Spectrochimica Acta Part A 90 (2012) 55–62

Contents lists available at SciVerse ScienceDirect

Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy

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

Neural network consistent empirical physical formula construction for density functional theory based nonlinear vibrational absorbance and intensity of 6-choloronicotinic acid molecule

Nihat Yildiz[a,*], Mehmet Karabacak[b], Mustafa Kurt[c], Serkan Akkoyun[a]

[a] Dept. of Physics, Cumhuriyet University, 58140 Sivas, Turkey
[b] Dept. of Physics, Afyon Kocatepe University, 03040 Afyonkarahisar, Turkey
[c] Dept. of Physics, Ahi Evran University, 40100 Kırşehir, Turkey

ARTICLE INFO

Article history:
Received 16 December 2011
Received in revised form 3 January 2012
Accepted 8 January 2012

Keywords:
Neural network
6-Choloronicotinic acid
Vibrational absorbance
Vibrational intensity
Molecular structure
Empirical physical formula

ABSTRACT

Being directly related to the electric charge distributions in a molecule, the vibrational spectra intensities are both experimentally and theoretically important physical quantities. However, these intensities are inherently highly nonlinear and of complex pattern. Therefore, in particular for unknown detailed spatial molecular structures, it is difficult to make ab initio intensity calculations to compare with new experimental data. In this respect, we very recently initiated entirely novel layered feedforward neural network (LFNN) approach to construct empirical physical formulas (EPFs) for density functional theory (DFT) vibrational spectra of some molecules. In this paper, as a new and far improved contribution to our novel molecular vibrational spectra LFNN-EPF approach, we constructed LFFN-EPFs for absorbances and intensities of 6-choloronicotinic acid (6-CNA) molecule. The 6-CNA data, borrowed from our previous study, was entirely different and much larger than the vibrational intensity data of our formerly used LFNN-EPF molecules. In line with our *another* previous work which *theoretically* proved the LFNN relevance to EPFs, although the 6-CNA DFT absorbance and intensity were inherently highly nonlinear and sharply fluctuating in character, still the optimally constructed train set LFFN-EPFs very successfully fitted the absorbances and intensities. Moreover, test set (i.e. yet-to-be measured experimental data) LFNN-EPFs consistently and successfully predicted the absorbance and intensity data. This simply means that the physical law embedded in the 6-CNA vibrational data was successfully extracted by the LFNN-EPFs. In conclusion, these vibrational LFNN-EPFs are of explicit form. Therefore, by various suitable operations of mathematical analysis, they can be used to estimate the electronic charge distributions of the unknown molecule of the significant complexity. Additionally, these estimations can be combined with those of theoretical DFT atomic polar tensor calculations to contribute to the identification of the molecule.

1. Introduction

Being directly related to the electric charge distributions in a molecule, the vibrational spectra intensities are both experimentally and theoretically important physical quantities [1]. Following the pioneering initial works [2,3] to interpret a large amount of experimental intensity data, studies were devoted into mainly the formal matters of the interpretation of IR (infrared) intensities [4,5]. After reasonable resolution of the formal aspects, efforts to explain the physical meaning of intensity curves enabled the researchers to correlate some intensity patterns with charge distributions of the molecule [6,7]. Briefly, for the *i*th normal mode the measured vibrational intensities I_i and atomic charges are related by the molecular dipole moment ($\vec{M}$) derivatives with respect to associated normal coordinate Q_i. Because *intensity–atomic charge relationship* is detailed in Section 2.1, here we simply mention that decades of efforts into this relationship resulted in recent accurate prediction of molecular charge distribution experimental data by ab initio IR intensity calculations [8–11].

On the other hand, for very complex structure molecules, both IR and Raman spectral intensities exhibit inherent high nonlinearity and complex pattern. In this case, it is difficult to make ab initio intensity calculations to compare with new experimental data. Additionally, the identity and three dimensional (3D) structure of the chemical compound may be totally unknown prior to the suitable spectral measurements. Indeed, data for 3D structure and IR spectra are available only for 0.5% of all known chemical molecules [12]. Clearly, satisfactory ab initio vibrational intensity calculations cannot be made for the most molecules since their detailed spatial structure are unknown. Therefore, based on the experimental

* Corresponding author. Tel.: +90 346 2191010x1378.
E-mail address: nyildiz@cumhuriyet.edu.tr (N. Yildiz).

1386-1425/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.saa.2012.01.018

data for the molecule of unknown detailed 3D structure, an explicit form of empirical physical formula (EPF) for this highly nonlinear vibrational intensity pattern is of great interest. Indeed, in line with our previous theoretical treatment [13], an appropriate vibrational intensity EPF, as we demonstrated in [14], can be consistently constructed by using a layered feedforward neural network (LFNN) [15]. As we give more details in Section 2.2, the LFNN is a universal nonlinear function approximator [16].

In this paper, as a continuation of our very recently initiated entirely novel LFNN-EPFs approach to vibrational intensities [14], suitable LFNN-EPFs were constructed for density functional theory (DFT) vibrational spectra absorbance and intensities of a different molecule, namely 6-choloronicotinic acid (6-CNA). The 6-CNA absorbance/intensity data was entirely different and much larger than the vibrational intensity data of our formerly used [14] LFNN-EPF molecules. The 6-CNA absorbance and intensity data was borrowed from our previous study [17]. Although our novel LFNN-EPFs approach vibrational intensities was clearly stated in our previous initial work [14], here for completeness we again re-state it after some brief remarks. A number of previous works [18–23] applying neural networks to molecular vibrational intensities exist. In particular, the studies of Gasteiger and coworkers [18–20] for 3D molecule structure derivation from IR spectra are very important. They successfully use radial distribution function for the simulation of IR spectra by suitable neural networks. But, they do not aim to construct an explicit mathematical function for the experimental IR spectra. On the other hand, just as we did for methanesulfonamides in our previous paper [14], in this paper we constructed consistent *explicit form* of LFNN-EPFs for vibrational absorbance and intensity of 6-CNA. Although the 6-CNA DFT vibrational absorbance and intensity were inherently highly nonlinear and sharply fluctuating in character, still train set LFFN-EPFs very successfully fitted these absorbances and intensities. Moreover, test set (i.e. yet-to-be measured experimental data) LFNN-EPFs consistently predicted the absorbance and intensity data. That is, the physical law embedded in the absorbance and intensity data was successfully extracted by the LFNN-EPFs. In conclusion, by various suitable operations of mathematical analysis, this explicit form of vibrational LFNN-EPFs can be used to estimate the electronic charge distributions of the unknown molecule of the significant complexity. Additionally, these estimations can be combined with those of theoretical DFT atomic polar tensor calculations to contribute to the identification of the molecule.

2. Theories

2.1. Atomic polar tensors and vibrational intensities

Atomic polar tensors (APTs) [10] are the fundamental parameters of any IR intensity predictions. As defined in Eq. (1), the elements of APTs are the molecular dipole moment derivatives with respect to Cartesian displacements of atoms ξ_k. In calculating Raman intensities molecular polarizability tensor $\boldsymbol{\alpha}$ is used instead of dipole moments.

$$I_i = C\left[\left(\frac{\partial \vec{M}}{\partial Q_i}\right)^0\right]^2 = C\left[\sum_k \left(\frac{\partial \vec{M}}{\partial \xi_k}\right)^0 L_{ki}\right]^2 \tag{1}$$

where superscript '0' indicates that the derivatives are evaluated at the minimum (equilibrium) geometry. I_i is the absolute IR intensity of ith normal mode, Q_i is the normal coordinate, L_i is the Cartesian eigenvector relative to Q_i, C is a constant whose value depends on the units used. The vector components of $(\partial \vec{M}/\partial \xi_k)^0$ in Eq. (1) can be gathered into 3×3 matrices called APTs. For the atom β, the general uv component of AFT can be expressed as in Eq. (2), where u and v are for x, y, or z components, respectively.

$$P^{uv} = \left(\frac{\partial M_u}{\partial v_\beta}\right)^0 \tag{2}$$

Here the starting point is the basic hypothesis of the equilibrium charges and charge fluxes (ECCF) theory [24], namely the identification of the molecular dipole moment $\vec{M}$ as the sum over point atomic charges q_α of Eq. (3).

$$\vec{M} = \sum_\alpha q_\alpha \vec{r}_\alpha \tag{3}$$

where $\vec{r}_\alpha$ is the position vector of αth atom. Using Eqs. (2) and (3), we obtain Eq. (4).

$$P^\beta_{uv} = q^0_\beta \delta_{uv} + \sum_\alpha \left(\frac{\partial q_\alpha}{\partial v_\beta}\right)^0 u^0_\alpha \tag{4}$$

Eq. (4) is the general equation to calculate the uv component of APT. If the Cartesian coordinate system is suitably rotated, simpler equations for APT components are obtained. As a result, relevant equilibrium charges and charge fluxes can be easily calculated. For instance, let us consider the planar molecules on a CH bond. Using a suitable local Cartesian coordinate system, we obtain Eq. (5) to relate yy APT component to hydrogen equilibrium charge q^0_H.

$$P^H_{yy} = q^0_H \tag{5}$$

Eq. (5) is the fundamental equation which relates the measured IR absolute intensity and molecular atomic charges (see for more [10]).

2.2. Brief LFNN fundamentals and its relevance to vibrational intensity EPF construction

LFNN relevance to general EPFs is analyzed in depth in [13]. Specific LFNN relevance to vibrational intensity EPF construction is mentioned in [14]. Still, here we again give the minimum LFNN fundamentals. We also mention briefly LFNN-EPFs in general and vibrational intensity LFNN-EPFs in particular.

2.2.1. The LFNN fundamentals

An artificial neural network (ANN) [15] resembles the brain and consists of interconnecting artificial neurons which have adaptive synaptic weights. The ANN acquires new knowledge by modifying weights. Being a particular kind of ANN, LFNN is a one input–many intermediate (hidden)–one output layer device, all layers of which are interconnected by adaptable weights (Fig. 1).

It has been rigorously proved that a *single hidden layer LFNN* is sufficient for excellent nonlinear function approximation [16]. However, contrary to what we did in most of our previous LFNN-EPFs papers, in this paper, due to extremely fluctuating nature of training IR and Raman data, we generally had to use *multi-hidden layer* LFNN. Still, here for simplicity and without loss of generality; we only explain the *single hidden layer* LFNN functionality. Borrowing from Ref. [16], for a LFNN with single hidden layer, in Fig. 1, the desired output vector $\vec{y}$ is approximated by a *network multi-output vector* $\vec{f}$ which is defined by Eq. (6)

$$\vec{f} : R^p \to R^r : \vec{f}_k(\vec{x}) = \sum_{j=1}^{h_1} \beta_j G(A_j(\vec{x})), \quad \vec{x} \in R^p, \quad \beta_j \in R, \quad A_j \in A^p,$$
$$\text{and} \quad k = 1, \ldots, r, \tag{6}$$

where A^p is the set of all functions of $R^p \to R$ defined by $A(\vec{x}) = \vec{w} \cdot \vec{x} + b$. '·' is the dot product, $\vec{w}$ is input to hidden layer weight

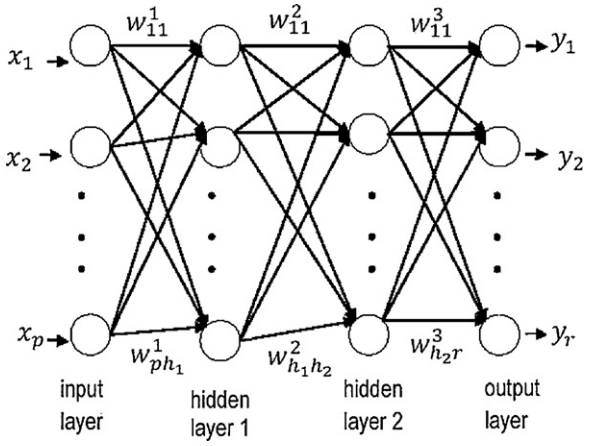

Fig. 1. Fully connected one input–many hidden–one output layer LFNN. Only two hidden layers are shown. x_i $(i = 1, \ldots, p)$ and y_i $(i = 1, \ldots, r)$ are, respectively, input and output vector components. Circles: artificial neurons, arrows: adaptable synaptic weights. w^i_{jk}: weight vector component, where i is a layer index, jk weight component from the jth neuron of ith layer and to kth neuron of $(i+1)$th layer. Hidden layer neurons are, respectively, h_1 and h_2.

vector, $\vec{x}$ is the LFNN input vector in Fig. 1, and b is the bias weight. In Fig. 1, the columns of the weight matrices w^1 and w^2 correspond to weight vectors defined in $A(\vec{x})$ and $\vec{\beta}$ in Eq. (6). However, note that, as is obvious from Fig. 1 and Eq. (6), the correspondences $w^1 \to A(\vec{x})$ and $w^2 \to \vec{\beta}$ are valid *only* for single hidden layer LFNN. For the LFNN with more than single hidden layer, both Eq. (6) and the correspondences must be modified accordingly. *Additionally, in Eq. (6), the hidden neuron activation function $G : R \to R$ is theoretically any well-behaved nonlinear function; proving that a LFNN is a universal nonlinear function approximator.* In applications, G is frequently chosen as a kind of nonlinear sigmoid function defined by Eq. (7),

$$G : R \to [0, 1] \quad \text{or} \quad [-1, 1], \quad \text{non-decreasing}, \quad \lim_{\lambda \to \infty} G(\lambda) = 1,$$
$$\text{and} \quad \lim_{\lambda \to -\infty} G(\lambda) = 0 \quad \text{or} \quad -1. \tag{7}$$

Using the LFNN constructed in line with Eqs. (6) and (7), sample train data is simultaneously introduced to both input and output layers. The LFNN suitably modifies its weights until an acceptable error level between predicted and desired outputs is attained. Then, by using LFNN of the final weights, the test set performance of the network is tested over a previously unseen data set. If test data predictions are good enough, the LFNN is considered to have consistently learned or generalized the inherent functional relationship existing between input and output data.

2.2.2. LFNN relevance to vibrational intensity EPF construction

Because a deterministic or random EPF is usually a mathematical vector function $\vec{y} : R^p \to R^r$ between the physical variables under interest, particularly LFNN (not any other ANN) is relevant to EPF construction. Hence, being a general input–output function estimator, the LFNN defined by Eq. (6) is particularly suitable in this context. But, in physics, although there can be several independent variables [$p \succ 1$ in Eq. (6)], the number of the dependent variable is usually one [$r = 1$ in Eq. (6)]. Train sample data for independent and dependent physical variables are presented to the input and output layers, respectively. Then, through weight adaptation process, the LFNN finally estimates the unknown generally nonlinear EPF. Note that EPF is a general abstract term, and in this paper it is specifically used for IR absorbance or Raman *vibrational intensity* (see Section 3.4). It must also firmly stated that depending on the number of hidden layer, hidden units, the kind of activation functions etc., we can construct infinitely many LFNN-EPFs. But, as shown in [13], in practice any of the final approximation function in Eq. (6) can be safely chosen as the desired EPF.

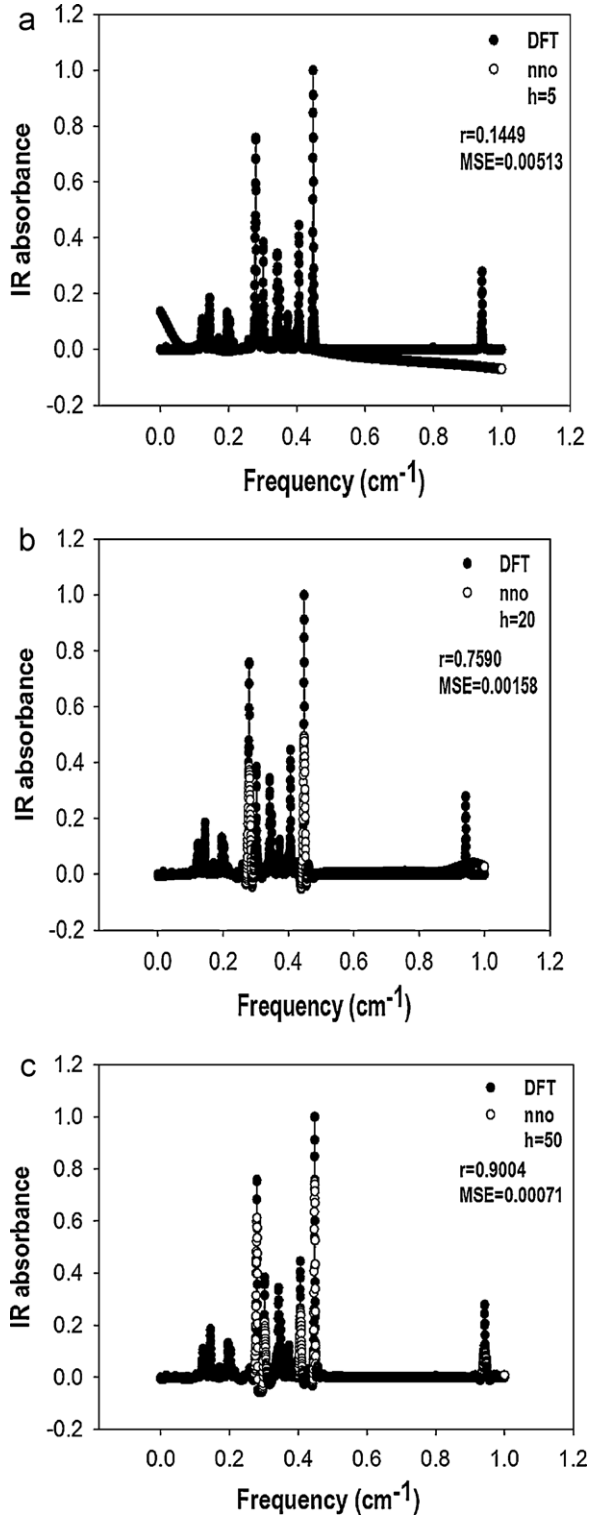

Fig. 2. 6-CNA IR absorbance versus frequency. DFT based absorbance (DFT) and their single hidden layer LFNN-EPF *train set* fittings (*nno*). h: hidden layer neuron number. (a) $h = 5$, (b) $h = 20$, (c) $h = 50$.

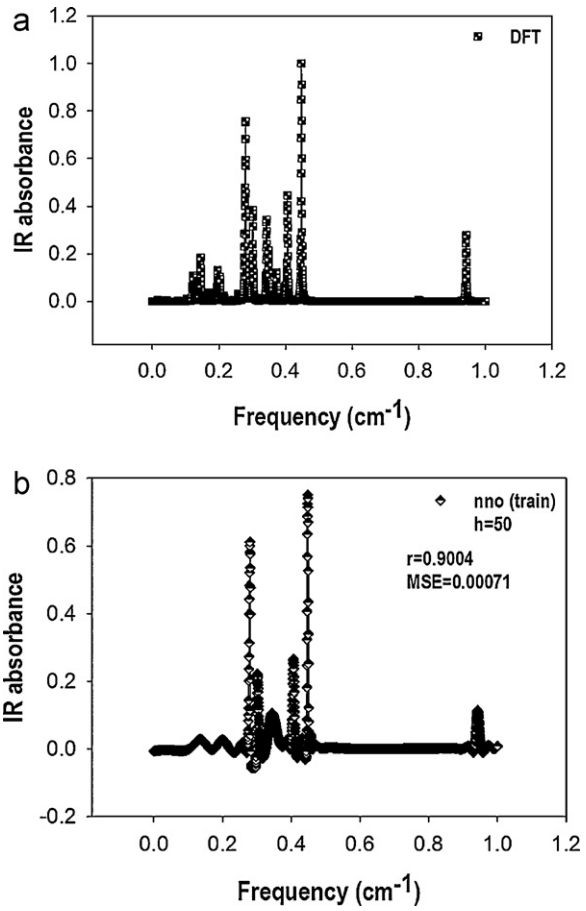

Fig. 3. 6-CNA IR absorbance versus frequency. DFT based absorbance (DFT) and their single hidden layer LFNN-EPF *train set* fittings (*nno*) with $h = 50$. h: hidden layer neuron number. (a) DFT, (b) *nno*.

3. LFNN application details

3.1. The vibrational spectral data for LFNN-EPF

The DFT-based vibrational data for LFNN-EPF construction were IR absorbance and Raman intensities of 6-CNA versus scaled frequencies. The data was borrowed from our previous work [17].

3.2. The LFNN structure used and the pre-train data transformation

The neural network software used was NeuroSolutions V5.06. For LFNN fitting of DFT calculated IR absorbance (input) versus frequency (output) data, we used one input layer neuron ($p = 1$) and one output layer neuron ($r = 1$) with one-hidden layer LFNN of $h_1 = 5$, 20 and 50 hidden neurons (see Fig. 1). In this case, From Fig. 1, the total number of adjustable weights ($p \times h_1 + h_1 \times r = h_1 \times (p + r) = h_1(1 + 1) = 2h_1$) was min $= 2 \times 5 = 10$, max $= 2 \times 50 = 100$, respectively. For LFNN fitting of DFT calculated Raman intensity (input) versus frequency (output) data, we again used one input layer neuron ($p = 1$) and one output layer neuron ($r = 1$) but with varying hidden layers. We first tried one-hidden layer LFNN. However, the fittings were not good enough as we mention more in Sections 4.1 and 4.2. Therefore, after many trials with varying hidden layer number, finally we had to use as large as ten (10) hidden layer LFNN with 5 hidden neurons at each hidden layer. In this case, the number of hidden layer weights can be calculated in the following way. There are $5 \times 5 = 25$ weights for any consecutive hidden layers. As there are 9 consecutive hidden layers, we have total of $25 \times 9 = 225$ hidden weights. Additionally, there are 5 weights from input to first hidden layer and 5 weights from last hidden layer to output layer. Therefore, we totally have $225 + 5 + 5 = 235$ weights (see Fig. 1). The number 235 for total weights may at first seem much larger, for instance when compared with 54 total weights in our previous LFNN-EPFs paper [14] for vibrational intensities. But, there are two important justifications for using much larger number of weights here. First, in comparison with our initial paper [14] consisting of totally only 63 DFT data points, in this paper we used 3999 total of absorbance/intensity versus frequency DFT data points. We see that LFNN data here is about 63 times larger in size. This naturally involves much larger number of weights. Second, as we mention more in Sections 4.1 and 4.2, both fitting and predictions of LFNN in this paper were much better than previous paper [14] results. Indeed, measured by correlation coefficients r, in this paper $r = 0.9$ for train set, while it was 0.5 for IR and 0.8 for Raman at most in Ref. [14]. No bias weight was used. Both train and test set input–output spectral data in this paper were normalized into *multi-dimensional unit intervals* before any LFNN processing. The type of activation functions G in Eq. (6) were, respectively, hyperbolic tangent $\tanh = (e^x - e^{-x})/(e^x + e^{-x})$ for hidden and linear for output layer. The LFNN weight adapting algorithm was back-propagation with Levenberg–Marquardt. For all LFNN processing cases, the whole DFT-vibrational data were *uniformly* partitioned into two separate sets (80% and 20%) to use as LFNN training set for fitting and test set for prediction, respectively.

3.3. The error function and its minimization

The error function to measure the difference between desired and actual neural network outputs was the mean square error (MSE). More clearly, in terms of the components of the user specified desired output vector $\vec{y}$ of Fig. 1 and of the LFNN predicted output vector $\vec{f}$ of Eq. (6), MSE is defined by Eq. (8),

$$\text{MSE} = \frac{\left[\sum_{k=1}^{r}\sum_{i=1}^{N}(y_{ki} - f_{ki})^2\right]}{N}, \tag{8}$$

where N is the number of training or test samples. Epoch (one complete presentation of the all input-desired output data to the network being trained) number was usually 1000. The best final LFNN approximation errors were nearly 0.0007 (for both train and test data) IR and 0.001 (for both train and test data) Raman.

3.4. The concrete algorithm for 6-CNA vibrational spectra LFNN-EPF construction

In order to construct suitable EPFs for highly nonlinear DFT calculated 6-CNA vibrational absorbance and intensity, we used *one neuron-output* LFNN vector function $\vec{f}$ in Eq. (6). However, Eq. (6) is itself not enough for the complete construction of *the desired nonlinear* EPF, because it gives only the *crude structure* of the LFNN without producing the final EPF parameters/final LFNN optimal weights. Therefore, in order to obtain both the final weight vector $\vec{w}_f$ (consisting of the final components of w^1 and w^2 of Fig. 1) and the corresponding LFNN output vector function $\vec{f}_{\min} = \vec{f}(\vec{w}_f)$ of Eq. (6), we simultaneously used Eqs. (6) and (8). More concretely, given the desired input–output experimental data, $\vec{f}_{\min}$ is the network output vector function of Eq. (6) giving the minimum MSE of Eq. (8) by a suitable LFNN weight adaptation. Also note that, $\vec{f}_{\min}$ is the best nonlinear estimation vector of the theoretically unknown desired output function $\vec{y} : R^p \to R^r$. In other words, the theoretically unknown vector function $\vec{y}$ is estimated by $\vec{f}_{\min}$. Briefly, $\vec{f}_{\min}$ is actually *the desired nonlinear* EPF, which we aim to ultimately

obtain. Because $\vec{f}_{\min}$ is the extremely important ultimate quantity, so we re-state it as a separate expression in Eq. (9),

$\vec{f}_{\min}$ of this paper : Given input–output data ($\vec{x}$ and $\vec{y}$ samples) and final weight vector $\vec{w}_f$, LFNN $\vec{f}_{\min} = \vec{f}(\vec{w}_f)$ of Eq. (6) and of Eq. (8) is our desired vibrational absorbance or intensity EPF. In this paper, LFNN input vector $\vec{x}$ (one dimensional) was the scaled frequency, one dimensional desired vector $\vec{y}$ was either IR absorbance or Raman intensity of the 6-choloronicotinic acid (6-CNA). Final details for $\vec{f}_{\min}$ of this paper are given in Section 3.5. (9)

3.5. Final $\vec{f}_{min}$ details

As stated in Eq. (9), we now give the final $\vec{f}_{\min}$ details. In Eq. (9), $\vec{f}_{\min}$ totally depends on the structure of the network output vector function $\vec{f}$ and the final weight vector $\vec{w}_f$. In Eq. (6), the weight components are embedded in $A(\vec{x})$ and $\vec{\beta}$ (w^1 and w^2 in this paper in Fig. 1). In Eq. (6), $\vec{f}$ depends on the explicit forms of G and A functions. In this paper, setting $\vec{\beta} = w^2$ of Fig. 1, G is *nonlinear tangent hyperbolic* and of A is the scalar product of w^1 and $\vec{x}$ of Fig. 1. Thus, we can construct the explicit form of $\vec{f}$. Then, by the minimization of Eq. (8), we finally obtain $\vec{f}_{\min} = \vec{f}(\vec{w}_f)$. The concrete LFNN-EPF construction algorithm for nonlinear DFT-based vibrational absorbance and intensity of 6-CNA is now complete. The actual LFNN-EPFs results obtained by this algorithm are given in Section 4.

4. Results and discussion: LFNN-EPFs construction

In this paper, one input–one output layer LFNN was used. The hidden layer number, after many suitable pre-train trials, was either *one* or *ten*. In figures where it applies, the abbreviation DFT is used for the DFT calculated data based on vibrational IR absorbance or Raman intensity. The DFT data was borrowed from our previous work [17] and used in this paper as both LFNN train and test data. Hence, neural network output (*nno*) was either vibrational IR absorbance or Raman intensity. The abbreviation *nno* is for both train or test set results.

4.1. LFNN-EPFs for train set fittings

As mentioned in Section 3.2, for all train and test set LFNN processing in this paper, the original DFT data were first normalized into *multi-dimensional unit* ([0, 1]) *intervals*. Let h be the hidden neuron number in a particular hidden layer (see Fig. 1). For a single hidden layer LFNN, the train set $h = 5$, 20 and 50 *nno* fittings of 6-CNA DFT IR absorbance versus frequency are given in Fig. 2a–c. As can be seen, with increasing number of h, the fittings greatly improves as measured by MSE and r correlation coefficient values. Here the correlation coefficient is an indicator of the goodness of the fit. For instance the $h = 50$ MSE (0.00071) is significantly lower than $h = 5$ MSE (0.00513). Additionally and more importantly, correlation coefficient of the fit $r = 0.9004$ in $h = 50$ and $r = 0.1449$ in $h = 5$, showing that with greater h value, the LFNN fittings are much better. Indeed, in Fig. 2c, the very complex DFT IR absorbance pattern was excellently fitted by the LFNN with $h = 50$. Note that the every single absorbance peak was excellently fitted. Moreover, the background absorbance values not corresponding to any peak values were also excellently fitted. However, in order to assess the fitting capabilities of the LFNN, particularly note that even the fittings with $h = 5$ LFNN is still acceptably well although the correlation coefficient is low. As can be seen by comparing Fig. 2a and c, the much lower correlation coefficient with $h = 5$ cannot generally be attributed to peak fittings. They seem to be reasonably fitted. The only deviations are with back ground fittings, producing naturally lower correlation coefficient. When compared with the best correlation coefficient ($r = 0.503$) of the fit in our previous initial work of LFNN-EPFs for vibrational intensities [14], in this paper we obtained much higher best fitting correlation coefficient $r = 0.9004$. This is pleasingly in very good agreement with our general postulations in Ref. [14] in the following way. In Ref. [14], we had to use much *less denser* DFT data, so the LFNN vibrational spectra fittings were not

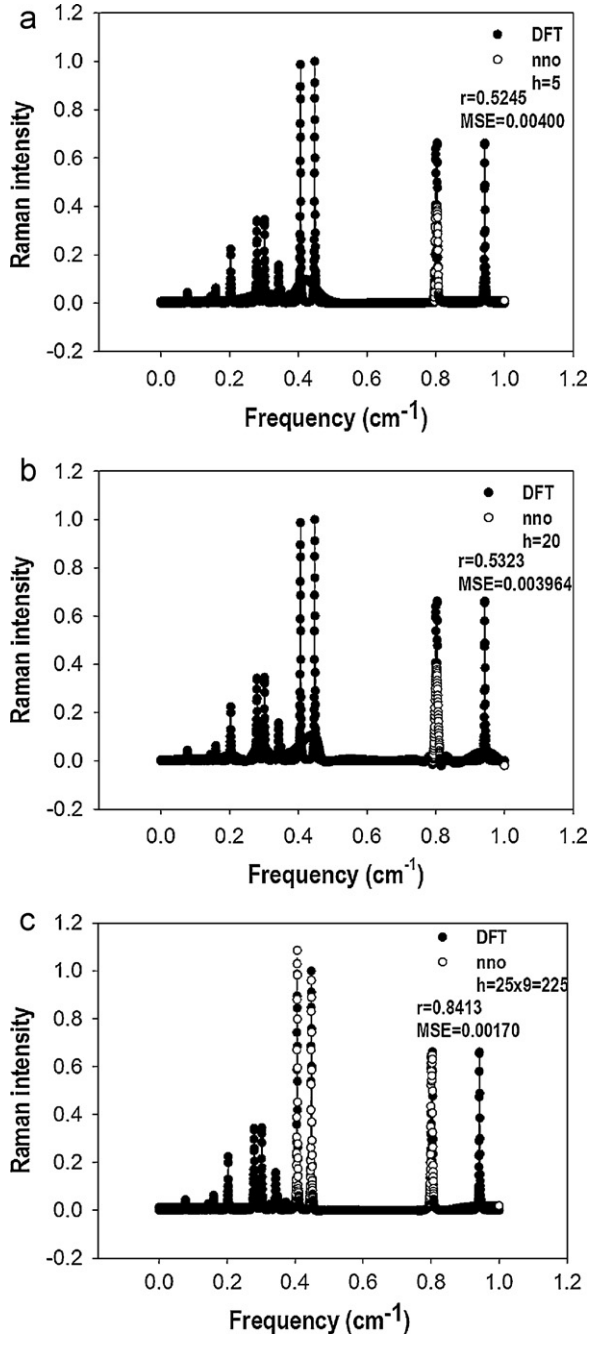

Fig. 4. 6-CNA Raman intensity versus frequency. DFT based absorbance (DFT) and their varying (single or ten) hidden layer LFNN-EPF *train set* fittings (*nno*). h: hidden layer neuron number. (a) $h = 5$ (single), (b) $h = 20$ (single), (c) $h = 225$ (ten).

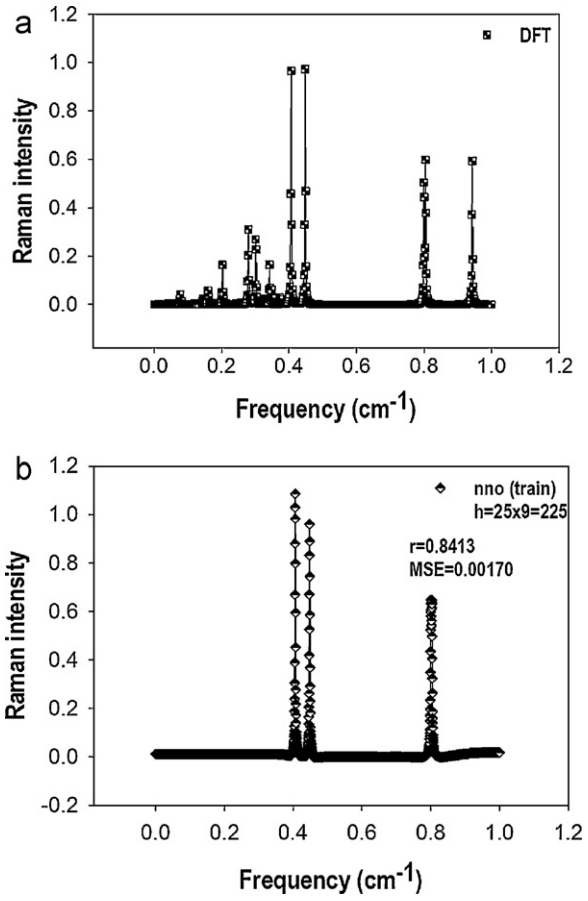

Fig. 5. 6-CNA Raman intensity versus frequency. DFT based absorbance (DFT) and ten hidden layer LFNN-EPF *train set* fittings (*nno*). *h*: hidden layer neuron number. (a) DFT, (b) *nno*. $h = 225$ (ten).

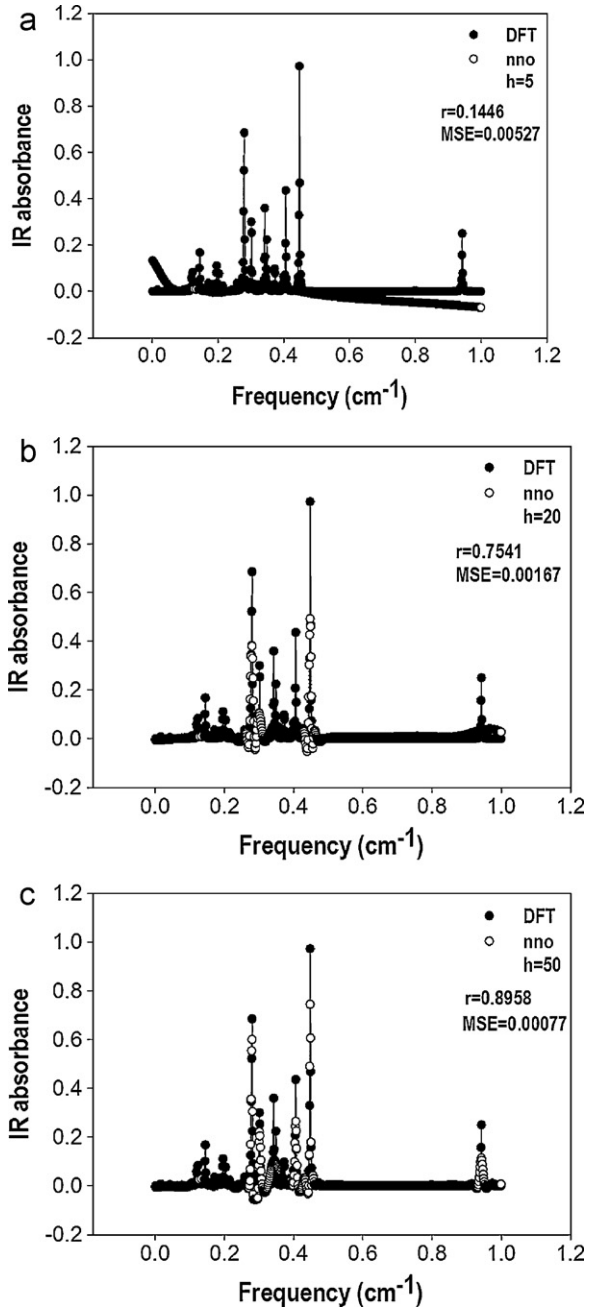

Fig. 6. 6-CNA IR absorbance versus frequency. DFT based absorbance (DFT) and their single hidden layer LFNN-EPF *test set* predictions (*nno*). *h*: hidden layer neuron number. (a) $h = 5$, (b) $h = 20$, (c) $h = 50$.

so good as desired, although the general tendency of the spectrum was still acceptable. The details of the peaks were really missing in LFNN fittings in Ref. [14]. But, as we firmly pointed out in Ref. [14], the previous study was intended only for a crude start into LFNN-EPFs of vibrational spectra. We also very clearly postulated in Ref. [14] that provided that spectral data were measured over much denser frequency interval, much more suitable LFNN-EPFs could be ultimately constructed. Compared with Ref. [14] data, the spectral data was much denser in this paper (63 times denser). So, we were anticipated to obtain much better LFNN-EPFs in this paper. Indeed, in this paper the LFNN-EPFs in Fig. 2a–c, most noticeably in Fig. 2c, are excellently better in details when compared with the LFNN-EPFs of Ref. [14] which only pay attention to crude functional tendency fitting. Briefly, Fig. 2c alone is an early proof of the postulate of Ref. [14] that much denser DFT spectral data can be fitted in much better manner by a suitable LFNN-EPFs. To be much more confident and conclusive in this assertion, we clearly have to provide more supporting LFNN-EPFs results. We will now do this. First, we must very explicitly further clarify an interesting and even potentially doubtful point with Fig. 2c. In Fig. 2c, because the number of data points was exceedingly high (about 3200 measurements) and also the *nno* values were excellent agreement with theoretically calculated DFT values, two distinctive patterns (DFT and *nno*) may be difficult to discriminate. Perhaps, one might even be doubtful as to whether there is really *only one pattern* (DFT) or two different patterns (DFT and *nno*) in Fig. 2c. In Fig. 3a and b, the two making up patterns of Fig. 2c (DFT and *nno*) were separately reproduced for a comparison so that no doubt should arise.

The train set *nno* fittings of 6-CNA DFT Raman intensity versus frequency are given in Fig. 4a–c. We first tried single-hidden layer LFNN with $h = 5$ and 20 (see Fig. 4a and b), but by measuring from correlation coefficients ($r = 0.5245$ and 0.5323, respectively) the fittings were not as good as desired although the peak and back ground fitting were still acceptable. After many trials, we found that 10 hidden layer with $h = 5$ (total of 225 hidden weights) LFNN produced the best result ($r = 0.8413$ and MSE = 0.00170) in Fig. 4c. As can be clearly seen from Fig. 4c, the *nno* fitting of Raman DFT data is exceptionally well. Particularly note that every single Raman peak was fitted with a great success. The same applies to the background fitting. Compared with $r = 0.9004$ in Fig. 2c of IR, $r = 0.8413$ in Fig. 4c of Raman is slightly lower because Raman intensity data seem to be more

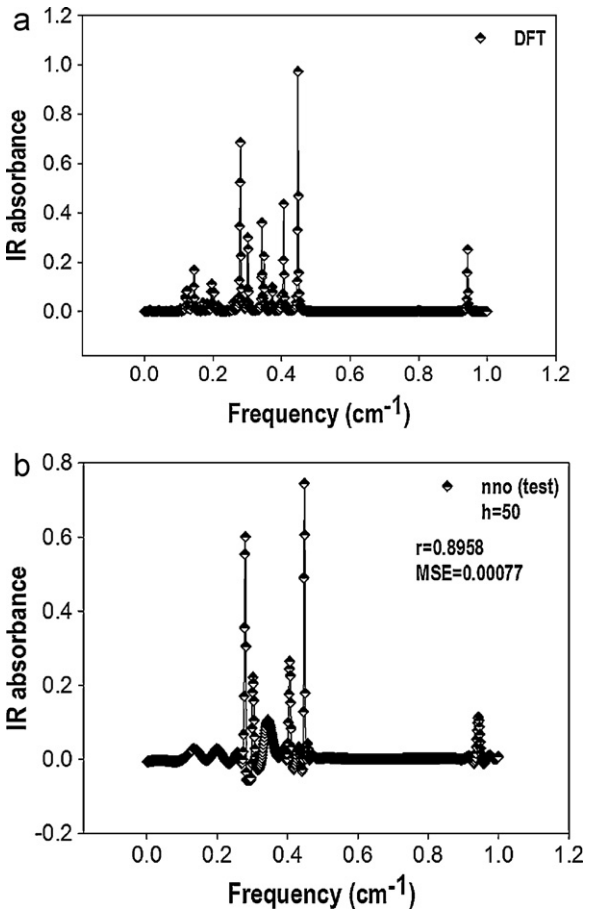

Fig. 7. 6-CNA IR absorbance versus frequency. DFT based absorbance (DFT) and their single hidden layer LFNN-EPF *test set* predictions (*nno*) with $h = 50$. h: hidden layer neuron number. (a) DFT, (b) *nno*.

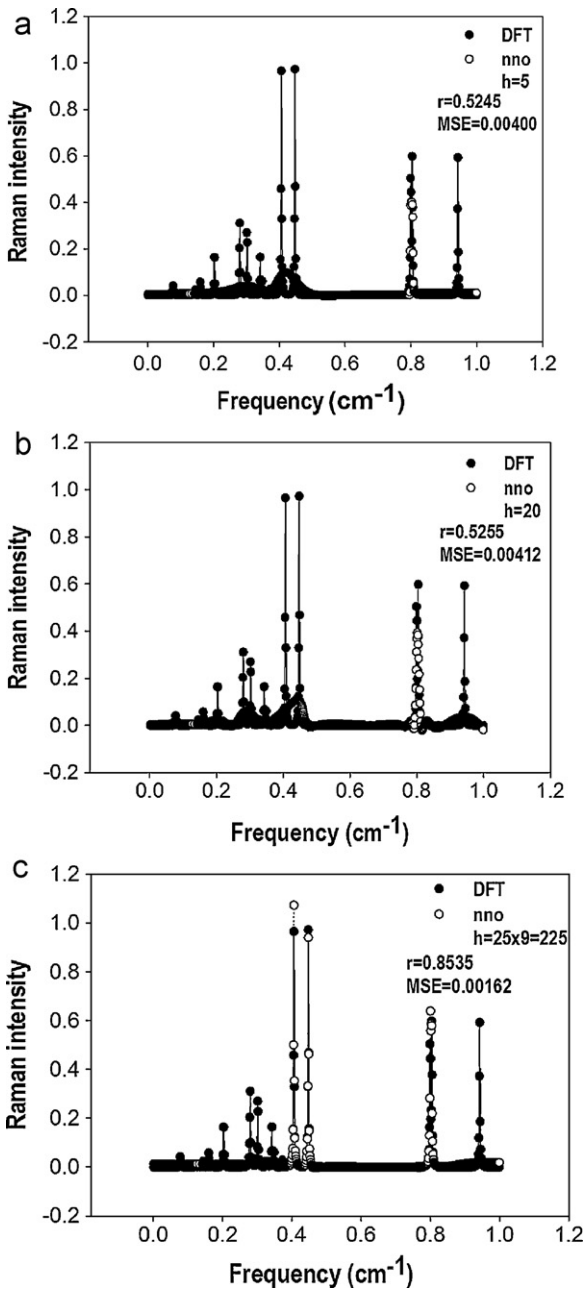

Fig. 8. 6-CNA Raman intensity versus frequency. DFT based absorbance (DFT) and their varying (single or ten) hidden layer LFNN-EPF *test set* predictions (*nno*). h: hidden layer neuron number. (a) $h = 5$ (single), (b) $h = 20$ (single), (c) $h = 225$ (ten).

complex in terms of the magnitude and location of peaks. This relative complexity of Raman spectra may also explain the fact that why we had to use much larger hidden weights for Raman spectra ($h = 225$) while using much lower IR spectra hidden weights ($h = 50$). Also note that both in Figs. 2c and 4c, the very fact that vibrational DFT data inherently consist of sharply fluctuating peaks, they are *naturally* cannot be fitted at the greatest accuracy with r very close to unity. This is only possible for usual smooth functions, for instance for polynomials, exponentials, etc. In other words, as far as the utmost complexity of the vibrational data is concerned, the correlation coefficients both in Figs. 2c and 4c ($r = 0.9004$ and $r = 0.8413$, respectively) can be safely regarded high enough, indicating excellent fittings. Again, due to the utmost complexity of the patterns in Fig. 4c, the two making up patterns of Fig. 4c (DFT and *nno*) are separately reproduced in Fig. 5a and b for a comparison so that no doubt should arise as to ultimate discrimination of DFT data and *nno* fittings.

4.2. Consistency of the constructed LFNN-EPFs: test set predictions

If the 6-CNA train set LFNN-EPFs do not be further tested over "yet-to-be measured" 6-CNA DFT data, these fitted EPFs cannot be used *consistently* over a desired range of DFT values. In other words, if the 6-CNA train vibrational absorbance/intensity LFNNs well predict previously unseen *test set data*, then the LFNN have regarded to have successfully generalized the data, proving consistent estimations. If the estimations are consistent with the test data values, then the LFNNs can be taken as suitable LFNN-EPFs.

For 6-CNA DFT IR absorbance versus frequency, the corresponding test set predictions of Fig. 2a–c are given in Fig. 6a–c. Again, the single hidden layer train set LFNNs with $h = 5$, 20 and 50, which led to Fig. 2a–c were also used for *nno* test set predictions. As can be seen, with increasing number of h, the predictions greatly improves as measured by MSE and r correlation coefficient values, $h = 50$ producing in Fig. 6c the best prediction with lowest MSE (0.00077) and highest r (0.8958). As the number of test data (about 800) was much lower than its corresponding test set data, *nno* predictions and DFT values can be discriminated in Fig. 6c without much difficulty. But, the *nno* predictions in Fig. 6c have still pleasingly and expectedly significant overlap with the DFT data. Therefore,

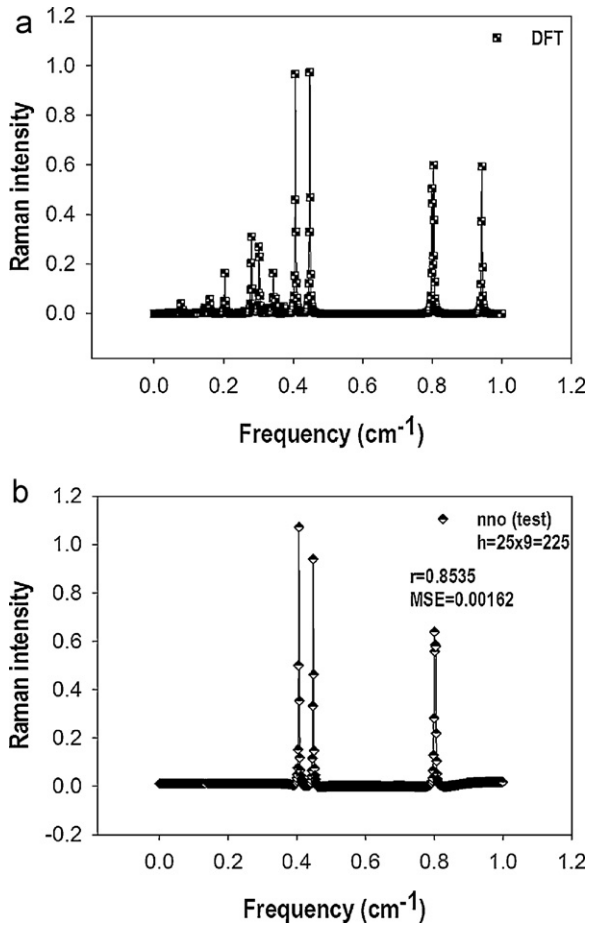

Fig. 9. 6-CNA Raman intensity versus frequency. DFT based absorbance (DFT) and ten hidden layer LFNN-EPF *test set* predictions (*nno*). *h*: hidden layer neuron number. (a) DFT, (b) *nno*. $h = 225$ (ten).

it is separately reproduced in Fig. 7a and b for a comparison. As can be seen for instance in Fig. 6c, the test set *nno* predictions are very good agreement with DFT data. This clearly shows that the test set LFNNs of IR absorbance have consistently generalized the train LFNN fittings. Therefore, LFNNs obtained can be safely used as LFNN-EPFs because the physical law embedded in DFT IR data has been successfully extracted by the LFNN constructed.

For 6-CNA DFT Raman intensity versus frequency, the corresponding test set predictions of Fig. 4a–c are given in Fig. 8a–c. Again, the corresponding train set LFNNs of final weights which led to Fig. 4a–c were also used for *nno* test set predictions. As can be seen, with increasing number of *h*, the predictions greatly improves as measured by MSE and *r* correlation coefficient values, 10 hidden layer with $h = 5$ (total of 225 hidden weights) LFNN producing the best predictions ($r = 0.8535$ and $\mathrm{MSE} = 0.00162$) in Fig. 8c. As the number of test data (about 800) was much lower than its corresponding test set data, *nno* predictions and DFT values can be discriminated in Fig. 8c without much difficulty. But, the *nno* predictions in Fig. 8c have still pleasingly and expectedly significant overlap with the DFT data. Therefore, it is separately reproduced in Fig. 9a and b for a comparison. As can be seen for instance in Fig. 6c, the test set *nno* predictions are very good agreement with DFT data. This clearly shows that the test set LFNNs of Raman intensity have consistently generalized the train LFNN fittings. Therefore, LFNNs obtained can be safely used as LFNN-EPFs because the physical law embedded in DFT Raman intensity has been successfully extracted by the LFNN constructed.

5. Conclusions and potential applications

In this paper, as a continuation of our very recently initiated entirely novel approach, LFNN-EPFs were constructed for density functional theory (DFT) vibrational spectra absorbances and intensities of a different molecule, 6-choloronicotinic acid (6-CNA). The major conclusion is stated in Section 5.1. Also, because, the vibrational spectral absorbances or intensities are important physical quantities directly related to distributions of the electric charges in a molecule, promising potential applications can be mentioned (Section 5.2).

5.1. Conclusion

Test set (i.e. yet-to-be measured experimental data) LFNN-EPFs consistently and successfully predicted the absorbance and intensity data. This simply means that the physical law embedded in 6-CNA absorbance and intensity data was successfully extracted by the LFNN-EPFs.

5.2. Potential applications

1. The vibrational LFNN-EPFs constructed in this paper are of explicit functional form. Therefore, by various suitable operations of mathematical analysis, they can be used to estimate the electronic charge distributions of the unknown molecule of the significant complexity. Additionally, these estimations can be combined with those of theoretical DFT atomic polar tensor calculations to contribute to the identification of the molecule.
2. The vibrational spectra LFNN-EPFs can be combined together with the findings of other existing vibrational spectra estimations by neural networks, for instance radial distribution function (RDF) coded IR spectra estimation mentioned in the introduction of this paper.

References

[1] B.S. Galabov, T. Dudev, in: J.R. Durig (Ed.), Vibrational Spectra and Structure: Vibrational Intensities, vol. 22, Elsevier Science B.V., Amsterdam, 1996.
[2] J. Overend, Quantitative intensity studies and dipole moment derivatives, in: M. Davies (Ed.), Infrared Spectroscopy and Molecular Structure, Elsevier, Amsterdam, 1963, p. 345.
[3] B. Crawford, J. Chem. Phys. 20 (1952) 977.
[4] M. Gussoni, S. Abbate, G. Zerbi, Prediction of infrared and Raman intensities by parametric methods, in: A.J. Barnes, W.J. Orville-Thomas (Eds.), Vibrational Spectroscopy: Modern Trends, Elsevier, Amsterdam, 1977, pp. 205–222.
[5] M. Gussoni, Infrared and Raman intensities from electrooptical parameters, in: R.J. Clark, R.E. Hester (Eds.), Advances in Infrared and Raman Spectroscopy, vol. 6, Heyden, London, 1979.
[6] M. Gussoni, P. Jona, G. Zerbi, J. Chem. Phys. 78 (1983) 6802.
[7] M. Gussoni, J. Mol. Struct. 141 (1986) 63.
[8] M. Gussoni, C. Castiglioni, J. Mol. Struct. 521 (2000) 1.
[9] R.L.A. Haiduke, Y. Hase, R.E. Bruns, Spectrochim. Acta A 57 (2001) 1369.
[10] A. Milani, C. Castiglioni, J. Phys. Chem. A 114 (2010) 624.
[11] A. Milani, D. Galimberti, C. Castiglioni, G. Zerbi, J. Mol. Struct. 976 (2010) 342.
[12] J. Gasteiger, Chemometr. Intell. Lab. 82 (2006) 200.
[13] N. Yildiz, Phys. Lett. A 345 (1–3) (2005) 69.
[14] N. Yildiz, M. Karabacak, M. Kurt, J. Mol. Struct. 1006 (2011) 642.
[15] S. Haykin, Neural Networks: A Comprehensive Foundation, 2nd ed., Prentice-Hall, NJ, 1999.
[16] K. Hornik, M. Stinchcombe, H. White, Neural Netw. 2 (1989) 359.
[17] M. Karabacak, M. Kurt, Spectrochim. Acta A 71 (2008) 876.
[18] J. Gasteiger, J. Sadowski, J. Schuur, P. Selzer, L. Steinhauer, V. Steinhauer, J. Chem. Inf. Comput. Sci. 36 (1996) 1030.
[19] M.C. Hemmer, V. Steinhauer, J. Gasteiger, Vib. Spectrosc. 19 (1999) 151.
[20] P. Selzer, J. Gasteiger, H. Thomas, R. Salzer, Chem. Eur. J. 6 (5) (2000) 920.
[21] L. Duponchel, C. Ruchebusch, J.P. Huvenne, P. Legrand, J. Mol. Struct. 480–481 (1999) 551.
[22] R. Goodacre, Vib. Spectrosc. 32 (2003) 33.
[23] C.B. Cai, H.W. Yang, B. Wang, Y.Y. Tao, M.Q. Wen, L. Xu, Vib. Spectrosc. 56 (2011) 202.
[24] M. Gussoni, C. Castiglioni, G. Zerbi, J. Chalmers, in: P. Griffiths (Ed.), Handbook of Vibrational Spectroscopy, vol. 3, John Wiley & Sons, Chichester, UK, 2001, p. 2040 and references therein.