Journal of Molecular Structure 1006 (2011) 642–649

Contents lists available at SciVerse ScienceDirect

Journal of Molecular Structure

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

Neural network consistent empirical physical formula construction for DFT based nonlinear vibrational spectra intensities of N-(2-methylphenyl) and N-(3-methylphenyl) methanesulfonamides

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

[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 1 July 2011
Received in revised form 11 October 2011
Accepted 12 October 2011
Available online 23 October 2011

Keywords:
Neural network
Vibrational intensity
Molecular structure
Empirical physical formula

ABSTRACT

Vibrational intensities are both experimentally measured and theoretically estimated important physical quantities which are directly related to distributions of the electric charges in a molecule. In this paper, as a novel approach, by a layered feedforward neural network (LFNN), empirical physical formulas (EPFs) were constructed for density functional theory (DFT) vibrational spectra intensities of N-(2-methylphenyl) and N-(3-methylphenyl) methanesulfonamides. The spectral data was obtained from our previous study. Although the DFT spectral data was inherently *extremely difficult-to-fit* (sparse frequency intervals, highly nonlinear and sharply fluctuating intensities), still the optimally constructed LFFN-EPFs succeeded in fitting this data to medium and higher level of satisfaction. Moreover, LFNN–EPFs test set (i.e. yet-to-be measured experimental data) intensity predictions were also moderate to higher level. This briefly means that the general tendency of the intensity data was consistently estimated by the LFNN to an acceptable degree. In conclusion, provided that vibrational spectral data measured over sufficiently dense frequency intervals are available for any unknown molecule of significant complexity, suitable LFNN–EPFs can be constructed. Then, by various mathematical tools such as differentiation, integration, minimization, these vibrational LFNN–EPFs can be used to estimate the electronic charge distributions of the molecule. Moreover, these estimations can be compared and combined with those of theoretical DFT atomic polar tensor calculations to contribute to the identification of the molecule.

1. Introduction

Vibrational spectral intensities are both experimentally measured and theoretically estimated important physical quantities which are directly related to distributions of the electric charges in a molecule [1]. The pioneering early studies [2,3] to interpret a wealth of experimental intensity data were later followed by works dealing with mainly the formal aspects of the interpretation of IR intensities [4,5]. After the formal matters were resolved reasonably, attempts into the physical meaning of intensity curves enabled to correlate some intensity patterns with charge distributions inside the molecule [6,7]. Very briefly, the measured vibrational intensities I_i of the ith normal mode are related to the atomic charges by the molecular dipole moment ($\vec{M}$) derivatives with respect to associated normal coordinate Q_i. More details into *intensity-atomic charge relationship* are given in Section 2.1. Here it suffices to say that, thanks to over five decade of efforts into this relationship, recent ab initio calculations have accurately predicted IR (infrared) intensity related molecular charge distributions, in very good agreement with experimental data [8–11].

Nevertheless, for the molecules of very complex structures, typical IR or Raman spectral intensities are naturally highly nonlinear and of complex pattern. In this case, it may be more difficult to perform appropriate ab initio intensity calculations in order to compare with the experimental data. Moreover, the identity and three dimensional (3D) structure of the chemical compound may be totally unknown before a set of appropriate spectral experiments are done. Indeed, although 40 million chemical compounds are known, experimental data for 3D structure are available only for 250,000 compounds. The largest database for IR spectra contains only 220,000 spectra. In other words, data for 3D structure and IR spectra are available only for 0.5% of all known chemical molecules [12]. Briefly, satisfactory ab initio vibrational intensity calculations cannot be made for the most molecules because of their unknown detailed spatial structures. Therefore, it would be helpful if the experimenter had a kind of explicit form of empirical physical formula (EPF) for this highly nonlinear vibrational intensity pattern

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

0022-2860/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.molstruc.2011.10.021

which experimentally obtained for the molecule of unknown detailed 3D structure. An appropriate EPF by using a layered feedforward neural network (LFNN) [13] can indeed be consistently constructed, as we previously theoretically [14] and experimentally [15,16] showed. Actually, the LFNN, as we mention more in Section 2.2, is a universal nonlinear function approximator [17].

In this paper, as a novel approach, by a LFNN, we constructed consistent EPFs for nonlinear density functional theory (DFT) vibrational spectra intensities of N-(2-methylphenyl) (2MPMSA) and N-(3-methylphenyl) methanesulfonamides (3MPMSA). The spectral data was obtained from our previous study [18]. Before explicitly mentioning the novelty in this paper, some preliminary remarks are necessary. There have been a number of previous works [19–24] which applied neural networks to molecular vibrational intensities. Among these works, the very systematic and pioneering studies of Gasteiger and coworkers [19–21] into deriving 3D molecule structures from their corresponding IR spectra must be mentioned with a particular stress. They successfully use radial distribution function (RDF) code for the simulation of an IR spectra by a suitable counterpropagation (CPG) neural network [25]. As a reverse operation, they also employ CPG neural network to obtain a 3D structure from a large database by using IR spectra as an input for the neural network. However, they do not particularly aim to construct an explicit mathematical functional form for the experimental IR spectra pattern. We can now state explicitly the novelty in this paper. This paper is the first to construct consistent *explicit form* of LFNN–EPFs for vibrational intensities, particularly for methanesulfonamides. Although the DFT spectral data was inherently *extremely difficult-to-fit* (sparse frequency intervals, highly nonlinear and sharply fluctuating intensities or activities), still the optimally constructed LFFN-EPFs succeeded in fitting this data to medium and higher level of satisfaction. Moreover, LFNN–EPFs test set (i.e. yet-to-be measured experimental data) intensity predictions were also moderate to higher level. This briefly means that the general tendency of the spectral intensity data were consistently estimated by the LFNN to an acceptable degree. We conclude that, provided that sufficient (non-sparse) vibrational spectral data available for any given molecule of significant complexity, suitable LFNN–EPFs can be constructed. Then, by various mathematical tools such as differentiation, integration, minimization, these LFNN–EPFs can be used to obtain further vibrational intensity related molecular structural parameters.

2. Theories

2.1. Atomic polar tensors and vibrational intensities

The basic parameters of any IR intensity predictions are the atomic polar tensors (APTs) [10] , the elements of which are the molecular dipole moment derivatives with respect to Cartesian displacements of atoms ξ_k as defined in Eq. (1). For Raman intensities molecular polarizability tensor $\boldsymbol{\alpha}$ is employed instead of dipole moments.

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

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

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

The underlying hypothesis of the equilibrium charges and charge fluxes (ECCF) theory [26] is the starting point, namely the identification of the molecular dipole moment $\vec{M}$ as the sum over point atomic charges q_α of:

$$\vec{M} = \sum_\alpha q_\alpha \vec{r}_\alpha, \quad (3)$$

where $\vec{r}_\alpha$ is the position vector of αth atom. From Eqs. (2) and (3), one obtains:

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

Eq. (4) is the general equation for uv component of APT. Rotation of the APT in a suitable Cartesian reference system gives simpler equations for APT components. Accordingly, relevant equilibrium charges and charge fluxes can be calculated in a straightforward manner. For instance, consider the planar molecules on a CH bond. With suitable local reference Cartesian system, one obtains the relationship of Eq. (5) between yy APT component and hydrogen equilibrium charge q^0_H.

$$P^H_{yy} = q^0_H. \quad (5)$$

Eq. (5) is the basic equation relating the measured IR absolute intensity and molecular atomic charges (see, for more [10]).

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

Although analyzed in depth in [14], here we again give the minimum basics of LFNN and its relevance to EPF construction.

2.2.1. The LFNN basics

An artificial neural network (ANN) [13], resembling the brain functionality, consists of artificial neurons which interconnect with each other by adaptive weights. The ANN can learn new knowledge by modifying weights. As a particular type of ANN, LFNN is a one input – many intermediate (hidden) – one output layer device, with all layers being interconnected by weights (Fig. 1).

Theoretically speaking, although a LFNN can be with any number of hidden layers, even a suitable *single hidden layer LFNN* is sufficient for excellent nonlinear function approximation [17]. Therefore, in this paper, we report *only single hidden layer* LFNN–EPFs, although

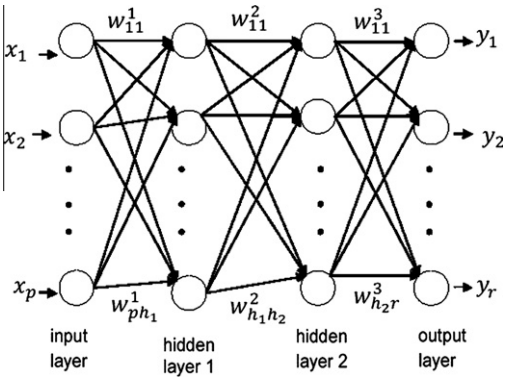

Fig. 1. Fully connected one input-many hidden-one output layer LFNN. Only two hidden layers are shown. $x_i(i = 1, \ldots, p)$ and $y_i(i = 1, \ldots, r)$ are, respectively, input and output vector components. Circles: artificial neurons, arrows: adaptable synaptic weights. w^i_{jk}: weight vector component, 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 .

we also used *three hidden layer* LFNNs to see whether there was any significant improvement in LFNN–EPFs when compared with single hidden layer ones. As we also mention Sections 3.2 and 4, we did not see any significant improvement. For a single hidden layer LFNN, in Fig. 1, the *network multi-output vector* $\vec{f}$ to approximate the desired output vector $\vec{y}$ is given by Eq. (6) (see [17]),

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

where A^p is the set of all functions of $R^p \to R$ defined by $A(\vec{x}) = \vec{w}.\vec{x} + b$. '.' is the scalar product, $\vec{w}$ is input to hidden layer weight vector, $\vec{x}$ is the network input vector in Fig. 1, and b is the bias. The columns of the weight matrices w^1 and w^2 in Fig. 1 actually correspond to weight vectors defined in $A(\vec{x})$ and $\vec{\beta}$ in Eq. (6). But, note that, as can be seen from Eq. (6) and Fig. 1, the correspondences $w^1 \to A(\vec{x})$ and $w^2 \to \vec{\beta}$ are valid *only* for single hidden layer LFNN. For more than single hidden layer LFNN, both Eq. (6) and the correspondences must be modified accordingly. *Also, in Eq. (6), the hidden neuron activation function G:R → R can be any well-behaved nonlinear function; therefore a LFNN is a universal nonlinear function approximator*. In applications, G is usually a kind of nonlinear sigmoid defined by:

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

With the LFNN of Eqs. (6) and (7), sample train data is simultaneously presented to both input and output layers. The network appropriately modifies its weights until a reasonable error level between predicted and desired outputs is attained. Then, by using the final weights, the performance of the network is tested for a previously unseen test data set. If test data predictions are successful, then the LFNN is considered to have learned (generalized) the inherent functional relationship between input and output data.

2.2.2. Relevance of LFNN to vibrational intensity EPF construction

Particularly LFNN (not any other ANN) is relevant to EPF construction, because a deterministic or random EPF is generally expressed as a mathematical vector function $\vec{g} : R^p \to R^r$ between the physical variables considered. Therefore, as a general input–output function estimator, the LFNN in 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)]. Still, in this paper, we used *multi-component-output* ($r = 2$) LFNN for two specific reasons. First, we wanted to deal with only one LFNN error function rather than dealing with several separate error functions for each vibrational method (IR and Raman). Second, it would be awkward to present too many LFNN visual data plotted separately for each vibrational method. Sample data for independent and dependent physical variables are presented to the input and output layers respectively. Then, the LFNN finally estimates the unknown generally nonlinear EPF by a weight adaptation process. Note that EPF is a general abstract term, and in this paper it is specifically used for the *vibrational intensity* (see, Section 3.4). Also note that, depending on the number of hidden layer, hidden units, activation functions, etc., we can obtain infinitely many LFNN–EPFs. But, as shown in [14], any of the final approximation function in Eq. (6) can be safely used as the desired EPF.

3. LFNN application details

3.1. The original vibrational spectral data for LFNN–EPF

The DFT-based vibrational data for LFNN–EPF construction were IR intensities and Raman scattering activities of 2MPMSA and 3MPMSA versus scaled frequencies. The original data was taken from our previous work [18].

3.2. The LFNN structure used and the original data transformation

The neural network software used was NeuroSolutions V5.06. Although we also used *three hidden layer* LFNNs to see whether there was any significant improvement in LFNN–EPFs when compared with single hidden layer ones. We did not see any improvement. Therefore, in this paper, we report only for the *single hidden layer* LFNN (with $h_1 = 4$, 7, 12 and 18 hidden neurons), one input layer neuron ($p = 1$) and *two output neuron* $r = 2$. The input was "scaled frequency" and the output components were "the IR and Raman intensities". From Fig. 1, in this paper the number of adjustable weights $(p \times h_1 + h_1 \times r = h_1 \times (p + r) = h_1(1 + 2) = 3h_1)$ are min = $3 \times 4 = 12$, max = $3 \times 18 = 54$. No bias weight was used. Although in our previous work [16] we did not normalize the train and test data into multi-dimensional unit intervals, in this paper we had to normalize the spectral data before LFNN processing. The sole reason for this normalization was that the original spectral data not only consisted of high magnitude of intensities but also was extremely sharply fluctuating. However, even the normalization did not much improve the LFNN fitting because the sharply fluctuating character of the original data was also inherited in the normalized data. Therefore, in order to smooth out considerably the sharp fluctuations in the original data, we performed log–log transformation over input–output original vibrational data. As can be seen from the corresponding figures of Section 4, with log–log transformation the LFNN–EPFs were considerably improved. However, as expected, the original sharp fluctuations were also present in log–log data, although the log–log data was easier to fit by suitably constructed LFNN–EPFs. The 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 momentum. The final weight components varied from about −3 to 3. In all cases leading to LFNN–EPF results, the 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 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:

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

where N is the number of training or test samples. Epoch (one complete presentation of the all input-desired data to the network being trained) number was usually 1000. The best final network approximation error was about in the range 0.030–0.050 for both train and test data.

3.4. The concrete algorithm for vibrational intensity LFNN–EPF construction

In order to construct appropriate EPF estimates for nonlinear DFT-vibrational intensities, we used multi-output LFNN vector function $\vec{f}$ in Eq. (6). However, to actually obtain *the desired nonlinear* EPF, Eq. (6) is itself not enough as it gives only the structure of the LFNN without saying anything about final EPF parameters or equivalently 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 employed Eqs. (6) and (8). More concretely, given the desired input–output experimental data, $\vec{f}_{\min}$ is network output vector function of Eq. (6) which gives minimum *MSE* of Eq. (8) through a proper LFNN weight adaptation process. Also note that, $\vec{f}_{\min}$ is the best nonlinear estimation vector of the theoretically unknown response vector function $\vec{g} : R^p \to R^r$. In other words, $\vec{f}_{\min}$, is *the desired nonlinear* EPF, which we aim to ultimately obtain. Because $\vec{f}_{\min}$ is so important for our ultimate purpose that we re-state it in Eq. (9),

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 intensity EPF. In this paper, LFNN input vector $\vec{x}$(one dimensional) was the scaled frequency , two dimensional desired vector $\vec{y}$ consisted of the IR intensity and Raman activities of the methanesulfonamides. Final details for $\vec{f}_{\min}$ of this paper are given in the text. (9)

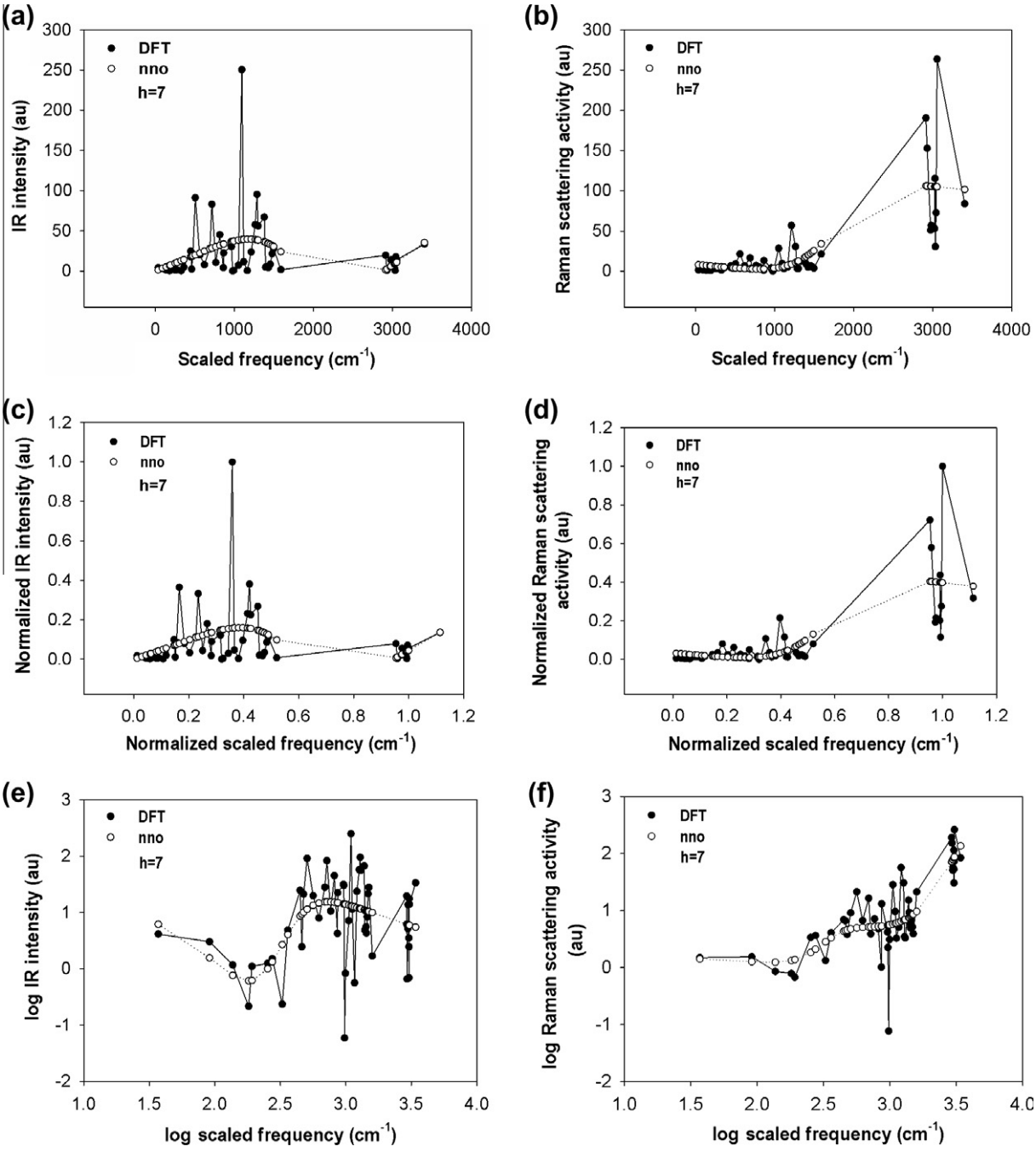

Fig. 2. For 2MPMSA, vibrational (IR and Raman) intensities in arbitrary units (au) versus scaled frequency. DFT based vibrational intensity calculations (DFT) and their LFNN–EPF *train set* fittings (*nno*). *h*: hidden layer neuron number. (a) IR-original untransformed data (b) Raman-original untransformed data (c) IR-normalized data (d) Raman-normalized data (e) IR-log data (f) Raman-log data.

As stated in Eq. (9), we now give the final details of concrete construction of $\vec{f}_{\min}$. In Eq. (9), the desired vector function $\vec{f}_{\min}$ totally depends on the structure of the network output vector function $\vec{f}$ and the final weight vector $\vec{w}_f$. Now, 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, the explicit form of 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 $\vec{f}$. Now, through 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 intensities of N-(2-methylphenyl) and N-(3-methylphenyl) methanesulfonamides is now complete. For actual LFNN–EPFs results obtained by this algorithm, see Section 4.

4. Results and discussion: LFNN–EPFs construction

Because we did not see any considerable improvement by using *three hidden layer* LFNNs when compared with single hidden layer ones, in this paper, we report *only single hidden layer* LFNN–EPFs. In Figures, the abbreviation DFT is for the DFT calculated vibrational intensities taken from our previous work (Ref. [18]) with which we compare our LFNN neural network output (*nno*) vibrational intensities obtained this paper. The abbreviation *nno* is for both train or test set results. The abbreviation *au* is for arbitrary units.

4.1. LFNN–EFPs for train set fittings

Train set $h = 7$ *nno* fittings of 2MPMSA DFT vibrational intensities were presented for untransformed original data (scaled frequency versus IR or Raman intensity of Fig. 2a and b), normalized data (Fig. 2c and d) and log–log transformed data (Fig. 2e and f). As already mentioned above and also can be clearly seen from Fig. 2a and b, the original vibrational intensities and activities had three following noticeable characteristics which made them *inherently extremely difficult-to-fit*. They were first: sparsely taken frequency intervals, second: relatively high intensity and activity magnitudes, third and last: extremely sharp fluctuations (IR in particular). In spite of these highly unfavorable characteristics of the DFT data, as can be seen from Fig. 2a and b, the LFNN–EPFs succeeded in fitting the data in the following sense. If we do not take the sharp values into account, the *nno* fittings in both Fig. 2a and b are considered to be good as far as the general tendency of the DFT data is considered. However, the correlation coefficients r measuring the goodness of fits were 0.349 (lower medium) for Fig. 2a (IR) and 0.761 (medium and higher) for Fig. 2b (Raman). Admittedly, $r = 0.349$ is not much satisfying. However, in Fig. 2a, given that the frequency intervals were sparse and several sharp pick values existed, the best the LFNN can do is to extract the general tendency without concentrating on fitting extremely sharp picks. In other words, if the frequency intervals were much denser, then the LFNN–EPFs would be much better, producing satisfactorily high r values. For instance, in our previous work [27] where wavelength intervals of the light source were desirably denser, the LFNN–EPFs fitted the vividly fluctuating transmittance data to a very high correlation coefficients about $r = 0.96$. As for Fig. 2b, the $r = 0.761$ is a pleasingly high value, in spite of sharp picks and sparse frequencies. Given that characteristics of IR (Fig. 2a) and Raman (Fig. 2b) data are similar, one naturally wonders why r of Fig. 2b is much higher than that of Fig. 2a. One explanation for this is that 37–1500 cm^{-1} frequency region of Raman (Fig. 2b) much smoother than that of Fig. 2a, if the one or two extremely sharp picks in both Fig. 2a and b are not taken into account. So, 37–1500 cm^{-1} is fitted better in Fig. 2b, contributing the higher r value. One more contribution to higher r value of Raman in Fig. 2b when compared with IR in Fig. 2a is as follows. In Fig. 2b, about 3000 cm^{-1} the LFNN also attempts to a degree to fit two extremely sharp picks of close in value, while not much attempting to fit the single extremely sharp value in Fig. 2a located at about 1100 cm^{-1}. Briefly, it may be said that the LFNN also cares about the regional fitting, even if there may be grouped intensities of relatively high magnitudes in noticeably separated spectral regions. In order to deal with much smaller frequency and vibrational magnitudes, the original data were normalized into [0,1] multiunit intervals. However, as can be seen from Fig. 2c and d, the normalization did not change the sharply fluctuating character of the original data at all. This is because the normalization is essentially a linear transformation. So, it only re-produces the original data in smaller scale without actually changing the original characteristics. Therefore, it is striking that in Fig. 2c and d, the similar LFNN–EPFs were produced with the ones in Fig. 2a and b. As expected, the correlation coefficients r were also similar, being 0.35 for IR and 0.76 for Raman in Fig. 2c and d. However, when log–log transformation as in Fig. 2e and f was applied to untransformed original data of Fig. 2a and b, a considerably smoother data was obtained. Here the phrase "smoother" was used to point that, although the inherent fluctuations still unavoidably existed, the ratios of sharp pick intensities became much smaller. For example, between two sharp picks of 1000 and 200 au, the ratio is 5, while for logarithms the ratio is only 3/2.3 = 1.3. Accordingly, in Fig. 2e and f of smoother data, the LFNN–EPFs fittings were considerably better, in particular for IR intensities. The correlation coefficients r were 0.503 for IR and

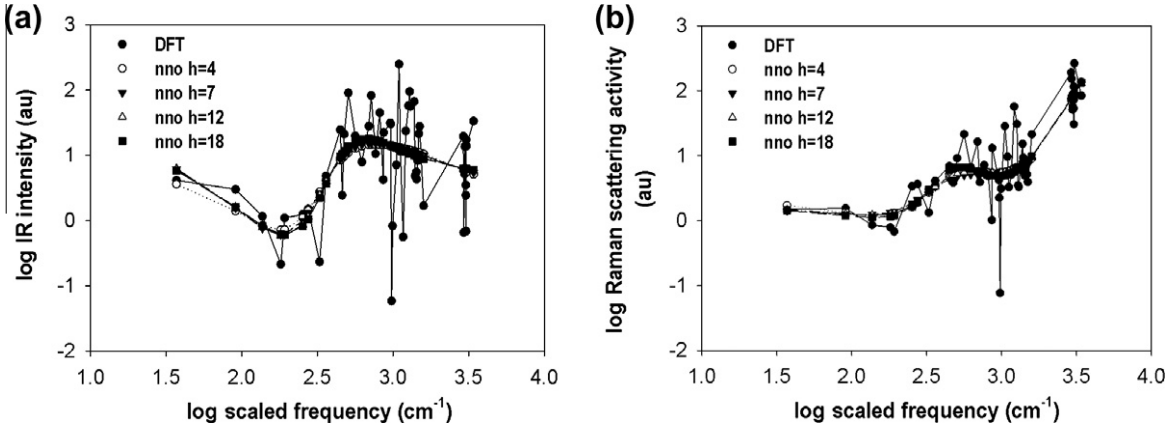

Fig. 3. For 2MPMSA, with increasing number of hidden layer neuron number h, LFNN–EPF *train set* fittings (*nno*) of DFT vibrational (IR and Raman) intensities versus scaled frequency. (a) IR-log data (b) Raman-log data.

0.800 for Raman in Fig. 2e and f. Notice that log–log $r = 0.503$ and 0.797 are respectively 1.44 and 1.05 times higher than their original data correlation coefficients. Also, notice that when compared with Fig. 2a and b, the LFNN–EPFs in Fig. 2e and f extract better the general tendency of average fluctuations hidden in the physical vibrational data. It is reasonable to think that LFNNs with significantly increased number of hidden neurons h fit the sharply fluctuating data much better. To test this hypothesis, in Fig. 3a and b, the *nno* log–log fittings with $h = 4, 7, 12, 18$ were given. Contrary to what was expected, as can be seen from Fig. 3a and b, there was no significant improvement in using more hidden neurons. Indeed, as few as $h = 4$ hidden neurons are sufficient to extract the general average trend embedded in the fluctuating vibrational intensities. Quantitatively speaking, the *MSE* errors, and correlation coefficients (the IR and Raman) for $h = 4, 7, 12, 18$ were almost the same, 0.03, 0.500 and 0.800 respectively. Also, almost the same set of error values and correlation coefficients were observed for three hidden layer LFNN with $h = 4, 7, 12, 18$. Briefly, the inherent sharp fluctuations cannot be approximated by the LFNN–EPFs within arbitrarily low fitting errors.

We also constructed suitable train set LFNN–EPFs for 3MPMSA DFT vibrational intensities of Ref. [18]. Similar numerical results with *nno* fittings of 2MPMSA above were also obtained for 3MPMSA when measured by *MSE* errors, and correlation coefficients of IR and Raman vibrational intensities (see Figs. S1a–f and S2a and b) (Supplementary Material). More clearly, in Fig. S1a–f

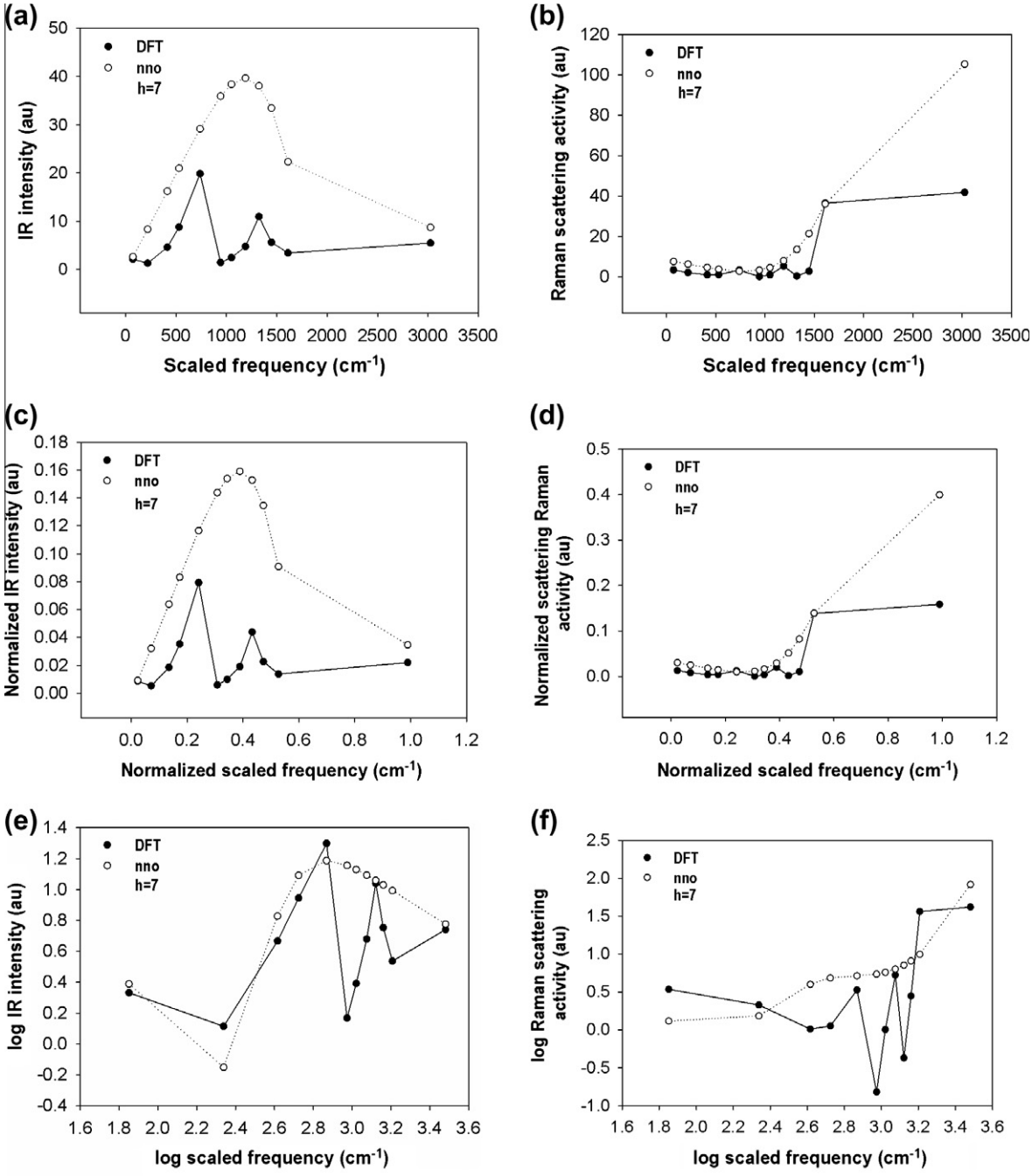

Fig. 4. For 2MPMSA, vibrational (IR and Raman) intensities versus scaled frequency. DFT based vibrational intensity calculations (DFT) and their LFNN–EPF *test set* predictions (*nno*). *h*: hidden layer neuron number. (a) IR-original untransformed data (b) Raman-original untransformed data (c) IR-normalized data (d) Raman-normalized data (e) IR-log data (f) Raman-log data.

(Supplementary Material), $h = 7$ for average *MSE* was 0.04. The correlation coefficients were r (IR) = 0.34 (original and normalized data), r (Raman) = 0.760 (original and normalized data), r (IR) = 0.45–0.50 for log–log data and r (Raman) = 0.753. Also, in Fig. S2a and b (Supplementary Material), the log–log *MSE* errors, and the correlation coefficients (the IR and Raman) for $h = 4, 7, 12, 18$ were nearly the same, namely log–log $MSE \simeq 0.05$, r (IR) $\simeq 0.500$ and r (Raman) $\simeq 0.760$ respectively. Therefore, the general remarks made for 2MPMSA LFNN–EPFs must be also valid for 3MPMSA LFNN–EPFs. Briefly, both 2MPMSA and 3MPMSA train LFNN–EPFs extract the hidden average general trend in vibrational intensities to medium and higher level of satisfaction.

4.2. Consistency of the constructed LFNN–EFPs: Test set predictions

If the train set LFNN–EPFs of 2MPMSA and 3MPMSA do not be further tested over "yet-to-be measured" DFT values, these fitted EPFs cannot be used consistently over a desired range of DFT values. In other words, if the train vibrational intensity LFNNs predict satisfactorily previously unseen *test set data*, then we say that the LFNN have successfully generalized the data, indicating consistent estimations. If the estimations are consistent with the test data values, then the LFNNs can be taken as suitable LFNN–EPFs. For 2MPMSA, in Fig. 4a–f, the corresponding test set predictions of Fig. 2a–f were given. The test set *MSE* errors were nearly the same with their train set values, while the test correlation coefficients were slightly lower than their corresponding train set values. In particular, log–log r (Raman) values for tests were nearly 0.485 (medium range), when compared their train set r values 0.750 (satisfactorily high). Still, as far as the ability to extract the general average of fluctuating intensities, the test set LFNN–EPFs of Fig. 4a–f can still be considered satisfactory to medium or higher lever. Also for 2MPMSA, in Fig. 5a and b, the test set log–log IR and Raman intensity LFNN–EPFs predictions were given for increasing h. The corresponding train set fittings were given in Fig. 3a and b. Again, without taking into account the predictions over sharp picks, we see that the test set LFNN–EPFs of Fig. 5a and b have actually succeeded in extracting the general average trend of vibrational intensities. Combining all the test set predictions for both Figs. 4a–f and 5a and b, we say that judging from medium to higher correlation coefficients of test set predictions (0.5–0.75) of the train sets concerned, the constructed LFNN–EPFs are consistent. However, a persistent argument concerning the sharp pick predictions must be stressed once again here. If the data were smooth (not containing extremely sharp picks) and dense (not containing sparsely taken data points) enough, then the correlation coefficients for consistent test sets predictions had to be expected to be higher than 0.95 in value. However, for the vibrational data of difficult-to-fit characteristics here, the best the LFNN–EPFs can (/must) do is to extract the average fluctuating behavior. This naturally produces lower correlation coefficients than 0.95. Lastly, what was said about 2MPMSA test set predictions generally applies also for 3MPMSA test set predictions in Figs. S3a–f and S4a and b (Supplementary Material). More clearly, the test *MSE* error values, as well as correlation coefficients of 3MPSSA were nearly the same with those of 2MPMSA, differing only about 5% in value. Briefly, the hidden general average behavior of sharply fluctuating vibrational intensities can be extracted by suitably constructed LFNN–EPPs to medium or higher degree of satisfaction.

5. Conclusions, potential applications and future remarks

In this paper, as a novel approach, satisfactorily consistent LFNN–EPFs were constructed for DFT vibrational spectra intensities of the molecules 2MPMSA and 3MPMSA. Because, the vibrational intensities are important physical quantities directly related to distributions of the electric charges in a molecule, the following major conclusions and potential applications can be drawn. Some future remarks for improving the predictive power of LFNN–EPFs are also given.

5.1. Conclusions

1. The average behavior of difficult-to-fit DFT vibrational intensities of both 2MPSA and 3MPMSA were fitted and predicted by the LFNN–EPFs to a satisfactory degree. This indicates that the physical laws embedded in the DFT vibrational data can be faithfully extracted by the LFNN–EPFs.
2. Even for vibrational intensities containing sparse data points and extremely sharp picks, suitable LFNN–EPFs for average vibrational intensity behaviors can be successfully constructed. Because, the LFNN–EPFs are of explicit functional form, they can be analytically used to obtain further vibrational intensity related molecular parameters.

5.2. Potential applications and future remarks

1. For any unknown molecule of significant complexity, provided that vibrational intensities are measured over sufficiently dense frequency intervals, explicit form of consistent LFNN–EPFs can

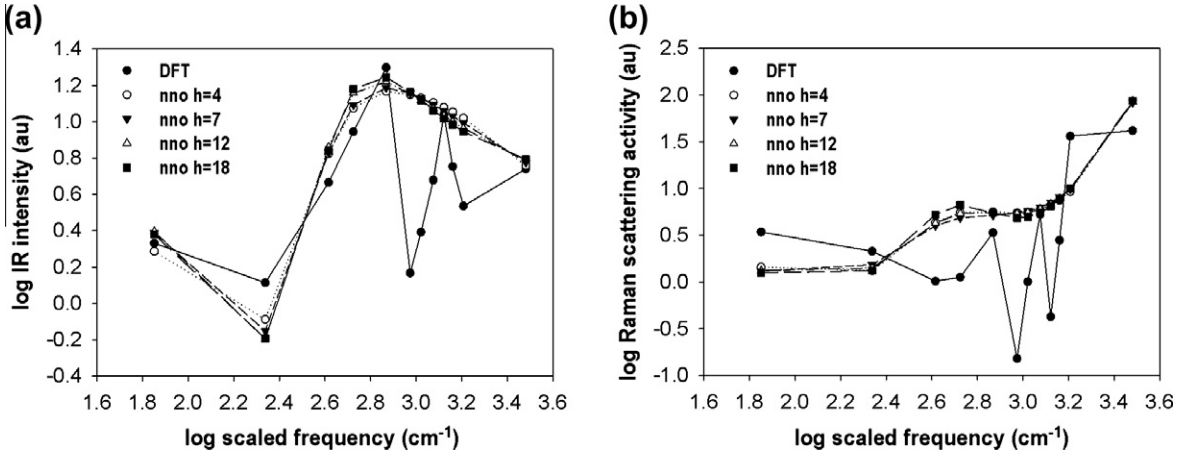

Fig. 5. For 2MPMSA, with increasing number of hidden layer neuron number h, LFNN–EPF *test set* predictions (*nno*) of DFT vibrational (IR and Raman) intensities versus scaled frequency. (a) IR-log data (b) Raman-log data.

be constructed to a very high satisfaction. Then, by various mathematical tools such as differentiation, integration, minimization, these vibrational LFNN–EPFs can be used to estimate the electronic charge distributions of the molecule. Moreover, these estimations can be compared and combined with those of theoretical DFT APT calculations to contribute to the identification of the molecule.

2. The IR spectra LFNN–EPFs can be combined together with the findings of other existing IR spectra estimations by neural networks, for instance radial distribution function (RDF) coded IR spectra estimation mentioned in the introduction of this paper.
3. To improve the predictive power of LFNN, one can consider other methods, for example partial least squares or another black box methods like support vector machines. One can then compare different methods with LFNN findings.
4. As a very general and persisting issue, when a lot of local minima emerge, the minimization of Eq. (8) can be tackled with other optimization algorithms, for example Metropolis–Hasting method, Monte-Carlo sampling, or simulated annealing.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.molstruc.2011.10.021.

References

[1] B.S. Galabov, T. Dudev, in: J.R. Durig (Ed.), Vibrational Spectra and Structure: Vibrational Intensities, vol. 22, Elsevier Science BV, 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] S. Haykin, Neural Networks: A Comprehensive Foundation, second ed., Prentice-Hall, New Jersey, 1999.
[14] N. Yildiz, Phys. Lett. A 345 (1–3) (2005) 69.
[15] N. Yildiz, S.E. San, Ö. Polat, Opt. Commun. 284 (2011) 2173.
[16] N. Yildiz, Ö. Polat, S.E. San, N. Kaya, J. Mol. Struct. 991 (2011) 127.
[17] K. Hornik, M. Stinchcombe, H. White, Neural Networks 2 (1989) 359.
[18] M. Karabacak, M. Cinar, M. Kurt, J. Mol. Struct. 968 (2010) 108.
[19] J. Gasteiger, J. Sadowski, J. Schuur, P. Selzer, L. Steinhauer, V. Steinhauer, J. Chem. Inform. Comput. Sci. 36 (1996) 1030.
[20] M.C. Hemmer, V. Steinhauer, J. Gasteiger, Vib. Spectrosc. 19 (1999) 151.
[21] P. Selzer, J. Gasteiger, H. Thomas, R. Salzer, Chem. Eur. J. 6 (5) (2000) 920.
[22] I. Duponchel, C. Ruchebusch, J.P. Huvenne, P. Legrand, J. Mol. Struct. 480–481 (1999) 551.
[23] R. Goodacre, Vib. Spectrosc. 32 (2003) 33.
[24] C.B. Cai, H.W. Yang, B. Wang, Y.Y. Tao, M.Q. Wen, L. Xu, Vib. Spectrosc. 56 (2011) 202.
[25] R. Hecht-Nielsen, Appl. Opt. 26 (1976) 4979.
[26] M. Gussoni, C. Castiglioni, G. Zerbi, in: J. Chalmers, P. Griffiths (Eds.), Handbook of Vibrational Spectroscopy, vol. 3, John Wiley & Sons, Chichester, UK, 2001, p. 2040 (and references therein).
[27] N. Yildiz, S.E. San, O. Köysal, Opt. Commun. 283 (2010) 3271.