Contents lists available at ScienceDirect

Fundamental Research

journal homepage: http://www.keaipublishing.com/en/journals/fundamental-research/

Article

A machine learning vibrational spectroscopy protocol for spectrum prediction and spectrum-based structure recognition

Hao Ren[a], Hao Li[b], Qian Zhang[b], Lijun Liang[c], Wenyue Guo[a], Fang Huang[b], Yi Luo[d,*], Jun Jiang[d,*]

[a] School of Materials Science and Engineering, China University of Petroleum (East China), Qingdao, Shandong 266580, China
[b] Center for Bioengineering and Biotechnology, China University of Petroleum (East China), Qingdao, Shandong 266580, China
[c] College of Automation, Hangzhou Dianzi University, Hangzhou, Zhejiang 310018, China
[d] Hefei National Laboratory for Physical Sciences at the Microscale, CAS Center for Excellence in Nanoscience, School of Chemistry and Materials Science, University of Science and Technology of China, Hefei, Anhui 230026, China

ARTICLE INFO

Article history:
Received 18 March 2021
Received in revised form 7 May 2021
Accepted 25 May 2021
Available online 27 June 2021

Keywords:
Vibrational spectroscopy
Machine learning
Spectrum analysis
Neural networks
Symmetry function

ABSTRACT

Vibrational spectroscopy is one of the most commonly applied techniques for determining molecular structures. Conventional applications often involve extensive expertise or expensive first principles computational effort in order to establish one to one spectrum structure relationships. Here we developed a machine learning protocol to correlate spectral fingerprints with local molecular structures. Our protocol enables not only quick and accurate prediction of infrared (IR) absorption and Raman vibrational spectra based on molecular structures, but more importantly, also enables structure recognition of chemical groups from vibrational spectral features. IR and Raman spectral features arising from different selection rules were recurrently fed to the model to achieve a nearly zero error rate in structure recognition. Both the spectrum prediction and structure recognition models have good transferability, implying a high possibility of being extended to various spectral or non spectral characteristics. This machine learning protocol may provide impovements to real-time field applications in many areas of spectroscopy.

1. Introduction

Spectroscopy is a major tool for revealing the microscopic structures and dynamics of many physical and chemical processes. It is extensively applied in science, industry as well as daily life, with an ever-growing demand for fast and precise spectrum prediction and spectrum-based structure recognition [1,2]. Vibrational spectroscopy is particularly useful owing to its multiple detection methods [3–5] and a high sensitivity to the local environment enabling high spatial and temporal resolution of structural details [6–9]. Prediction and interpretation of vibrational spectra commonly adopt a trial-and-error method to correlate an experimental spectrum with a (series of) probable microstructure(s). This conventional methodology requires deep chemical intuition and broad experience to “guess” the most probable microstructures, followed by expensive calculation, mostly at the first-principles level [1], of vibrational spectra of the “guessed” microstructures.

Machine learning (ML) has greatly advanced the application of quantum and chemical physics [10,11] for determining electronic structures [12–14], energetics [15–19], reaction activities [20–22], drug discovery [23] and materials design [24–27]. Extensive ML efforts have been reported for correlating spectral features with molecular structures. For example, photoionization [28,29], X-ray absorption (XAS) [30,31], UV-visible [32], infrared (IR) [33–35], and nuclear magnetic resonance (NMR) spectra [36] can now be predicted by ML models based on either detailed geometrical structure (i.e., 3D coordinates) or abstract structural descriptors only containing atomic connectivities. The latter approach [37], however, is not capable of capturing fine spectroscopic variations due to subtle differences in local environments, thereafter limiting their practical applications. Recently, Yao et al. used ML-accelerated molecular dynamics to simulate IR spectra of water clusters and small molecules, with results comparable to density functional theory (DFT) results [33]. Kananenka et al. constructed ML models to predict IR signals of the hydroxyl stretch in condensed phase water clusters [34]. Either of these ML models requires only a marginal computational cost to obtain results at the same level of accuracy as DFT calculations. It has been demonstrated that the accuracy of ML models can reach the same level as the data set on which they are trained; the accuracy can be further improved based on datasets of higher accuracy by transfer-learning or Δ-machine learning techniques [38].

* Corresponding authors.
E-mail addresses: yiluo@ustc.edu.cn (Y. Luo), jiangj1@ustc.edu.cn (J. Jiang).

https://doi.org/10.1016/j.fmre.2021.05.005

Compared to the process of spectrum prediction from known structures, spectrum-based structure recognition is more important and of greater practical interest. In principle, one can straightforwardly calculate spectral responses from known structures, while the reverse process is in general more difficult. Efforts to construct ML models to extract structural features from IR or NMR spectra of small collections of molecules can be traced back to the 1990s [39]. The spectrum-structure correlations then obtainable were not very plausible, because each functional group examined was based on only a mere dozens of small molecules. The "shallow" neural networks (NNs) used in previous work likewise had restricted ability to incorporate multiple spectral data for analysis. The recent development of NN algorithms, especially those with deep and complex hidden layers, has brought opportunities to establish the complex correlations inherent in multiple spectroscopic fields from large and reliable datasets. For example, the DP4-AI system developed very recently aids in the interpretation of NMR spectra based on characteristic chemical shifts [40].

Developing ML techniques for vibrational spectroscopy is actually challenging. Many key vibrational spectral signals arise collectively from displacements of several atoms around an equilibrium configuration (i.e. vibrational modes or phonons). In addition, delocalized molecular structures, e.g., conjugated functional groups, and long-range chemical connectivities/interactions are involved in vibrational spectra. Moreover, spectral intensities are determined by quantum selection rules, leading to different spectral features of the same mode when measured by different techniques. These challenges, on the other hand, give us advantageous and complementary means for structure recognition. For example, IR and Raman spectra are often used complementarily for improving structure characterization.

Hydroxyl (–OH) and carbonyl (C=O) groups widely exist in important molecules (Supplementary Material Fig. S1). They participate in many key reactions in synthesis, life processes, and industrial catalysis. Their characteristic stretching modes lying in the spectral ranges of 3000–4000 and 1400–2000 cm^{-1}, respectively, depend strongly on the local environment in which they reside. The appearance (increase) or disappearance (decrease) of corresponding spectral features signal the formation or destruction of these bonds in related chemical reactions. Fast and precise prediction and recognition of these groups are hence of pivotal importance.

In this work, we constructed several NN models to predict stretching frequencies and IR/Raman intensities of –OH and C=O in a dataset containing more than twenty-one thousand molecules. Including the effort devoted to model training, ML prediction is nearly 1000 times faster than direct first-principles calculations. More importantly, we are able to emulate human-expert-based structure recognition by interrelating chemical information extracted from IR and Raman spectra to recognize the occurrence of these groups in unknown molecules. Our NN models can incorporate other spectral or non-spectral data that contain additional structure-property correlations to improve performance. Both NN protocols for spectrum prediction and for structure recognition exhibited good transferability, demonstrating great potential for application to other spectroscopic measurements and chemical identification.

2. Structure-based spectrum prediction

The molecules involved in this work were taken from the QM9 dataset [41], consisting of up to 9 heavy atoms (C, O, N, F). Since the number of molecules increases exponentially as the number of constituent atoms increases (Fig. S1(a,b)), we used QM8, a subset of QM9, for model training and testing. There were in total 21988 molecules; however, F-containing molecules were excluded due to low occurrence (38 out of 21988). Each of the remaining 21950 (denoted as 22K) molecules were calculated by density functional theory (DFT) at B3LYP/6-31G(2df,p) level using Gaussian 09 [42]. Another two groups, each with 3000 molecules, with 9 or 10 heavy atoms were randomly selected from the QM9 and QM10 datasets, respectively, for transferability study.

Among the 22K molecules, 5703 –OH-containing and 6788 C=O-containing molecules were employed to construct NN models for predicting the corresponding stretching vibrational signatures. As shown in Fig. 1, the local chemical environment of the hydroxyl (or carbonyl) group was encoded with atomic centered symmetry functions (SFs) [43,35]. All the SFs were restricted by cutoff functions given by

$$f_c\left(R_{ij}\right) = \begin{cases} 0.5 \cdot \left[\cos\left(\pi R_{ij}\right)\right] & \text{for}\, R_{ij} \le R_c \\ 0 & \text{for}\, R_{ij} > R_c \end{cases}, \tag{1}$$

where R_{ij} is the distance between the i-th and j-th atoms, and R_c is the cutoff radius. Two sets of SFs, radial symmetry functions (RSFs, Fig. 1(b)) and angular symmetry functions (ASFs, Fig. 1(c)), respectively, were used to describe the two-body and three-body interactions between atoms. The RSFs for the neighboring element X (C, H, O, or N) were written as a sum of Gaussians restricted by f_c [35]:

$$G_i^{\gamma}, X = \sum_{j \neq i} e^{-\eta\left(R_{ij}-R_s\right)^2} f_c\left(R_{ij}\right), \tag{2}$$

where the widths of the Gaussians were controlled by η. Each Gaussian collects the two-body interactions arising from a spherical shell controlled by R_s, and we used 26 RSFs for each of the elements C, H, O and N. The ASF has the form [35]

$$G_i^{a} = 2^{1-\xi}\left[\left(1+\lambda\cos\theta\right)_{ijk} \cdot e^{-\eta\left(R_{ij}^2+R_{ik}^2\right)} \cdot f_c\left(R_{ij}\right) \cdot f_c\left(R_{ik}\right)\right]^{\xi} \tag{3}$$

with the angles $\theta_{ijk} = cos^{-1}(R_{ij} \cdot R_{ik}/R_{ij}R_{ik})$ centered at the ith atom. For each of the elements C, H, O and N, 6 ASFs were used to collect the three-body interactions. In the end, a total of 128 (26 × 4 RSFs + 6 × 4 ASFs) SFs were used to build up the structural descriptors for each hydroxyl (carbonyl) group. Detailed values of the parameters η and cutoff radii can be found in Lists S1 and S2 in the Supplemental Materials.

We built separate NN models for each of two spectral properties, i.e. vibrational frequency and IR/Raman intensities. All the spectrum prediction models adopt the same architecture, namely a fully connected feedforward neural network with three hidden layers, configured with 256, 128 or 64 nodes. These models were directly fed with 128 structural descriptors, and use the corresponding spectral properties as regression targets. Details about model initialization and training can be found in Section 4 of the Supplemental Materials.

In Fig. 2, we compare NN-predicted frequencies and IR/Raman intensities with the corresponding DFT results for –OH stretches. Two measures of prediction accuracy, mean absolute error (MAE) and mean relative error (MRE), were used. MAE is suitable for quantifying the prediction accuracy of quantities with values lying within a certain range, e.g. frequencies of a specific mode; while MRE is more suitable for values that vary significantly, e.g. intensities, or frequencies of different modes. As shown in the right panel of Fig. 2(a), –OH stretching frequencies of most molecules occur in the 3650–3850 cm^{-1} region, and the distributions of the frequencies predicted by NN and by DFT agree well. The middle panel of Fig. 2 shows the correlation between NN and DFT frequencies, with a Pearson correlation coefficient (PCC) of $r = 0.99$ and a MRE of 0.09%. Note that the MAE of the NN frequencies is only 3.4 cm^{-1}, which is comparable with many first-principles methods, and is much lower than the root mean square error of approximately 30–40 cm^{-1} between DFT and experiments [44]. This result thus demonstrates that the NN model indeed captured the structure-spectrum correlations with DFT level accuracy, suggesting it may provide a promising way of revealing correlations from reliable experimental data.

Fig. 2(b) and (c) depict NN-predicted IR and Raman intensities of –OH stretches. We observe that most molecules have relatively low IR intensities. Since IR intensity is proportional to the squared derivative of the dipole moment with respect to the normal mode, a low IR intensity implies a small variation in dipole moment caused by this vibration. The MRE of the NN-predicted IR intensities is 11%. Unlike the

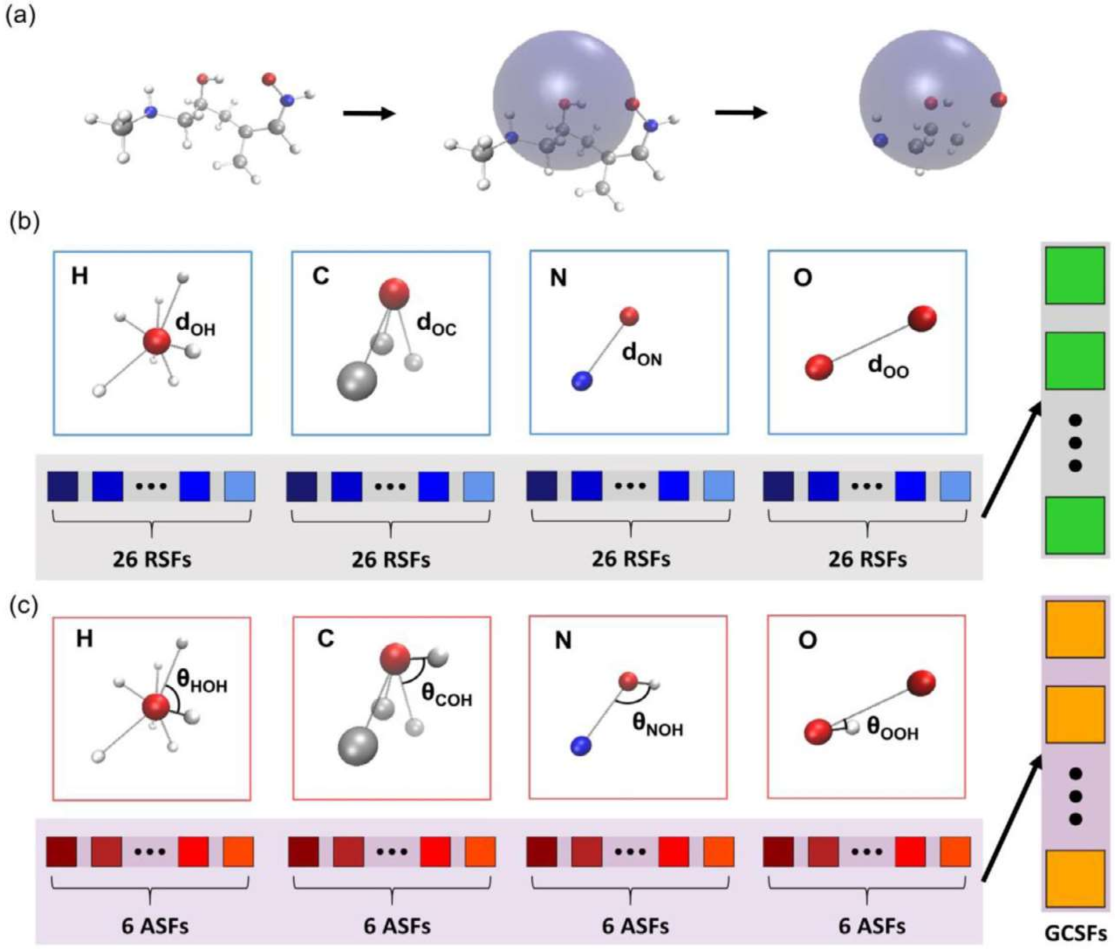

Fig. 1. Schematic of the symmetry function descriptor for encoding the local chemical environment into feature vectors. (a) An appropriate cutoff radius for the hydroxyl group is selected; only atoms within the cutoff sphere are considered. (b) Radial symmetry functions (RSF) centered on the hydroxyl oxygen atom. 26 RSFs are used to encode the contribution of atoms of each element H, C, N, and O. There are 104 (26 × 4) elements in total arising from the radial part. (c) Similar to (b), but for angular symmetry functions (ASFs), with 6 ASFs for each element.

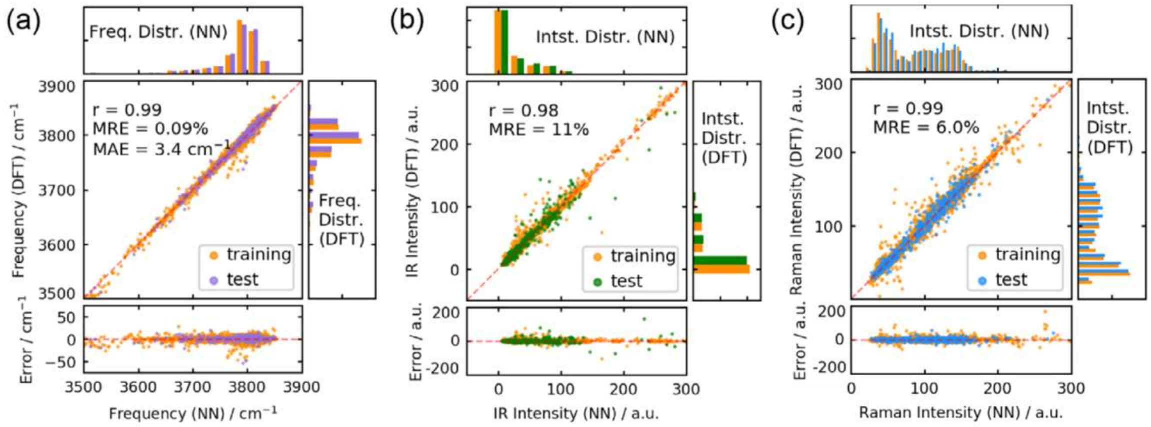

Fig. 2. Correlation plots of (a) NN and DFT frequencies, (b) IR, and (c) Raman intensities for –OH stretching vibrations. Top and right panels depict distributions, while bottom panel shows errors.

IR case, Raman intensities vary widely, making NN prediction easier and resulting in better PCC (r = 0.99) and MRE (6.0%) values. Similar prediction accuracies were also achieved for C=O stretching vibrations, as shown in Fig. S4. Here prediction of C=O stretching frequencies is better than for –OH, possibly due to the much heavier reduced mass of C=O stretches (6.3 AMU) compared to –OH stretches (0.94 AMU): perturbing C=O stretching vibrations requires stronger interactions, which can more easily be captured by the model. The prediction accuracies of IR and Raman intensities were largely affected by the distribution of the data: the more even the distribution and the wider its range (Figs. 2(c), and S4(c)), the more accurate the prediction.

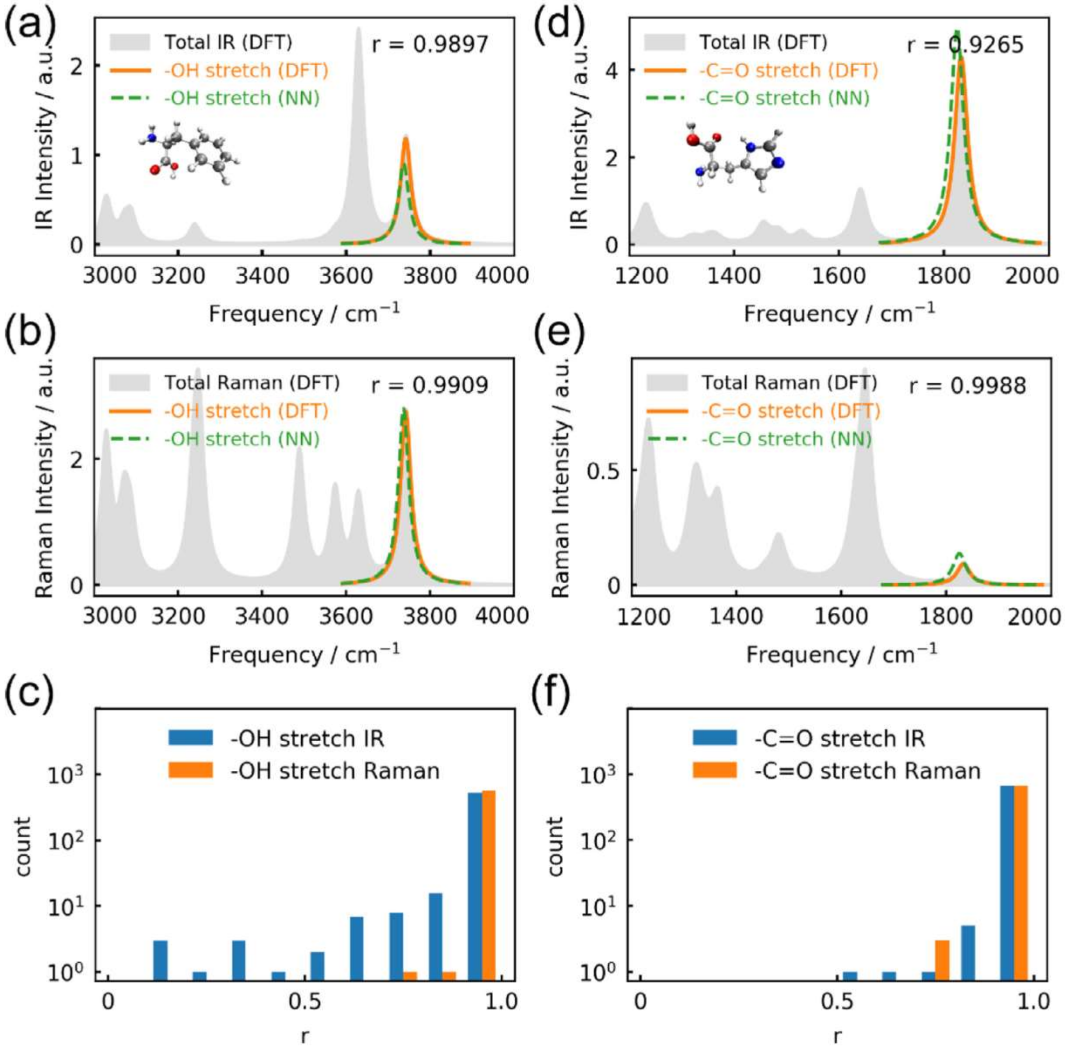

Fig 3. Comparisons between NN predicted (green dashed) and DFT calculated (orange solid) (a) infrared absorption and (b) Raman spectra of histidine –OH stretching vibrations. Total spectra are presented as gray shading. The displayed *r* values are Pearson's correlation coefficients between NN and DFT spectral lines. (c) The distribution of Pearson's correlation coefficients for the –OH stretching vibrations in molecules of the QM8 set. Similar comparisons were performed for the phenylalanine C=O stretching vibration [(d) and (e)]. (f) Distribution of Pearson's correlation coefficients for C=O stretching vibrations of molecules in the QM8 set. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

To measure the spectrum prediction accuracy, we convoluted the NN and DFT frequency-intensity pairs in the same frequency range, wherein the spectra were represented by intensity sequences. The prediction accuracy was then evaluated as the similarity between NN and DFT sequences of intensities. A Lorentzian lineshape with the full-width-at-half-maximum (FWHM) of 30 cm^{-1} was used in the convolution and the similarities between NN and DFT results were evaluated as the PCCs. Fig. 3(a) depicts the NN and DFT IR spectra of the –OH stretching vibration in histidine (containing 12 heavy atoms, thus not included in the QM8 set). The full spectrum is shaded gray. The NN-predicted spectrum agrees well with the DFT result, with a PCC of $r = 0.990$. The Raman spectra of the same mode were compared as shown in Fig. 3(b), producing a similar high consistency of $r = 0.991$. Fig. 3(c) depicts the distribution of PCCs of –OH stretching vibrations in the test set. The PCC reached 0.9 or higher in most entries, i.e. 95.9% (548/571) in IR spectra and 99.6% (568/571) in Raman. Similar comparison results are shown in Fig. 3(d) and (e) for NN predictions of IR and Raman spectra of the C=O stretching vibration in phenylalanine (11 heavy atoms). The PCCs for IR and Raman comparison are 0.927 and 0.999, respectively, indicating that our NN models perform well for larger molecules not in the QM8 set. The distribution of PCCs for C=O stretching vibrations is shown in Fig. 3(f), where more than 98.8% (671/679, IR) and 99.5% (675/679, Raman) of the test set show PCCs higher than 0.9. It has been found that the choice of the FWHM parameter has a limited effect on the comparison results, as shown in Fig. S5.

Table 1

Optimal cutoff radii R_c [Å] of the NN models for hydroxyl and carbonyl stretching vibrations.

	Frequency	IR intensity	Raman intensity
–OH	4.2	5.4	6.2
C=O	4.0	5.0	6.0

Table 2

Overall –OH and C=O recognition accuracies and error rates (in parentheses) using IR, Raman, and combined (IR+Raman) spectra.

	IR	Raman	IR+Raman
–OH	98.50% (1.50%)	98.58% (1.42%)	99.36% (0.64%)
C=O	98.04% (1.96%)	98.49% (1.51%)	98.50% (1.50%)

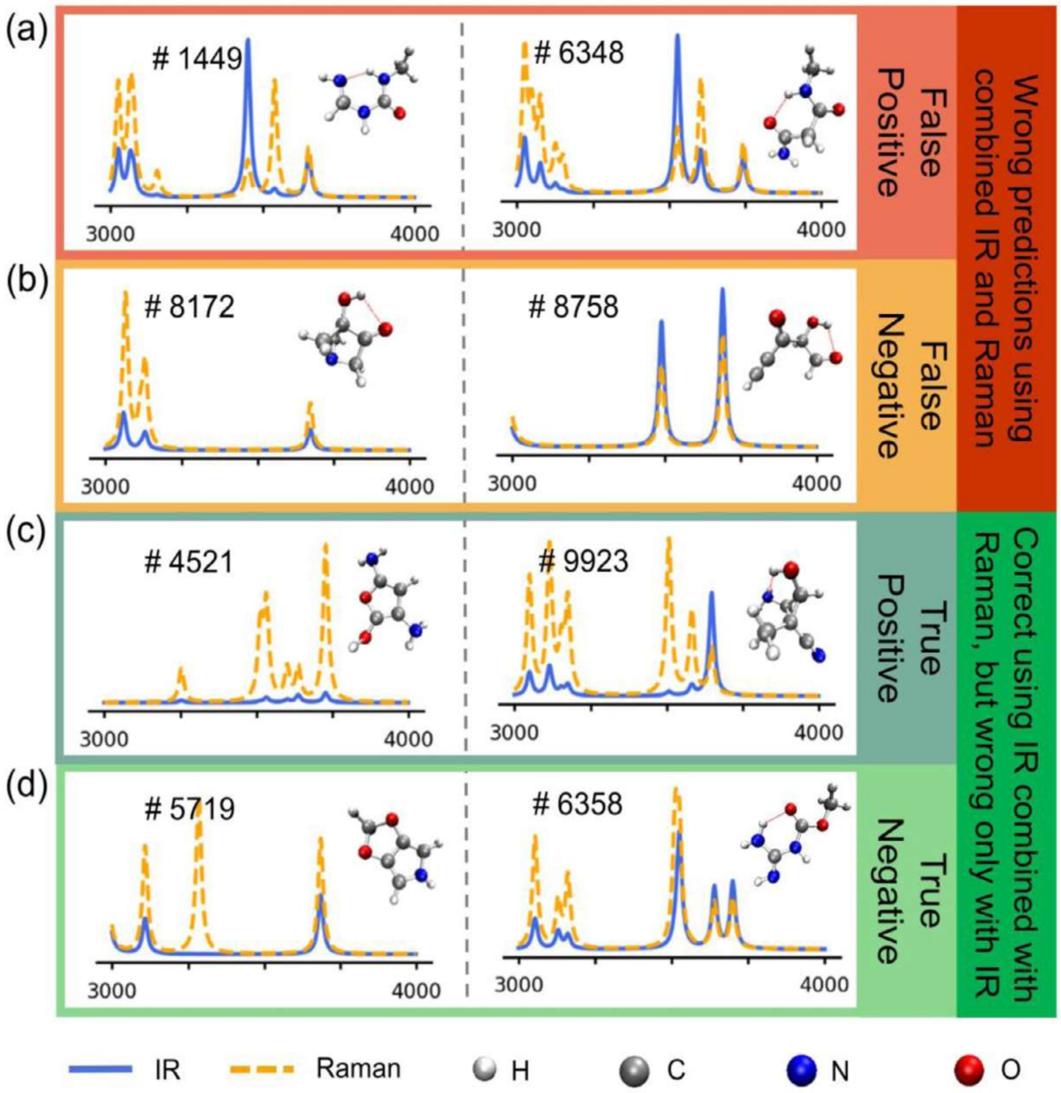

Fig. 4. Structures and vibrational spectra of two randomly selected molecules that resulted in (a) false positive (FP), (b) false negative (FN), (c) true positive (TP), and (d) true negative (TN) recognition.

While SF descriptors are chosen to represent molecular structures, the cutoff radii R_c restrict the extent of these descriptors and indicate how much of the neighborhood around a given group interactions were considered. Choosing too small an R_c results in inadequate information, while choosing one too large would include too many irrelevant structural details. Providing too much information not only increases computational demands, but also introduces the risk of over-fitting. As estimated by prediction accuracy, an optimal R_c exists for each quantity and provides a measure of locality, as shown in Figs. S3 and S4. Interestingly, the optimal R_c values for frequency prediction, as listed in Table 1, are shorter than those for IR intensity prediction, whereas Raman intensity prediction requires longer ones. This may be ascribed to the physical nature of the three quantities. In the harmonic approximation, the frequency of a bond stretch exhibits second order variation in energy with respect to structure deformation. It is hardly affected by environmental variations in distant regions. Relatively short R_c values are thus expected to be optimal for frequency prediction. On the other hand, IR intensity relates to the charge redistribution versus bond stretching, which is a delocalized property, requiring a larger R_c. Raman intensity relates to the second-order derivative of the molecular dipole, consequently, it needs the largest R_c.

It required approximately 80 min to train a prediction NN model for one property of the QM8 set on a GTX 1080 Ti GPU. Once properly trained, a model only requires several milliseconds to predict a property for a batch of 7000 molecules. In comparison, full first-principles calculations for 7000 molecules consume 1.4×10^5 min (Section 5 of Supplementary Materials). In other words, our ML protocol is 1000 times faster than quantum mechanical methods at a similar level of precision.

3. Spectrum based structure recognition

We now focus on recognizing –OH and C=O in a molecule based on vibrational spectra. IR and Raman spectra are complementary for structure characterization due to their different selection rules. To interrelate the information extracted from these spectra, we chose the long short-term memory (LSTM) network. A LSTM is capable of accepting data from multiple spectra recurrently, and then combining the chemical knowledge extracted from each sequence to make decisions. The DFT-calculated IR and Raman spectra in the 3000–4000 cm^{-1} region for –OH and 1000–2000 cm^{-1} for C=O were used as inputs. Each spectrum was sampled at 2 cm^{-1} resolution, resulting a 500-dimensional vector of intensities. The model contains two LSTM layers. The first LSTM layer takes the IR spectra, the second takes the result of the first together with the Raman spectra, and feeds its output to a fully-connected layer. Lastly, a softmax layer makes the final decision, *Yes* or *No*, as to whether the group exists or not in the molecule. As shown in Fig. S6(b), the –OH recognition model achieved an accuracy of 99.36%; only 14 out of the 2194 test molecules were incorrectly recognized, with 12 false negatives (FN) and 2 false positives (FP). Structure recognition accuracy of C=O, shown in Fig. S6(d), is 98.50%, lower than that of –OH. This can be ascribed to the more complex IR and Raman spectra in the range 1000–2000 cm^{-1}: many other modes, e.g., C=C, N=O and $–NO_2$ stretches, water bending, etc. occur in this range. In contrast, only C–H, O–H, and N–H stretching modes lie sparsely in the 3000–4000 cm^{-1} region.

Table 2 lists the overall recognition accuracies of our LSTM models when trained using either IR, Raman, or combined (IR+Raman) spectra. For –OH recognition, individual IR- or Raman-trained models have error rates of 1.50% and 1.42%, respectively; the error is significantly improved to 0.64% when trained using both. Similar behavior was observed for C=O. Since LSTM is a generic procedure capable of accepting multiple inputs simultaneously, we would expect further improvement by feeding it more features such as NMR or XAS.

Fig. 4(a) depicts two FPs in which N–H stretch signals have been incorrectly identified as being O–H. Tables S1 and S2 in Supplementary Materials list all –OH and C=O recognition errors. Normal O–H and N–H stretching modes lie approximately at 3600 and 3400 cm−1, respectively. In the two FP errors, the N–H stretching mode is blue-shifted, i.e., there are stronger interactions between N and H because of the more electronegative local environment around N in these two molecules. However, blue-shifted N–H stretches do not always generate errors. In Fig. 4(d), the model incorrectly predicted that a –OH exists when only IR spectral data were provided; this assignment was corrected when Raman signals were included. Interestingly, the correct assignment based on combined (IR+Raman) spectra shown in Fig. 4(d) are more typical than those shown in (a). Molecule #5719 is heterocyclic, #6358 has a peptide bond, while the falsely recognized #1449 has an unsaturated terminal *N* and is not stable. Further analysis shows that another cause of recognition errors is presence of an intramolecular hydrogen bond (HB) between a –OH group and a nearby carbonyl oxygen, as shown in Fig. 4(b). However, not all molecules with HBs are incorrectly recognized. For example, #9923 is incorrectly recognized when only using IR, but corrected by adding Raman information. However, for molecules #8172 and #8758, as shown in Fig. 4(b), even combined spectra are still insufficient to produce correct recognition, possibly due to the lack of distinctive IR and Raman lineshapes in those moleucles. We conclude that IR+Raman spectra are able to distinguish most but not all –OH groups, with notable exceptions being blue-shifted NH stretches and molecules with direct HBs. However, incorporating additional kinds of spectra-such as ^{1}H NMR or N/O XAS-that can provide additional selection rules, is expected to eliminate these ambiguities.

The reported recognition results were achieved using the QM8 dataset. Since vibrational spectra depend most strongly on the local environment, spectrum-structure correlations in small molecules are expected to be transferable to larger molecules. We tested the transferability of the models (trained with the QM8 dataset) to two groups of 3000 larger molecules randomly selected from the QM9 and QM10 datasets, respectively. As shown in Fig. S7, the respective recognition accuracies for –OH and C=O are 98.97 and 97.47%, for QM9 molecules; and 96.07 and 93.27% for QM10 molecules. The decrease in performance on larger molecules is expected since larger size introduces higher complexity, which might not be sufficiently described by models developed with smaller molecules. The transferability is improvable by incorporating more complex local structural patterns in the training stage, e.g. sequentially augmenting more data with an online machine learning technique.

4. Conclusions

In summary, we present an initial effort toward using ML to correlate spectroscopic knowledge of a molecule to its structure and local intramolecular environments. Our ML models can predict vibrational spectra of known structures at a computational cost that is three orders of magnitude lower than first-principles calculations while maintaining the same level of accuracy. More importantly, based on IR and Raman spectra, the models can recognize the existence of hydroxyl or carbonyl groups in a molecule at accuracies of 99.4% and 98.5%, respectively. We demonstrate that the recognition accuracy can be significantly improved by interrelating information extracted from both IR and Raman spectroscopies. We expect the accuracy could be further improved by involving other spectroscopic or non-spectroscopic characteristics. The established ML protocols trained with small molecules can be transferred to larger ones, suggesting a bright future for fast and precise spectrum prediction and intelligent structure recognition in various important applications, such as detection of structural variations/fluctuations in chemical reactions, automatic identification of interstellar molecules, or real-time recognition of molecular groups in biomedical diagnosis.

Declaration of Competing Interest

The authors declared that they have no conflict of interest to this work.

Acknowledgments

This work was financially supported by the National Natural Science Foundation of China (21773309, 21776315, 21633006 and 21633007), the National Key Research and Development Program of China (2018YFA0208603, 2017YFA0303500), the Fundamental Research Funds for the Central Universities (19CX05001A), the Anhui Initiative in Quantum Information Technologies (AHY090200), and Hefei National Laboratory for Physical Sciences at Microscale (KF2020004). Part of the numerical calculations have been carried out in the National Supercomputing Center in Shanghai and the Supercomputing Center of University of Science and Technology of China. The implemented hydroxyl and carbonyl recognition functionalities have been deployed on the DCAIKU project. All the data involved in this work were also deposited to DCAIKU (http://dcaiku.com:13000).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.fmre.2021.05.005.

References

[1] U.N. Morzan, D.J. Alonso de Armiño, N.O. Foglia, et al., Spectroscopy in complex environments from QM–MM simulations, Chem. Rev. 118 (2018) 4071–4113.
[2] K.E. Dorfman, F. Schlawin, S. Mukamel, Nonlinear optical signals and spectroscopy with quantum light, Rev. Mod. Phys. 88 (2016) 045008.
[3] B.C. Stipe, M.A. Rezaei, W. Ho, Localization of inelastic tunneling and the determination of atomic-scale structure with chemical specificity, Phys. Rev. Lett. 82 (1999) 1724–1727.
[4] S. Duan, Z. Rinkevicius, G. Tian, et al., Optomagnetic effect induced by magnetized nanocavity plasmon, J. Am. Chem. Soc. 141 (2019) 13795–13798.
[5] G. Reecht, N. Krane, C. Lotze, et al., Vibrational excitation mechanism in tunneling spectroscopy beyond the franck-condon model, Phys. Rev. Lett. 124 (2020) 116804.
[6] Z. Han, G. Czap, C.L. Chiang, et al., Imaging the halogen bond in self-assembled halogenbenzenes on silver, Science 358 (2017) 206–210.
[7] J. Lee, K.T. Crampton, N. Tallarida, et al., Visualizing vibrational normal modes of a single molecule with atomically confined light, Nature 568 (2019) 78–82.
[8] C. Sánchez Muñoz, F. Schlawin, Photon correlation spectroscopy as a witness for quantum coherence, Phys. Rev. Lett. 124 (20) (2020) 203601.
[9] G. Radtke, D. Taverna, N. Menguy, et al., Polarization selectivity in vibrational electron-energy-loss spectroscopy, Phys. Rev. Lett. 123 (2019) 256001.
[10] K.T. Schütt, S. Chmiela, O.A. von Lilienfeld, et al., Machine Learning Meets Quantum Physics, vol. 968 of Lecture Notes In Physics, Springer International Publishing, Cham, 2020.
[11] O.A. von Lilienfeld, K.-R. Müller, A. Tkatchenko, Exploring chemical compound space with quantum-based machine learning, Nat. Rev. Chem. 4 (2020) 347–358.
[12] G. Carleo, M. Troyer, Solving the quantum many-body problem with artificial neural networks, Science 355 (2017) 602–606.
[13] A. Canabarro, S. Brito, R. Chaves, Machine learning nonlocal correlations, Phys. Rev. Lett. 122 (2019) 200401.

[14] M.S. Scheurer, R.J. Slager, Unsupervised machine learning and band topology, Phys. Rev. Lett. 124 (2020) 226401.

[15] M. Rupp, A. Tkatchenko, K.R. Müller, et al., Fast and accurate modeling of molecular atomization energies with machine learning, Phys. Rev. Lett. 108 (2012) 058301.

[16] T. Jacobsen, M. Jørgensen, B. Hammer, On-the-fly machine learning of atomic potential in density functional theory structure optimization, Phys. Rev. Lett. 120 (2018) 026102.

[17] J. Han, L. Zhang, R. Car, et al., Deep potential: a general representation of a many–body potential energy surface, Commun. Comput. Phys. 23 (2018) 629–639.

[18] K. Yao, J.E. Herr, S.N. Brown, et al., Intrinsic bond energies from a bonds-in–molecules neural network, J. Phys. Chem. Lett. 8 (12) (2017) 2689–2694.

[19] S.-D. Huang, C. Shang, P.-L. Kang, et al., Atomic structure of boron resolved using machine learning and global sampling, Chem. Sci. 9 (2018) 8644–8655.

[20] A.F. Zahrt, J.J. Henle, B.T. Rose, et al., Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning, Science 363 (2019) 247.

[21] X. Wang, S. Ye, W. Hu, et al., Electric dipole descriptor for machine learning prediction of catalyst surface–molecular adsorbate interactions, J. Am. Chem. Soc. 142 (2020) 7737–7743.

[22] D.T. Ahneman, J.G. Estrada, S. Lin, et al., Predicting reaction performance in c–n cross-coupling using machine learning, Science 360 (2018) 186–190.

[23] S. Ekins, A.C. Puhl, K.M. Zorn, et al., Exploiting machine learning for end-to-end drug discovery and development, Nat. Mater. 18 (2019) 435–441.

[24] T. Xie, J.C. Grossman, Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties, Phys. Rev. Lett. 120 (2018) 145301.

[25] J.W. Barnett, C.R. Bilchak, Y. Wang, et al., Designing exceptional gas-separation polymer membranes using machine learning, Sci. Adv. 6 (20) (2020) eaaz4301.

[26] Q. Zhou, S. Lu, Y. Wu, et al., Property-oriented material design based on a data–driven machine learning technique, J. Phys. Chem. Lett. (2020) 3920–3927.

[27] L.A. Griffin, I. Gaponenko, S. Zhang, et al., Smart machine learning or discovering meaningful physical and chemical contributions through dimensional stacking, NPJ Comput. Mater. 5 (2019) 85.

[28] S. Kumar Giri, U. Saalmann, J.M. Rost, Purifying electron spectra from noisy pulses with machine learning using synthetic hamilton matrices, Phys. Rev. Lett. 124 (2020) 113201.

[29] K. Ghosh, A. Stuke, M. Todorović, et al., Deep learning spectroscopy: Neural networks for molecular excitation spectra, Adv. Sci. 6 (2019) 1801367.

[30] M.R. Carbone, M. Topsakal, D. Lu, et al., Machine-learning x-ray absorption spectra to quantitative accuracy, Phys. Rev. Lett. 124 (15) (2020) 156401.

[31] C. Zheng, K. Mathew, C. Chen, et al., Automated generation and ensemble-learned matching of x-ray absorption spectra, NPJ Comput. Mater. 4 (2018) 12.

[32] S. Ye, W. Hu, X. Li, et al., A neural network protocol for electronic excitations of *n*-methylacetamide, Proc. Natl. Acad. Sci. USA (2019) 201821044.

[33] K. Yao, J.E. Herr, D. Toth, et al., The tensormol-0.1 model chemistry: a neural network augmented with long-range physics, Chem. Sci. 9 (2018) 2261–2269.

[34] A.A. Kananenka, K. Yao, S.A. Corcelli, et al., Machine learning for vibrational spectroscopic maps, J. Chem. Theory Comput. 15 (2019) 6850–6858.

[35] M. Gastegger, J. Behler, P. Marquetand, Machine learning molecular dynamics for the simulation of infrared spectra, Chem. Sci. 8 (2017) 6924–6935.

[36] F.M. Paruzzo, A. Hofstetter, F. Musil, et al., Chemical shifts in molecular solids by machine learning, Nat. Commun. 9 (1) (2018) 4501.

[37] C. Affolter, J. Clerc, Prediction of infrared spectra from chemical structures of organic compounds using neural networks, Chemom. Intell. Lab. Syst. 21 (2-3) (1993) 151–157.

[38] R. Ramakrishnan, P.O. Dral, M. Rupp, et al., Big data meets quantum chemistry approximations: the δ-machine learning approach, J. Chem. Theory Comput. 11 (5) (2015) 2087–2096.

[39] D. Ricard, C. Cachet, D. Cabrol-Bass, et al., Neural network approach to structural feature recognition from infrared spectra, J. Chem. Inf. Model. 33 (2) (1993) 202–210.

[40] A. Howarth, K. Ermanis, J.M. Goodman, DP4-AI automated NMR data analysis: straight from spectrometer to structure, Chem. Sci. 11 (2020) 453–4539.

[41] R. Ramakrishnan, P.O. Dral, M. Rupp, et al., Quantum chemistry structures and properties of 134 kilo molecules, Sci. Data 1 (2014) 140022.

[42] M.J. Frisch, G.W. Trucks, H.B. Schlegel, et al., Gaussian 09 Revision E.01, Gaussian Inc., 2009 Wallingford CT.

[43] J. Behler, First principles neural network potentials for reactive simulations of large molecular and condensed systems, Angew. Chem. Int. Ed. 56 (42) (2017) 12828–12840.

[44] J.P. Merrick, D. Moran, L. Radom, An evaluation of harmonic vibrational frequency scale factors, J. Phys. Chem. A 111 (2007) 11683–11700.

Hao Ren is an associate professor at China University of Petroleum (East China). He received his B.S. degree in Chemistry in 2003 at Shandong University, a Ph.D. degree in Physical Chemistry under the tutelage of Prof. Jinlong Yang in January 2010 at University of Science and Technology of China, a Ph.D. degree in Biotechnology under the tutelage of Prof. Yi Luo in June 2010 at Royal Institute of Technology, Sweden. From 2010 to 2013, he worked as a post-doc in Prof. Shaul Mukamel's group at University of California, Irvine. His current research mainly focuses on the development of *ab-initio* and machine learning methods for molecular spectroscopy.

Jun Jiang is a professor at Hefei National Laboratory for Physical Sciences at the Microscale, University of Science and Technology of China (USTC). He received a B.S. degree in Theoretical Physics in 2000 at Wuhan University, a Ph.D. degree in Theoretical Chemistry under the tutelage of Prof. Yi Luo in 2007 at Royal Institute of Technology (KTH), Sweden, a Ph.D. degree in Solid State Physics under the tutelage of Prof. Wei Lu in 2008 at Shanghai Institute of Technical Physics, CAS. From 2008 to 2011, he worked as a Post-doc at KTH and University of California Irvine under the tutelage of Prof. Shaul Mukamel. He joined USTC in 2011 as a Professor in Physical Chemistry. His research interests focus on the development and application of multi-scale methods and machine learning techniques in the study of charge kinetics in complex systems. He targets a wide range of physics or chemistry applications such as photocatalysis, biochemistry, photochemistry, molecular electronics and photonics. Dr. Jiang is a recipient of the "National Science Fund for Distinguished Young Scholars in China", and has won the "Young Theoretical Chemistry Investigator Award of Chinese Chemistry Society", "Distinguished Lectureship Award of the Chemical Society of Japan 2020".