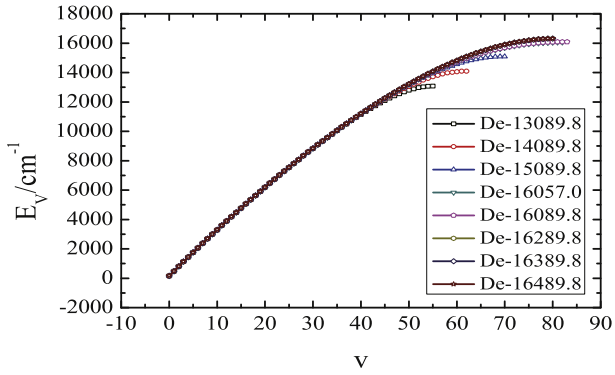

Fig. 5. The full vibrational spectra corresponding to different dissociation energies for the ground state of Br_2.

Eq. (12) the value of X could be firstly restricted by

$$\Delta X = \Delta E A^{-1} \tag{13}$$

On the other hand, experiments/calculations can usually provide dozens of energy levels, which means that scores of elements in X can be produced through Eq. (12). Occam's razor is used to further confine the size of X. The simpler model (X) with enough expression are preferred.

Data set in practice.

Dozens of $\{E_n^{exi}\}$ could be collected from experiments/calculations. Heterogeneous information such as dissociation energy (D_e in Eq. (8)) and molar vibrational heat capacity [29] (C_{mol} in Eq. (14)) are introduced to enhance the dataset, which are also accessible.

$$C_{mol}^{cal} = \frac{N_A}{kT^2}\left(\left\langle E_v^2\right\rangle - \langle E_v\rangle^2\right) \tag{14}$$

Finally, the spectroscopy learning task is to make good use of data set $\{E_n^{exi}\}$, D_e and C_{mol} to locate a suitable formula f in terms of Eq. (11) to predict the whole vibrational spectroscopy $\{E_n\}$ by learning algorithm described below.

2.3.2. Learning algorithm

• Learning is optimization

There are many possibilities of molecular constants X even if their scope are limited by Eq. (13). The learning is to find a prophetic one by minimize the distance between reconstructed data with the input data. In our case, there can be tree concerns:

$$X^* = \arg\min_X \|E^{exi} - AX\| \tag{15}$$

$$X^* = \arg\min_X \|D_e^{exi} - D_e^{cal}(X)\| \tag{16}$$

$$X^* = \arg\min_X \|C_{mol}^{exi} - C_{mol}^{cal}(X)\| \tag{17}$$

where X are restricted in Eq. (13). D_e^{cal} depends on X as a combination of Eqs. (8) and (10). C_{mol}^{cal} depends on X as a combination of Eqs. (14) and (10). The distance is defined as

$$\overline{\Delta E} = \sqrt{\frac{1}{m}\sum_{v=0}^{m-1}\left|E_{v,exi} - E_{v,cal}\right|^2} \tag{18}$$

for energy levels and

$$\overline{\Delta D_e} = \left| D_e^{cal} - D_e^{exi} \right| \tag{19}$$

$$\overline{\Delta C_{mol}} = \left| C_{mol}^{cal} - C_{mol}^{exi} \right| \tag{20}$$

for dissociation energy and heat capacity. All above three optimization goals can be used to get X^*. In order to predict spectra without experimental D_e, we use Eqs. (15) and (16) to obtain X^* while treating D_e as a pending parameter. The heat capacity is further introduced as an additional physical criterion to determine the D_e.

• Greedy algorithm

In the adjustment to finding X^* within tolerance, the greedy algorithm was adopted. Parameters are adjusted one by one from the lowest-order one. This compresses the search problem of n-dimensional space into n 1-dimensional spaces. In detail, two small variations ($\delta E = 1cm^{-1}$) are first added to get three try points $E_v^1 \in \{E_v^{ex} - \delta E_v, E_v^{exi}, E_v^{exi} + \delta E_v\}$ and select the one with the least error (according to Eqs. (18) and (19)), then halve the variations to get three new points $E_v^2 \in \{E_v^{exi} - 0.5\delta E_v, E_v^{exi}, E_v^{exi} + 0.5\delta E_v\}$ and choose the one with the least error. At last, iterate the operation until convergence (the last significant digit, $0.001cm^{-1}$ here) is reached.

• Initial values guessing

The selection of initial value plays an important part in optimization problems. If the initial value is close to the optimal value, the final results are easy to locate by simpler searching algorithm like we used here. According to Eq. (10), if m parameters are undetermined, only m

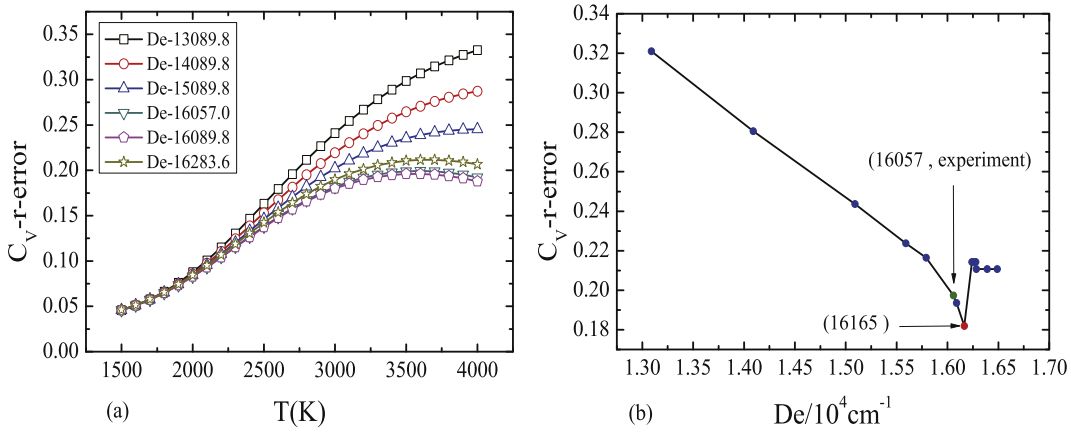

Fig. 6. (a) The vibrational molar heat capacity relative errors based on different dissociation energy comparing to experiment for the ground state of Br_2. (b) The relative error correspondence between dissociation energy and heat capacity (3800 K) in detail [33].

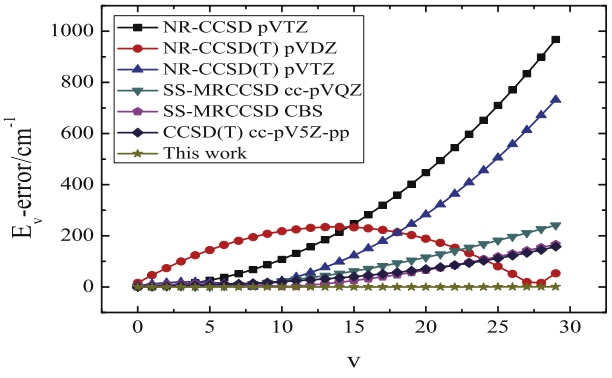

Fig. 7. Vibrational level errors of different methods for the ground state of Br2 [35–37].

elements of $\{E_n^{exi}\}$ are needed. In fact, the number of elements in $\{E_n^{exi}\}$ is usually larger than the size of X^*. There are C_n^m alternatives to solve Eq. (10) and produce X^*. If $\{E_n^{exi}\}$ is exactly correct, then the C_n^m alternatives are equivalent. Whereas, the correct answers distribute around $\{E_n^{exi}\}$ in a narrow breadth, which makes C_n^m alternatives become different qualified starting points.

• Testing

The dissociation energy can server as an independent test for the spectrum constructed from optimized X^* because its exact value is the most difficult to predict. By comparing it with experimental counterpart, one can show the reliability of the method and avoid over-fitting.

2.3.3. Calculating details

1) For a certain dissociation energy D_e, try 5 low-order parameters as the initial attempt for the size of X in Eq. (10).

2) Select five levels from the known energy levels and use greedy algorithm to solve the five parameters in step 1). As described in Initial values guessing, if there are m lines, there can be C_m^5 choice for the calculation, which finally gives C_m^5 different answers to the 5 parameters.

3) Verify the parameters that best meet the conditions $\overline{\Delta E}$<0.5cm^{-1} in Eq. (18) and $\overline{\Delta D_e}$<10cm^{-1} in Eq. (19). It is worth noting that the criterion can also ensure that the final error given by the parameter solution found by different initial values in step 2) is very small.

4) If step 3) is satisfied, the calculation is over. Conversely, if the condition cannot be met, increasing the number of parameters by 1 (to 6 this time) and repeat steps 1) to 4) until step 3) is satisfied.

5) When step 4) is down, one can calculate heat capacity according to the spectra with Eq. (14), then by changing the D_e used in step 1) one can draw the heat capacity error curve with different dissociation energy and determine the best D_e (first inflection point, see the following cases of *CO* and Br_2).

Table 3
The error of dissociation energy calculated by different methods of Br_2 in the ground state. [36,37].

Method	Basis	Error (cm^{-1})
SS-MRCCSD	cc-pVQZ	3270
SS-MRCCSD	CBS	3547
SS-MRCCSD	cc-pVTZ-PP	3169
CCSD(T)	cc-pVDZ-PP	2731
AF-QMC	cc-pVDZ-PP	1332
AF-QMC	cc-pVTZ-PP	2130
DC-HF	pVDZ	12,909
DC-HF	pVTZ	10,566
NR-CCSD	pVDZ	3396
NR-CCSD	pVTZ	1053
NR-CCSD(T)	pVDZ	2626
NR-CCSD(T)	pVTZ	67
This work		108

3. Application

Continuous researches on carbon monoxide across many years provide historical increasingly reliable data, which could be used as touchstone for our method. It can be seen from Table 1 that the dissociation energy varies greatly during different times.

Following the steps described in Section 2.3, the full vibrational spectrum corresponding to various dissociation energies in Table 1 are calculated and shown in Fig. 2. It can be seen from the figure that dissociation energy has a great influence on spectra prediction (overfitting) and levels closer to it are greatly affected (Difference in $10^4 cm^{-1}$). The latest ($D_e^{expt} = 90{,}679.1 cm^{-1}$) [31]) is listed in Table 2. The levels and dissociation energy errors are all within 0.1 cm^{-1}. The nine selected levels to calculate full spectrum are over-strike. It can be seen that the "Initial values guessing" algorithm in Section 2.3 works well. The optimal values are directly found by changing the choice.

Because data in [32] has been used to learn the parameters (see Eqs. (15) and (16)), additional information (C_{mol} in Eq. (14)) is introduced for further testing. As shown in Fig. 3 (a), the more accurate the dissociation energy is, the more reliable the calculated spectrum will be. The best C_{mol} still bound to latest D_e^{expt} (90,679.1cm^{-1} [31]).

The dependence on heat capacity for dissociation energy suggests a way to obtain the latter by the former and make dissociation energy a good criteria to test the reliability of the whole method. The correspondence between them is expanded to detail and illustrated with Fig. 3 (b). As the dissociation energy increases, the heat capacity error decreases to near latest D_e^{expt}. Considering the uncertainty of heat capacity, detail changes after the first turning point may be ignored and 91,179.1 cm^{-1} could be the estimate with an absolute error of 500 cm^{-1} (5.5‰), which is better than the second best in Table 1. The corresponding energy levels are compared with the MRCI method [34] in Fig. 4. The MRCI method obtains 21 vibrational levels from $\nu = 0$ to $\nu = 20$. The error (compared to [32]) increases with the vibration number rapidly and reaches 144.835 cm^{-1} at $\nu = 20$. In this work, the errors are all within the range of 0.13 cm^{-1} and tend to be stable. The detailed data sheet is listed in the supplementary materials (Table S1).

In order to further test the method, similar analysis is applied to Br_2 in ground state. Dissociation energy is treated as an unknown parameter as X in Eq. (10). Several candidates across the experimental one (16,057 cm^{-1} [35]) are illustrated with Fig. 5. It can be seen that the variance in dissociation energy has great influence on the construction of vibrational spectrum. On the heat capacity side, situation is also similar. The correspondence is shown in Fig. 6. It can be seen that higher or lower dissociation limit will raise the error. The best choice is 16,165 cm^{-1} with a shift of 108 cm^{-1} to experiment (6.7‰). The final results are further compared with several other method as shown in Fig. 7 for energy levels and in Table 3 for dissociation energy. In Fig. 7, the errors of this work are all within the range of 0.36 cm^{-1} and tend to be stable, while the errors of other methods are much greater. Like the situation in *CO* system, the error of other method increases with the vibration quantum number rapidly and reach around 100 cm^{-1} at $\nu = 20$. In Table 3, the results given here (error is 108 cm^{-1}) are more accurate than most other methods (error distributed from 67 cm^{-1} to 12,909 cm^{-1} and most of them are over 1000 cm^{-1}) in comparison with the experiment. The only exception is NR-CCSD(T) with pVTZ basis (67 cm-1). However, as shown in Fig. 7, although it yields impressive result in the calculation of dissociation energy, it performs poorly in the prediction of vibration levels.

4. Discussion

Suggestion on how to collect and use data efficiently.

The solution may come from Bayes' theory:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{21}$$

Suppose a particular model A is contained in a big model set ($P(A)$ is small) and evidence B can be derived from A ($P(B|A)$ is large). According to Eq. (21), if the evidence B is hard to happen ($P(B)$ is very small) and does take place, then the confidence in model A is greatly increased ($P(A|B) > P(A)$). In our method, only part of the energy levels is used to generate the model sets, the dozen others together with dissociation behavior and heat capacity are used as evidences (validation and testing) that are hard to guess from merely several levels. Their co-occurrence is even rarer. The learning method introduce here is actually to use great information carried by rare evidence to eliminate uncertainty in full spectrum prediction. Therefore, in order to collect and use data efficiently, when collecting data, attention should be paid to those with large amounts of information (rare evidence), and when using data, model sets should be generated with as little data as possible to ensure sufficient data for validation and testing.

5. Conclusion

A joint spectroscopy learning method for long-range vibrations including dissociation behavior is proposed, which allows us to obtain reliable full vibrational spectra that are notoriously difficult to achieve. The reliability of the method is guaranteed in two ways. 1) using quantum models to provide flexible parameter form to cover any subtle physical effects in long-range vibrations to solve the under-fitting problem. 2) several carefully selected evidence such as low-lying energy levels, dissociation energy and heat capacity together with machine learning strategy are used to solve the over-fitting problem. The systems of CO and Br_2 in the ground state are investigated by the method. It manages to reconstruct the entire vibrational spectrum including dissociation energy from several low-lying energy levels and heat capacity. The joint method proposed here can make good use of existing research findings to predict the ones in long-range vibrations that may not be available. Considering the widespread application of parametric form in physics, many fields may benefit from this study.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.saa.2020.118363.

CRediT authorship contribution statement

Jia Fu: Conceptualization, Methodology, Software, Writing - original draft, Funding acquisition, Project administration. **ShanShan Long:** Software, Data curation, Visualization, Investigation, Formal analysis. **Jun Jian:** Software, Visualization. **Zhixiang Fan:** Writing - review & editing, Software, Validation. **Qunchao Fan:** Writing - review & editing, Supervision, Resources. **Feng Xie:** Writing - review & editing, Resources. **Yi Zhang:** Writing - review & editing, Resources. **Jie Ma:** Writing - review & editing, Resources.

Acknowledgments

This research is supported by the ministry of education “Chunhui Plan” (Grant No: Z2016160), National Natural Science Foundation of China (Grant No. 11904295, 61722507), the Sichuan Education Department project (Grant No.17ZA0369), the Fund for Sichuan Distinguished Scientists of China (Grant No. 2019JDJQ0050), the Open Foundation of Key Laboratory of Advanced Reactor Engineering and Safety (Grant No. ares-2019-01), the State Key Laboratory Open Fund of Quantum Optics and Quantum Optics Devices, Laser Spectroscopy Laboratory (Grant No. KF201811).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] C. Puzzarini, M.P. de Lara-Castells, R. Tarroni, P. Palmieri, J. Domaison, Accurate ab initio prediction of the rovibrational energy levels and equilibrium geometry of carbonyl selenide (OCSe), Phys. Chem. Chem. Phys. 1 (17) (1999) 3955–3960.
[2] A.M. Kelley, Condensed-Phase Molecular Spectroscopy and Photophysics, John Wiley & Sons, Inc, Hoboken, NJ, USA, 2012.
[3] A.G.G.M. Tielens, The molecular universe, Rev. Mod. Phys. 85 (3) (2013) 1021–1081.
[4] T. Tanabe (Ed.), Tritium: Fuel of Fusion Reactors, Springer, Japan, 2017.
[5] P.-P. Zhang, Z.-X. Zhong, Z.-C. Yan, T.-Y. Shi, Precision spectroscopy of the hydrogen molecular ion D_2^+, Phys. Rev. A 93 (3) (2016) 032507.
[6] M. S. Safronova, D. Budker, D. DeMille, D. F. J. Kimball, A. Derevianko, C. W. Clark, Search for new physics with atoms and molecules, Rev. Mod. Phys. 90 (2).
[7] N. Picqué, Molecular constants mostly from infrared spectroscopy: Non-linear triatomic molecules, part 1: H_2O (HOH), part ε: $HD^{16}O$ ($H^{16}OD$), $HT^{16}O$ ($H^{16}OT$), $HD^{17}O$ ($H^{17}OD$), $HD^{18}O$ ($H^{18}OD$), $DT^{16}O$ ($D^{16}OT$), Molecules and Radicals, Springer-Verlag, Berlin Heidelberg, 2016.
[8] D. Christen, Molecular constants mostly from microwave, molecular beam, and sub-Doppler laser spectroscopy: Paramagnetic diatomic molecules (radicals), Part 1, Molecules and Radicals, Springer-Verlag, Berlin Heidelberg, 2017.
[9] K.M. Jones, E. Tiesinga, P.D. Lett, P.S. Julienne, Ultracold photoassociation spectroscopy: long-range molecules and atomic scattering, Rev. Mod. Phys. 78 (2) (2006) 483–535.
[10] H. Yang, D.-C. Zhang, L. Liu, Y.-X. Liu, J. Nan, B. Zhao, J.-W. Pan, Observation of magnetically tunable Feshbach resonances in ultracold ^{23} Na ^{40} K + ^{40} K collisions, Science 363 (6424) (2019) 261–264.
[11] W. Liu, J. Wu, J. Ma, P. Li, V. B. Sovkov, L. Xiao, S. Jia, Observation and analysis of the hyperfine structure of near-dissociation levels of the $c^3\Sigma^+$ state below the dissociation limit $3S_{1/2} + 6P_{3/2}$, Phys. Rev. A 94 (3).
[12] K.M. Jones, S. Maleki, S. Bize, P.D. Lett, C.J. Williams, H. Richling, H. Knöckel, E. Tiemann, H. Wang, P.L. Gould, W.C. Stwalley, Direct measurement of the ground-state dissociation energy of Na_2, Phys. Rev. A 54 (2) (1996) R1006–R1009.
[13] B. Ji, C. Tsai, L. Li, T. Whang, A.M. Lyyra, H. Wang, J.T. Bahns, W.C. Stwalley, R.J. LeRoy, Determination of the long-range potential and dissociation energy of the $1^3\Delta_g$ state of Na_2, J. Chem. Phys. 103 (17) (1995) 7240–7254.
[14] A. Koleżyński, M. Król (Eds.), Molecular Spectroscopy—Experiment and Theory: From Molecules to Functional Materials, Challenges and Advances in Computational Chemistry and Physics, Springer International Publishing, 2019.
[15] M. Casida, M. Huix-Rotllant, Progress in time-dependent density-functional theory, Annu. Rev. Phys. Chem. 63 (1) (2012) 287–323.
[16] A. Zaitsevskii, N. S. Mosyagin, A. V. Stolyarov, E. Eliav, Approximate relativistic coupled-cluster calculations on heavy alkali-metal diatomics: application to the spin-orbit-coupled $a^1\Sigma^+$ and $b^3\Pi$ states of $RbCs$ and Cs_2, Phys. Rev. A 96 (2).
[17] J. Tennyson, L. Lodi, L.K. McKemmish, S.N. Yurchenko, The ab-initio calculation of spectra of open shell diatomic molecules, J. Phys. B Atomic Mol. Phys. 49 (10) (2016), 102001. .
[18] M. Musiał, P. Skupin, A. Motyl, Chapter thirteen - potential energy curves of NaK molecule from all-electron multireference-coupled cluster calculations, in: P.E. Hoggan, T. Ozdogan (Eds.), Advances in Quantum Chemistry, Vol. 73 of Electron Correlation in Molecules Â€″ Ab Initio beyond Gaussian Quantum Chemistry, Academic Press 2016, pp. 249–262.
[19] T.M. Cover, J.A. Thomas, Elements of Information Theory, 2nd edition N.J, Wiley-Interscience, Hoboken, 2006.
[20] Y. Zhang, W. Sun, J. Fu, Q. Fan, J. Ma, L. Xiao, S. Jia, H. Feng, H. Li, A variational algebraic method used to study the full vibrational spectra and dissociation energies of some specific diatomic systems, Spectrochim. Acta Part A: Mol. Biomol. Spectrosc. 117 (2014) 442–448.
[21] Y. Zhang, W. Sun, J. Fu, Q. Fan, J. Ma, L. Xiao, S. Jia, Method for studying diatomic rovibrational spectra at a given vibrational state, Science China Physics, Mechanics & Astronomy 62 (4) (2019), 943011. .
[22] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016.
[23] D. Wu, L. Wang, P. Zhang, Solving statistical mechanics using variational autoregressive networks, Phy. Rev. Lett. 122 (8) (2019), 080602. .
[24] Y. Levine, O. Sharir, N. Cohen, A. Shashua, Quantum entanglement in deep learning architectures, Phy. Rev. Lett. 122 (6).
[25] K. Mills, M. Spanner, I. Tamblyn, Deep learning and the schrodinger equation, Phys. Rev. A 96 (4) (2017) 042113.
[26] W. Sun, S. Hou, H. Feng, W. Ren, Studies on the vibrational and rovibrational energies and vibrational force constants of diatomic molecular states using algebraic and variational methods, J. Mol. Spectrosc. 215 (1) (2002) 93–105.
[27] G. Herzberg, Molecular Spectra and Molecular Structure, vol I, Reitell Press, 2008.
[28] J.L. Dunham, The energy levels of a rotating vibrator, Phys. Rev. 41 (6) (1932) 721–731.
[29] J. Fu, Q. Fan, G. Liu, H. Li, Y. Xu, Z. Fan, Y. Zhang, Influence of different micro-vibrational behavior on the thermodynamic properties of SO gas, Comput. Theor. Chem. 1115 (2017) 136–143.
[30] M. Volkenstein, The Structure and Physical Properties of Molecules, Science Press, 1960.

[31] R. Kępa, M. Ostrowska-Kopeć, I. Piotrowska, M. Zachwieja, R. Hakalla, W. Szajna, P. Kolek, Ångström ($B^1{}^+A^1$) 0–1 and 1–1 bands in isotopic CO molecules: further investigations, J. Phys. B Atomic Mol. Phys. 47 (4) (2014), 045101. .
[32] J.A. Coxon, P.G. Hajigeorgiou, Born–Oppenheimer breakdown in the ground state of carbon monoxide: a direct reduction of spectroscopic line positions to analytical radial Hamiltonian operators, Can. J. Phys. 70 (1) (1992) 40–54.
[33] M. W. Chase, Nist-janaf thermochemical tables, J. Phys. Chem. Ref. Data 1.
[34] H.-Y. Xu, Y. Liu, Z.-Y. Li, Y.-J. Yang, Y. Bing, Rovibrational spectrum calculations of four electronic states in carbon monoxide molecule: comparison of two effect correction methods, Acta Phys. Sin. 67 (21) (2018), 213301. .
[35] C. Focsa, H. Li, P.F. Bernath, Characterization of the ground state of br2 by laser-induced fluorescence fourier transform spectroscopy of the $B^3\Pi_{0+u} - X^1\Sigma_g^+$ system, J. Mol. Spectrosc. 200 (1) (2000) 104–119.
[36] S. Chattopadhyay, U.S. Mahapatra, R.K. Chaudhuri, Dissociation of homonuclear diatomic halogens via multireference coupled cluster calculations, Mol. Phys. 112 (20) (2014) 2720–2736.
[37] L. Visscher, K.G. Dyall, Relativistic and correlation effects on molecular properties. I. the dihalogens F $_2$, Cl $_2$, Br $_2$, I $_2$, and At $_2$, J. Chem. Phys. 104 (22) (1996) 9040–9046.