Generalized Regression Neural Network Based Quantitative Structure-Property Relationship for the Prediction of Absorption Energy

Hui Li
School of Computer Science and Information Technology
Northeast Normal University
Changchun, Jilin, China
lihui@nenu.edu.cn

Yinghua Lu*
School of Computer Science and Information Technology
Northeast Normal University
Changchun, Jilin, China
luyh@nenu.edu.cn

Ting Gao
School of Computer Science and
Information Technology
Northeast Normal University
Changchun, Jilin, China

Hongzhi Li
School of Computer Science and
Information Technology
Northeast Normal University
Changchun, Jilin, China

Lihong Hu
School of Computer Science and
Information Technology
Northeast Normal University
Changchun, Jilin, China

***Abstract*—Generalized Regression Neural Network (GRNN) was used to develop a quantitative structure-property relationship (QSPR) model to improve the calculation accuracy of density functional theory (DFT). The model has been applied to evaluate optical absorption energies of 150 organic molecules based on the molecular descriptors. The entire dataset was divided into a training set of 120 molecules and a test set of 30 molecules according to the method, termed SPXY (Sample set Partitioning based on joint x–y distances), extended Kennard and Stones (KS) algorithm according to their differences in both x (instrumental responses) and y (predicted parameter) spaces in the calculation of inter-sample distances. Back-propagation neural network with SPXY partitioning algorithm (BPNN-SPXY) and GRNN with KS algorithm (GRNN-KS) were also utilized to construct model to compare with the results obtained by GRNN with SPXY algorithm (GRNN-SPXY). The root-mean-square errors in absorption energy predictions for the whole data set given by DFT, BPNN-SPXY , GRNN-KS and GRNN-SPXY were 0.47, 0.21, 0.17 and 0.13, respectively. The GRNN-SPXY prediction results are in good agreement with the experimental value of absorption energy.**

Keywords-Generalized Regression Neural Network; Sample subset partitioning; Kennard and Stones algorithm; Absorption energy; Density functional theory

I. INTRODUCTION

The main task of modern quantum chemistry (QC) is the generation of approximate solutions to the Schrödinger equation for molecular systems [1-3]. This task is performed almost exclusively by expanding the molecular orbitals in a Gaussian basis set located on the nuclei. The representation of the molecular electron density and, as a consequence, all molecular properties is improved as the basis set is enlarged. In practical calculations, however, the basis set is usually far from complete, meaning that modern QC calculations are greatly influenced by the incompleteness of the chosen basis set. In fact, Becke's three-parameter hybrid method (B3LYP) [4-6] has been widely recognized as a cost-effective method and has been successfully applied to many chemically interesting systems. However, the calculation results are not accurate enough for all systems, especially for large systems [7]. This limitation is caused by the electron correlation inclusion obtained and finite basis sets chosen in practical computations. To resolve this, simple yet efficient way to correct such errors is desired.

Quantitative structure–property relationship (QSPR) provides an alternative method for the prediction of impact sensitivity using descriptors derived solely from the molecular structure to fit experimental data. The QSPR method is based on the assumption that the variation of the behavior of the compounds, as expressed by any measured physicochemical properties, can be correlated with numerical changes in structural features of all compounds, termed "molecular descriptors" [8, 9]. The advantage of this method lies in the fact that it requires only the knowledge of the chemical structure and is not dependent on any experimental properties. Once a correlation is established and validated, it can be applicable for the prediction of the property of new compounds that have not been synthesized or found. Thus the QSPR method can expedite the process of development of new molecules and materials with desired properties [10]. Artificial neural network (ANN) techniques have recently been used with success to map the problem of solving complex physical differential equations to statistical models. In recent years, Chen and co-workers have developed a NN-based approach to improve the B3LYP heats of formation of 180 organic molecules [11] and Gibbs energy of formation [12, 13], etc., Wu and Xu [14, 15] recently introduced a similar NN-based method called X1.The latter was shown to drastically improve the prediction of B3LYP heats of formation. Our group

The authors gratefully acknowledge financial support from NSFC (20903020), 973 Program (2009CB623605), and the Science and Technology Development Planning of Jilin Province (20110364, 20100114).

developed a genetic algorithm and neural network approach to improve the calculation accuracy of absorption energies for organic molecules [16, 17]. The result is quite promising.

In the present work, we develop a nonlinear generalized regression neural network (GRNN) model for computing molecular absorption energies in chemical compound space. The accurate calculation for the electronic absorption energy is one of the important topics in computational chemistry. The entire dataset including 150 organic molecules was divided into a training set of 120 organic molecules and a test set of 30 organic molecules according to the method, termed SPXY (Sample set Partitioning based on joint x–y distances), extended Kennard and Stones (KS) algorithm according to their differences in both x (instrumental responses) and y (predicted parameter) spaces in the calculation of inter-sample distances [18]. The raw calculated absorption energies are evaluated by TDDFT/B3LYP method. In addition, the contributions of the involved descriptors to the models were discussed in detail.

II. Materials and method

A. Dataset Partitioning

In this work, a set of 150 organic molecules collected is investigated. Their experimental absorption energies are accurately known [16]. Kennard and Stones algorithm [19] has been widely used for splitting datasets into two subsets. This algorithm starts by finding two samples, based on the input variables that are the farthest apart from each other. For this purpose, the algorithm employs the Euclidean distances *dx*(*p*, *q*) between the x-vectors of each pair (*p*, *q*) of samples calculated as

$$d_x(p,q)=\sqrt{\sum_{j=1}^{J}[x_p(j)-x_q(j)]^2}\,,\ p,q\in[1,N] \qquad (1)$$

where $x_p(j)$ and $x_q(j)$ are the responses at the *j*th output for samples *p* and *q*, respectively.

These two samples are removed from the original dataset and put into the training set. Then, the remaining sample farthest away from the selected two samples is again included in training set. This step is repeated until the desired number of samples has been selected in the training set.

A shortcoming of KS in the multivariate calibration context lies in the fact that the statistics of the dependent variable (y) are not taken into account. It could be argued that the inclusion of y-information in the selection process might result in a more effective distribution of calibration samples in the multidimensional space, thus improving the predictive ability and robustness of the resulting model. The SPXY method extends the KS algorithm by encompassing both x- and y-differences in the calculation of inter-sample distances [18]. Such a distance $d_y(p, q)$ can be calculated for each pair of samples *p* and *q* as

$$d_y(p,q)=\sqrt{(y_p-y_q)^2}=\left|y_p-y_q\right|.p,q\in[1,N] \qquad (2)$$

In order to assign equal importance to the distribution of the samples in the x and y spaces, distances $d_x(p, q)$ and $d_y(p,q)$ are divided by their maximum values in the data set. In this manner, a normalized x*y* distance is calculated as

$$d_{xy}(p,q)=\frac{d_x(p,q)}{\max_{p,q\in[1,N]}d_x(p,q)}+\frac{d_y(p,q)}{\max_{p,q\in[1,N]}d_y(p,q)}.p,q\in[1,N] \qquad (3)$$

A stepwise selection procedure similar to the KS algorithm can then be applied with $d_{xy}(p, q)$ instead of $d_x(p, q)$ alone [18].

Using SPXY algorithm, the entire dataset was divided into two subsets: a training set of 120 organic molecules, and a test set including the remaining organic molecules.

B. Descriptors Selection

The most important issue is to select the proper physical descriptors, which are to be used as the input for the GRNN model. As stated in our previous work [16], the calculated value of excited-state electronic energy contains the essence of exact values of absorption energy, and is an obvious choice of the primary descriptor. Other physical descriptors are selected according to their correlation to absorption energies. The physical properties, such as the number of electrons, the oscillator strength, the dipole moment, the number of double bonds, the HOMO-LUMO energy gap, the orbital energy gap corresponding the dominant configuration of the excited state, the corresponding transitional coefficient, the number of aromatic rings, have been chosen as the other physical descriptors. The physical parameters are calculated using the quantum chemistry program package. The parameters are optimized geometrically for all degrees of freedom with the DFT-B3LYP method in the polarization basis set 6-31G (d) level and the frequency calculations are used to confirm the stable structure.

C. GRNN Method

Generalized regression neural network is proposed by the American scholar DF Specht [20]. The method uses vertical basis function as the basis of the hidden units to form the hidden layers. The hidden layer transforms the input vectors from the low-dimensional input data into a high dimensional space so that the problem can be separated linearly in the high dimensional space. It is good at function approximation and the network finally converges to the optimized regression plane which contains the most samples. It can predict well even with very few sample data and can handle the instability in the data.

Nonlinear models are then developed by submitting the selected descriptors to a GRNN model. The number of input neurons is nine descriptors. One output neuron is used to represent the experimental absorption energies. To avoid overtraining, 5-fold cross-validation was applied to train the GRNN model.

III. Results and Discussion

In order to evaluate the effectiveness of the GRNN model for evaluating the optical absorption energies of 150 organic molecule problem, it was compared with the TDDFT/B3LYP/6-31G (d) calculation and Back-propagation neural network (BPNN) on the same problems. We find that the best value of spread for GRNN set 0.21 results in the best output by 5-fold cross validation. The entire dataset was divided into a training set of 120 organic molecules and a test set of 30 organic molecules according to KS and SPXY algorithm, respectively.

The raw calculated absorption energies values versus their experimental data are shown in Fig.1 (a). The vertical coordinate is the experimental absorption energies, and the horizontal coordinate is the calculated values by DFT. The dashed line is where the vertical and horizontal values are equal. In Fig.1 (b), the horizontal coordinates are for the BPNN corrected absorption energies by SPXY partitioning algorithm (BPNN-SPXY). In Fig.1 (c) and 1 (d), the horizontal coordinates are for the GRNN corrected absorption energies by KS (GRNN-KS) and SPXY partitioning algorithm (GRNN-SPXY), respectively. Compared to the raw calculated values, the GRNN-SPXY corrected results are much closer to the experimental values. This can be shown clearly by the error analysis performed for all 150 organic molecules.

The root-mean-square (rms) errors in absorption energy predictions for the whole dataset given by DFT, BPNN-SPXY, GRNN-KS and GRNN-SPXY were 0.47, 0.21, 0.17 and 0.13, respectively (see table 1).

TABLE I. RMS deviation of TDDFT/ B3LYP/6-31G (d), BPNN-SPXY, GRNN-KS and GRNN-SPXY Corrections (in eV)

	TDDFT/B3LYP/ 6-31G(d)	**BPN-SPXY**	**GRNN-KS**	**GRNN-SPXY**
Absorption energies	0.47	0.21	0.17	0.13

As regards the comparison of BPNN-SPXY, GRNN-KS, and GRNN-SPXY performances, it can be seen that GRNN-SPXY yielded the smallest rms errors for absorption energies. The prediction results are in good agreement with the experimental value of absorption energy; also, the results reveal the superiority of the GRNN-SPXY over BPNN-SPXY and GRNN-KS models. The GRNN approach improved DFT calculation results.

IV. Conclusions

To summarize, GRNN was used to develop a QSPR model for the prediction of absorption energies of 150 organic molecules. BPNN was also utilized to construct model to compare with the results obtained by GRNN. The GRNN approach improved DFT calculation results and reduced the RMS deviations from 0.47 to 0.17 and 0.13 eV by KS and SPXY partitioning algorithm, respectively, while BPNN-SPXY is 0.21 eV. Very satisfactory results were obtained with the proposed methods. Additionally, models using GRNN with SPXY partitioning algorithm based on the same set of descriptors produced even better models with a good predictive ability than GRNN with KS partitioning algorithm models. This study of the QSPR model shows that the GRNN-SPXY is a very promising tool in the prediction of absorption energy when compared with BPNN-SPXY. The approach can be used as the approximation of experimental results where the experimental results are unavailable or uncertain. Furthermore, the proposed approach can also be extended to other QSPR investigations.

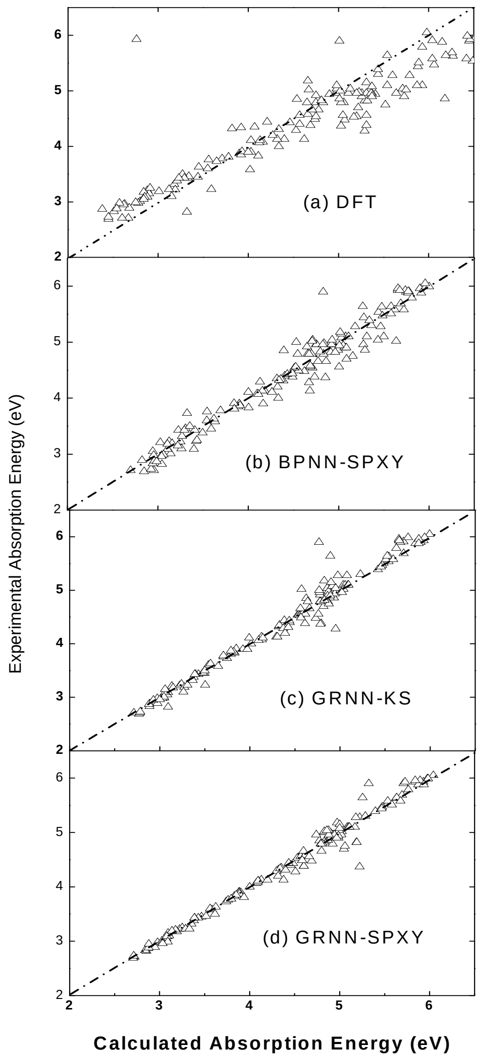

Figure 1. Calculated absorption energies versus experimental absorption energies for all 150 molecules.

References

[1] F. Jensen, The Magnitude of Intramolecular Basis Set Superposition Error, Chem. Phys. Lett., vol. 261, pp. 633-636, 1996.

[2] F. Jensen, Introduction to Computational Chemistry, John Wiley &Sons, 1999.

[3] S. M . Bachrach, Computational Organic Chemistry, Wiley-Interscience, 2007.

[4] A. D. Becke, Density-functional thermochemistry III. The role of exact exchange, J Chem Phys, vol 98, pp. 5648-5652, 1993.

[5] C Lee, W. Yang and R. G. Parr, Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density, Phys Rev B, vol. 37, pp. 785-789, 1988.

[6] A. D. Becke, Density-functional exchange-energy approximation with correct asymptotic behavior, Phys Rev A, vol. 38, pp. 3098-3100, 1988.

[7] K. K. Irikura and D. J. Frurip, Computational thermochemistry: prediction and estimation of molecular thermodynamics, American Chemical Society: Washington, DC, 1998.

[8] X.J. Yao, Y. W. Wang, X. Y. Zhang, R. S. Zhang, M. C. Liu, Z. D. Hu, B. T. Fan, Radial basis function neural network-based QSPR for the prediction of critical temperature. Chemom. Intell. Lab. Syst., vol. 62, pp. 217-225, 2002.

[9] J. Xu, L. Wang, L. Wang, X. Shen, W. Xu, QSPR study of Setschenow constants of organic compounds using MLR, ANN and SVM analyses, J. Comput. Chem., vol. 32, pp. 3241–3252, 2011.

[10] J. Xu, L. J. Zhu, D. Fang, L. X. Wang, S. L. Xiao, L. Liu, W. L. Xu, QSPR studies of impact sensitivity of nitro energetic compounds using three-dimensional descriptors, Journal of Molecular Graphics and Modelling,vol. 36, pp. 10–19, 2012.

[11] L. H. Hu, X. J. Wang, L. H. Wong, and G. H. Chen, Combined first-principles calculation and neural-network correction approach for heat of formation, J Chem Phys., vol. 119, pp. 11501-11507, 2003.

[12] X. J. Wang, L. H. Wong, L. H. Hu, C. Y. Chan, Z. M. Su, and G. H. Chen, Improving the accuracy of density functional theory calculation: the statistical correction approach, J Phys Chem A, vol. 108, pp. 8514-8525, 2004.

[13] X. J. Wang, L. H. Hu, L. H. Wong, and G. H. Chen, A combined first-principles calculation and Neural Networks correction approach for evaluating Gibbs energy of formation, Mol Simul, vol. 30, pp. 9 -15, 2004.

[14] J. M. Wu and X. Xu, The X1 method for accurate and efficient prediction of heats of formation, J Chem Phys, vol. 127, pp. 214105–214113, 2007.

[15] J. M. Wu and X. Xu, Improving the B3LYP bond energies by using the X1 method, J Chem Phys, vol 129, pp.164103-1-164103-11, 2008.

[16] H. Li, L. L. Shi, M. Zhang, Z. M. Su, X. J. Wang, L. H. Hu, and G. H. Chen, Improving the accuracy of density-functional theory calculation: The genetic algorithm and neural network approach, J Chem Phys, vol. 126, pp.144101-1-144101-8, 2007.

[17] T. Gao, L. L. Shi, H. B. Li, S. S. Zhao, H. Li, S. L. Sun, Z. M. Su and Y. H. Lu, Improving the accuracy of low level quantum chemical calculation for absorption energies: the genetic algorithm and neural network approach, Phys Chem Chem Phys., vol. 11, pp. 5124-5129, 2009.

[18] R. K. Galvão, M. C. Araujo, G. E. José, M. J. Pontes, E. C. Silva, T. C. Saldanha, A method for calibration and validation subset partitioning, Talanta, vol. 67, pp. 736–740, 2005.

[19] R. W. Kennard, L. A. Stone, Computer aided design of experiments, Technomet-rics, vol. 11, pp. 137-148, 1969.

[20] D. F. Specht, The General Regression Neural network-Rediscovered, vol. 6, pp. 1033-1034, 1993.