Improving the accuracy of density-functional theory calculation: The genetic algorithm and neural network approach

Hui Li, LiLi Shi, Min Zhang, and Zhongmin Su[a)]
Institute of Functional Material Chemistry, Faculty of Chemistry, Northeast Normal University, Changchun 130024, People's Republic of China

XiuJun Wang, LiHong Hu, and GuanHua Chen
Department of Chemistry, The University of Hong Kong, Hong Kong, China

(Received 5 June 2006; accepted 16 February 2007; published online 9 April 2007)

The combination of genetic algorithm and neural network approach (GANN) has been developed to improve the calculation accuracy of density functional theory. As a demonstration, this combined quantum mechanical calculation and GANN correction approach has been applied to evaluate the optical absorption energies of 150 organic molecules. The neural network approach reduces the root-mean-square (rms) deviation of the calculated absorption energies of 150 organic molecules from 0.47 to 0.22 eV for the TDDFT/B3LYP/6-31G(d) calculation, and the newly developed GANN correction approach reduces the rms deviation to 0.16 eV. © *2007 American Institute of Physics*. [DOI: 10.1063/1.2715579]

I. INTRODUCTION

First-principles quantum mechanical methods have become indispensable research tools in chemistry, condensed-matter physics, materials science, and molecular biology.[1–3] Experimentalists rely increasingly on these methods to interpret their experimental findings. Despite their successes, first-principles quantum mechanical methods are often not quantitatively accurate enough to predict the results of experimental measurements, in particular, on large systems. This limitation is caused by the inherent approximations adopted in first-principles methods. Because of computational cost, electron correlation has always been a difficult obstacle for first-principles calculations. Finite basis sets chosen in practical computations are not able to cover entire physical space and this inadequacy introduces also inherent computational errors.[4] The accuracy of a density functional theory (DFT) calculation is mainly determined by the exchange-correlation functional being employed,[2] whose exact form is, however, unknown. Therefore, better methods are required.

Nevertheless, the results of first-principles quantum mechanical calculation can capture the essence of physics. For instance, the calculated results, despite that their absolute values may agree poorly with measurements, are usually of the same tendency among different molecules as their experimental counterparts. The quantitative discrepancy between the calculated and experimental results depends predominantly on the property of primary interest and, to a less extent, also on other related properties of the material. There exists thus a sort of quantitative relation between the calculated and experimental results, as the aforementioned approximations to a large extent contribute to the systematic errors of specified first-principles methods. Although it is exceedingly difficult to be determined from the first principles, the quantitative relationship can be obtained empirically.

Recently, Chen and co-workers proposed a neural network based correction method DFT-NEURON to determine the quantitative relationship between the experimental data and the first-principles calculation results.[5–8] The determined relation will subsequently be used to eliminate the systematic deviations of the calculated results on the optical absorption energy.[5] Hutchison *et al.*[9] evaluated the absorption energies of 60 heterocyclic organic molecules using Zerner's intermediate neglect of differential overlap/configuration interaction singles (ZINDO/CIS), ZINDO/random phase approximation (ZINDO/RPA), Hartree-Fock/CIS (HF/CIS), HF/RPA, and time-dependent density functional theory/RPA (TDDFT/RPA) calculations. They concluded that TDDFT/CIS and TDDFT/RPA methods yield relatively accurate results upon linear regression fit. Chen and co-workers employ the TDDFT/B3LYP calculation to evaluate the absorption energies for those 60 organic molecules, too. Then the raw calculated absorption energies are corrected by the DFT-NEURON method and the multiple linear regression (MLR) correction approaches. They concluded that the DFT-NEURON method yield more accurate results. In their scheme, the raw calculated absorption energy is the primary descriptor. The number of electrons N_e in a molecule is explicitly included as an inputting physical descriptor. The oscillator strength O_s is a measure of absorption strength and is selected as the third and last descriptor. After the neural network correction, the root-mean-square (rms) deviation of the calculated absorption energies of 60 organic molecules is reduced from 0.33 to 0.09 eV for the TDDFT/B3LYP/6-31G(d) calculation. Their result is quite promising.

In neural network calculation, the result is determined on the synaptic weights that are determined iteratively by the

[a)]Author to whom correspondence should be addressed. Tel.: +86-431-5099108; Fax: +86-431-5684009; Electronic mail: zmsu@nenu.edu.cn

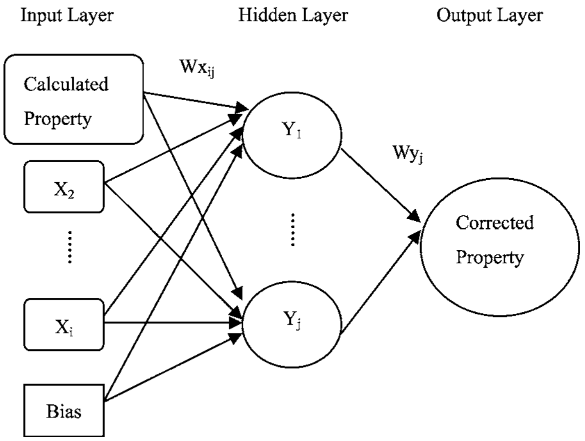


FIG. 1. The structure of our neural networks.

neural network training procedure. The synaptic weights include all properties of the neural networks. So the different synaptic weights of the neural networks may result in the different outputs. However, for conventional neural network algorithms, their initial synaptic weights are determined by the educated guess or randomly. These usually lead to slow convergence and poor performance. Also generally the final synaptic weights might be trapped in local optimal solution because of nonlinear multiextreme object function. So an improved procedure to determine the synaptic weights is desirable.

Genetic algorithm[10] (GA) is an efficient, parallel, and full search method with its inherent virtues of robustness, parallel, and self-adaptive characteristics. It is applicable for searching the optimization result in the large search space. It has been applied in many fields. Genetic algorithm uses biologically derived techniques such as inheritance, natural selection, and crossover. Genetic algorithm is typically implemented as a computer simulation in which a population of abstract representations (called chromosomes) of candidate solutions (called individuals) to an optimization problem evolves toward better solutions. So we adopt it to optimize the synaptic weights of neural networks.[11–13]

In the present work, we propose a genetic algorithm and neural network (GANN) approach to improve the calculation accuracy of absorption energies for 150 organic molecules. The raw calculated absorption energies are evaluated by TDDFT/B3LYP method. In this GANN approach, GA is adopted in searching the optimal initial synaptic weights for neural networks of prespecified topology, while back propagation (BP) is employed in further training the neural networks to find the optimal final synaptic weights. Most of the 150 molecules we considered in this paper are organic conjugated molecules.

II. DESCRIPTION OF GANN APPROACH

It is proved that neural networks of three-layer architecture can mimic any function.[14] We adopt the three-layer architecture for our neural networks (see Fig. 1). This architecture includes an input layer consisting of inputs from the physical descriptors $(X_1, X_2 \ldots, X_m)$ and a bias, a hidden layer containing a number of hidden neurons $(Y_1, \ldots, Y_n)$, and an output layer that outputs the corrected value for the property of interest. The numbers of descriptors and hidden neurons are to be determined. The most important issue is to select the proper physical descriptors, which are to be used as the input for the neural networks. The calculated value of absorption energy contains the essence of exact values of absorption energy and is thus an obvious choice of the primary descriptor for correcting values of absorption energy. Other physical descriptors are selected according to their correlation to optical absorption energies. If it is closely related to absorption energies, the property is chosen as a physical descriptor; otherwise, it is not. For the same set of oligomers, when the number of repeating units is small (for example, 1 or 2), the raw calculated absorption energies are higher than the experimental counterparts, and when the number of repeating units is large, the calculated absorption energies redshift strongly compared to their experimental counterparts. In other words, the oligomer size correlates strongly with the deviation between the raw result and experimental counterparts. The number of electrons N_t is thus taken as the second physical descriptor. The oscillator strength O_s is a measure of absorption magnitude and is selected as the third descriptor. The dipole moment D_m correlated with the oscillator strength and is taken as the fourth descriptor. The number of double bonds N_{db} is selected as the fifth descriptor to reflect the chemical structure of a molecule. The highest occupied molecular orbital–lowest unoccupied molecular orbital energy gap E_g affects the longest absorption spectrum and is selected as the sixth descriptor. The orbital energy gap corresponding to the dominant configuration of the excited state E_0 determined the absorption energies and is selected as the seventh descriptor. The corresponding transitional coefficient T_c is thus selected as the eighth descriptor. The number of aromatic rings N_a is selected as the ninth descriptor to reflect the conjugating degree.

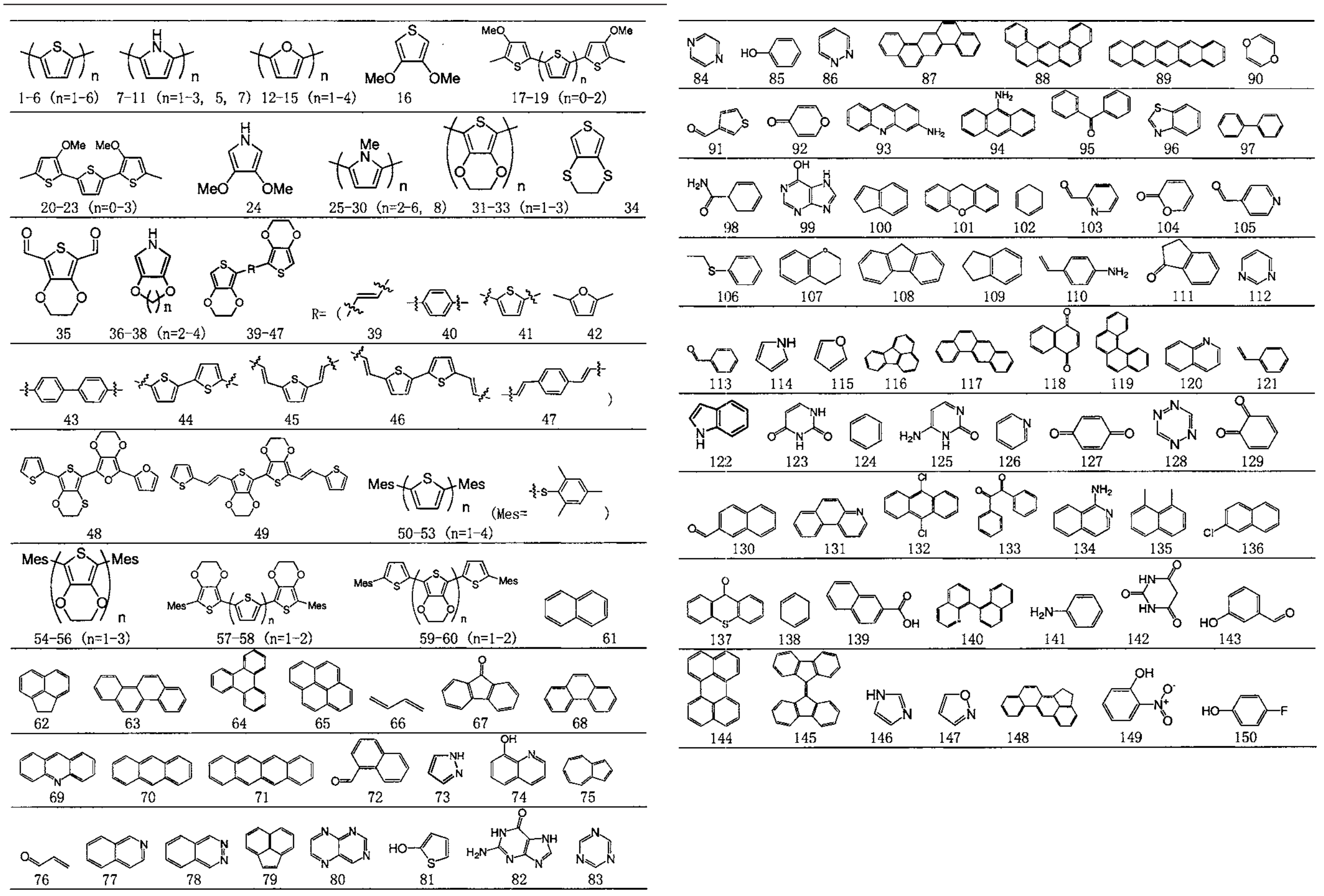
The number of neurons in the hidden layer is varied from 1 to 10 to decide the optimal structure of our neural networks. We find that the hidden layer containing five neurons yields the best overall results. Therefore, the 10-5-1 structure is adopted for our neural networks. The input values at the input layers, x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, and x_{10}, are scaled with the raw calculated absorption energy, N_t, O_s, D_m, N_{db}, E_g, E_0, T_c, N_a, and bias, respectively. The synaptic weights $\{Wx_{ij}\}$s connect the input descriptors $\{X_i\}$ and the hidden neurons $\{Y_j\}$ while $\{Wy_j\}$s connect the hidden neurons and the output Z $(i=1,\ldots,10;\ \textit{and}\ j=1,\ldots,5)$. The error BP learning algorithm[15] is used to optimize the synaptic weights. The output Z is related to the input $\{X_i\}$ as follows:

$$Z = \sum_{j=1}^{5} Wy_j \tan sig\left(\sum_{j=1}^{10} Wx_{ij} X_i\right), \tag{1}$$

where $\tan sig(v) = 2/(1+\exp(-2v)) - 1$. During the calculation, expect for the bias the input and output values are scaled. The scaled values are between -1 and 1.

In neural network calculation, their initial synaptic weights are determined randomly. The synaptic weights include all the properties of the neural networks. The different synaptic weights of the neural networks may result in the

TABLE I. The structure of the 150 organic molecules.



different outputs. These usually lead to slow convergence and poor performance. Also generally the final synaptic weights might be trapped in local optimal solution because of nonlinear multiextreme object function.

"In general, any abstract task to be accomplished can be thought of as solving a problem, which, in turn, can be perceived as a search through a space of potential solutions,"[16] Whereas traditional search techniques use characteristics of the problem to determine the next sampling point (e.g., gradients, Hessians, linearity, and continuity), stochastic search techniques make no such assumptions. Instead, the next sampled point(s) is (are) determined based on stochastic sampling/decision rules rather than a set of deterministic decision rules. Genetic algorithms, more intelligent stochastic search techniques than random restart, improve upon random restart by mimicking the evolutionary process through the use of a "survival of the fittest" strategy. In general, the fittest individuals of any population tend to reproduce and pass their genes on the the next generation, thus improving successive generations. However, some of the worst individuals do, by chance, survive and also reproduce. Genetic algorithm has been shown to solve linear and nonlinear problems by exploring all regions of the state space and exponentially exploiting promising areas through mutation, crossover, and selection operations applied to individuals in the population.[17] So we adopt the GA to optimize the initial values of the synaptic weights.

The following are the procedures where the synaptic weights are optimized by the GANN approach.

Step 1. Set generation counter $i=0$.

Step 2. Create the initial population, Pop(i), by generating N random the initial values of the neural network's synaptic weights.

Step 3. Determine the fitness of each individual (i.e., each initial values of the network's synaptic weights) by an evaluation function in the population.

Step 4. Increment to the next generation, $i=i+1$.

Step 5. Create the new population, Pop(i), by

(a) selecting N individuals stochastically based on the fitness from the previous population, Pop($i-1$), and

(b) randomly selecting R ($R<N$) parents to produce children through the application of genetic operators (crossover and mutation operation).

Step 6. Evaluate the fitness of the newly formed children as in step 3.

Step 7. If i is less than the maximum number of generations to be considered, go to step 4.

Step 8. The best individual is decoded to the values of the network's synaptic weights.

Step 9. The BP algorithm is used to further train the neural networks.

For any genetic algorithm, a chromosome representation is needed to describe each solution or individual in the population of interest. The representation scheme determines how the problem is structured and which genetic operators are used. Each individual (or chromosome) is made up of a sequence of genes from a certain alphabet. An alphabet could be binary numbers, real (floating point) numbers, integer numbers, symbols, matrices, etc. A floating point representation is used in the GA for the initial values of the synaptic weights. An individual in the GA population would represent all the synaptic weights of the neural networks. Each variable Wx_{ij} and Wy_j takes on real value

$$\text{Individual} \quad (Wx_{11}, Wx_{12}, \ldots, Wx_{1j}, Wx_{21}, Wx_{22}, ..., Wx_{2j}, ..., Wx_{i1}, Wx_{i2}, \ldots, Wx_{ij}, Wy_1, Wy_2 \ldots, Wy_j)$$
$$(i = 1, \ldots, 10; j = 1, \ldots, 5).$$

Once the initial population (of size N) is randomly created, each individual is evaluated using an evaluation function to determine its fitness value. Evaluation functions of many forms can be used, subject to the minimal requirement that the function can map the population into a partially ordered set. The evaluation functions for optimizing the initial synaptic weights of neural networks are expressed as follows:

$$f_k = 1/E_k, \tag{2}$$

$$E_k = \sum_{l=1}^{n} (Z_{lk} - dz_{lk})^2, \tag{3}$$

where f_k is the fitness of one individual ($k=1, \ldots, N$); n is the number of overall training examples. Z_{ik} is the training examples' outputs for each random individual; dz_{lk} represents the desired neural network response for the output neuron. E_k is the sum of squared errors.

After the population (of size N) has been evaluated, a new population of size N individuals is selected from the previous generation. The selection of individuals to produce successive generations plays an extremely important role in a genetic algorithm. The normalized geometric ranking scheme[17] was used in the GA procedure described in this paper. The individuals in the population are ranked from best to worst according to their fitness value. Then, each individual is assigned a probability of selection based on the normalized geometric distribution,

$$P[\text{selecting } i\text{th individual}] = q'(1-q)^{r-1},$$

where $q' = q/(1-(1-q)^N)$, q is the probability of selecting the best individual, r is the rank of the individual (where 1 is the best), and N is the size of the population.

After the new population is selected, each genetic operator is applied a discrete number of times to create new solutions based on existing solutions in the population. Mutation and crossover are the two basic types of genetic operators.[16] Mutation operators tend to make small random changes in one parent to form one child. Crossover operators combine information from two parents to form two offspring such that the two children contain a "likeness" (a set of building blocks) from each parent. The application of these two basic types of operators, and their derivatives, depends on the chromosome representation used. Two float operators described by Michalewicz[16] were employed to work with the floating point representation: arithmetic crossover and nonuniform mutation. The "arithmetic crossover" operator produces a complimentary pair of linear combinations produced from random proportions of the parents. The "nonuniform mutation" operator randomly selects one of the variables from a parent and sets it equal to a random number from a nonuniform distribution.[16]

The GA moves from generation to generation, repeating steps 4–7 until the termination criterion is met. The stopping criterion used is the specification of the maximum number of generations to iterate through.

After the GA stopped, the BP algorithm is used to further train the neural networks in our method. Many researchers have shown that GAs perform well for a global search but perform very pooly in a localized search.[16,18] So we proposed this combined GANN approach, where the good approximation performance of NN and effective and robust evolutionary searching ability of GA are applied in hybrid sense. That is, GA is adopted in searching the optimal initial synaptic weights, while BP is employed in further training the neural networks to find the optimal final synaptic weights. This combined algorithm can overcome the GA's shortcomings of the poorly localized adjustive ability. The GANN approach combines the good qualities of the GA and the BP neural networks and is shown to be effective at solving the problem.

III. RESULTS AND DISCUSSION

In order to evaluate the effectiveness of the GANN approach for evaluating the optical absorption energies of 150 organic molecule problem, it was compared with the TDDFT/B3LYP/6-31G(d) calculation and the BP neural network approach on the same problems.

All experimental data are randomly divided into a training set (120 molecules) and a testing set (30 molecules). The structures of 150 organic molecules are shown in Table I. The experimental absorption energies and differences between the calculated and experimental absorption energies for the BP neural network (BPN) and GANN correction results are tabulated in Table II.

TABLE II. The experimental absorption energies and the differences between the calculated and experimental values of 150 molecules (in eV).

No.	Expt.[a]	Deviation[b]	Deviation[c]	Deviation[d]
1	5.10	0.83	0.35	0.17
2	4.11	−0.08	−0.12	−0.16
3	3.50	−0.23	0.01	−0.01
4	3.18	−0.34	0.04	0.11
5	2.98	−0.41	−0.01	0.00
6[e]	2.87	−0.49	−0.75	0.01
7	5.96	0.8	0.22	0.07
8	4.49	0.25	0.03	0.04
9	3.91	0.02	0.04	0.00
10[e]	3.38	−0.17	0.09	−0.36
11	3.25	−0.34	0.00	0.09
12	5.93	0.58	0.03	−0.16
13[e]	4.40	0.17	0.00	0.00
14	3.78	−0.07	0.04	0.03
15[e]	3.43	−0.2	0.08	0.06
16[e]	4.90	0.47	−0.06	0.16
17	3.76	−0.2	−0.14	−0.20
18	3.19	−0.18	0.19	0.14
19	2.96	−0.33	0.17	0.19
20[e]	3.81	−0.03	−0.04	0.03
21	3.23	−0.11	0.03	−0.05
22[e]	2.99	−0.24	0.10	0.06
23[e]	2.83	−0.32	−0.61	−0.23
24	5.58	0.46	−0.09	0.06
25	4.96	0.01	−0.29	−0.31
26	4.58	0.08	−0.03	−0.07
27	4.44	−0.23	−0.16	−0.04
28	4.35	−0.28	−0.05	−0.01
29	4.34	−0.42	−0.06	0.07
30	4.32	−0.5	−0.23	−0.03
31	4.82	0.5	−0.02	−0.02
32	3.87	0.03	−0.12	0.02
33	3.10	0.05	0.04	−0.02
34	4.38	0.31	−0.08	0.08
35	3.83	0.28	0.03	−0.01
36[e]	5.58	0.84	0.22	0.28
37	5.93	0.53	−0.10	0.01
38	5.90	0.55	−0.09	0.05
39	3.45	−0.01	0.04	−0.01
40[e]	3.60	−0.05	−0.18	0.04
41	3.32	−0.14	−0.12	0.08
42	3.43	−0.14	−0.17	0.03
43	3.63	−0.18	−0.06	0.10
44	3.01	−0.21	0.02	−0.02
45	2.89	−0.21	0.01	−0.06
46	2.73	−0.28	−0.02	0.07
47	3.15	−0.25	−0.23	−0.04
48	2.98	−0.2	−0.01	−0.15
49	2.69	−0.24	0.00	0.00
50	4.11	0.02	0.16	0.05
51	3.46	−0.12	0.18	−0.05
52	3.21	−0.33	0.09	0.08
53	2.95	−0.33	0.20	−0.13
54	4.07	0.05	0.10	0.01
55	3.43	−0.1	0.09	−0.02
56	3.09	−0.24	−0.09	0.14
57	3.09	−0.21	0.02	−0.08
58	2.71	−0.11	0.00	−0.06
59[e]	3.05	−0.21	0.13	−0.14
60[e]	2.88	−0.34	0.00	0.26
61	5.63	0.64	0.13	−0.09
62	5.47	0.59	−0.05	−0.10
63	4.66	0.05	0.30	0.25
64	4.83	−0.02	−0.10	−0.09
65	5.18	−0.52	−0.22	−0.15
66	5.91	0.13	−0.21	0.01
67	4.85	−0.31	−0.15	−0.24
68	4.96	0.15	−0.02	−0.11
69	4.96	0.41	0.00	−0.18
70	4.96	0.35	−0.04	−0.03
71	4.54	0.21	0.00	−0.01
72[e]	5.88	0.27	−0.12	−0.09
73	5.90	0.9	0.08	−0.12
74	3.90	0.13	0.16	0.10
75[e]	4.53	0.64	0.20	0.36
76[e]	5.99	0.44	−0.37	−0.23
77[e]	5.64	0.55	0.00	0.04
78	5.69	0.57	−0.08	−0.01
79	3.85	0.07	0.15	0.01
80	4.13	0.49	0.33	0.17
81	5.64	−0.1	−0.22	0.05
82	4.54	0.66	0.00	0.08
83	4.46	0.01	−0.20	0.00
84[e]	4.75	0.77	0.38	0.60
85	4.70	0.52	0.16	0.44
86	5.02	0.73	0.17	−0.03
87	4.13	0.27	−0.07	0.08
88	4.08	0.08	−0.15	−0.07
89	4.00	0.34	0.00	0.02
90	4.96	0.3	0.23	0.06
91	4.94	0.47	−0.07	−0.31
92	5.04	0.67	0.10	−0.04
93	2.82	0.5	0.10	−0.03
94	4.28	1.01	0.57	0.22
95	4.92	−0.17	−0.20	−0.16
96[e]	4.96	0.24	−0.11	−0.28
97	5.02	−0.35	−0.52	−0.10
98	4.96	0.4	−0.12	−0.01
99	4.98	0.25	−0.12	−0.22
100[e]	5.00	0.02	−0.31	−0.14
101	5.04	−0.06	−0.17	0.00
102	4.79	0.04	0.01	0.02
103	4.79	0.27	−0.04	0.04
104	4.29	0.24	0.22	0.05
105	4.79	−0.14	−0.35	−0.05
106	4.82	0.47	0.11	−0.06
107	4.37	0.66	0.32	0.24
108	4.66	0.12	−0.07	0.07
109	4.56	0.75	0.28	−0.04
110	4.48	0.59	0.18	0.17
111	5.08	0.29	−0.18	−0.20
112	5.10	0.77	0.20	0.16
113	5.15	0.16	−0.23	−0.05
114	5.93	−3.17	−0.04	−0.04
115	5.96	0.55	−0.03	−0.05
116	7.65	0.14	−0.17	0.00
117[e]	4.31	0.03	−0.30	−0.16
118[e]	3.73	−0.08	0.02	−0.39
119	4.43	0.03	0.00	−0.04

TABLE II. (*Continued.*)

No.	Expt.[a]	Deviation[b]	Deviation[c]	Deviation[d]
120	6.05	−0.07	−0.35	0.07
121[e]	5.04	0.09	−0.29	0.07
122[e]	5.79	0.14	−0.34	−0.55
123	4.38	0.93	0.48	0.35
124	5.10	0.44	−0.01	0.08
125[e]	5.28	0.32	−0.26	−0.34
126[e]	4.96	0.68	0.17	0.23
127[e]	5.10	−0.12	−0.53	0.02
128	4.90	0.83	0.34	−0.05
129	3.22	−0.03	0.00	−0.02
130	3.23	0.36	0.00	−0.11
131	3.58	0.44	0.12	0.22
132	4.79	−0.04	−0.09	0.07
133	4.79	0.24	0.18	0.09
134	4.13	0.19	0.11	0.09
135	5.44	0.44	−0.10	0.08
136[e]	5.51	0.38	−0.14	−0.18
137	4.86	0.15	0.16	−0.07
138	5.54	0.95	0.12	0.09
139[e]	5.39	0.05	−0.23	−0.11
140	5.96	0.1	0.25	0.06
141[e]	5.28	0.51	−0.06	0.01
142[e]	4.86	1.32	0.53	0.00
143	4.90	0.36	−0.06	−0.02
144	5.54	−0.04	−0.03	0.00
145	6.05	−0.04	−0.13	0.00
146	5.99	0.75	−0.07	0.07
147	5.88	0.71	−0.05	0.04
148	6.78	0.07	0.29	−0.02
149	4.56	0.49	0.18	0.20
150	5.90	−0.89	−1.16	−0.80

[a] Experimental data.

[b] Differences between the raw calculated and experimental values.

[c] Differences between calculated and experimental values for DFT-BPN calculation.

[d] Differences between calculated and experimental values for DFT-GANN calculation.

[e] Molecules belong to the testing set.

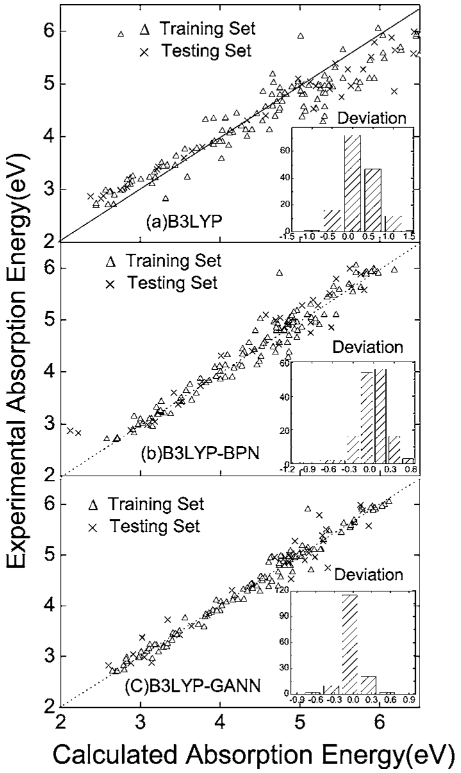


FIG. 2. Calculated absorption energies vs experimental absorption energies for all 150 molecules. Part (a) is for raw calculated absorption energies from the DFT approach. Part (b) is for neural network corrected absorption energies for the BPN approach. Part (c) is for the combined neural network and genetic algorithm corrected absorption energies for the GANN approach. Triangles (△) are for the training set and crosses (×) are for the testing set. Insets are the histograms for the differences between the experimental and calculated absorption energies. All values are in units of eV.

The raw calculated absorption energy values versus their experimental data are shown in Fig. 2(a). The vertical coordinate is the experimental absorption spectrum energies, and the horizontal coordinate is the calculated values by DFT. The dashed line is where the vertical and horizontal values are equal. In Figs. 2(b) and 2(c) the horizontal coordinates are for the BPN-corrected and GANN-corrected absorption energies, respectively. Compared to the raw calculated values, the GANN-corrected results are much closer to the experimental values for both training and testing sets. More importantly, the systematic deviation in Figs. 2(a)–2(c) is eliminated. The insets are the histograms for the deviations of various approaches in Figs. 2(a)–2(c). Obviously, the raw calculated absorption energies have large systematic deviations, while the BPN- and GANN-corrected absorption energies have smaller systematic deviations. The rms deviation of the BPN approach is slightly larger than that of the GANN approach. This can be shown clearly by the error analysis performed for all 150 organic molecules. For the training set, the rms deviations before and after the BPN correction are 0.48 and 0.20 eV, respectively, while for the testing set, they are 0.41 and 0.28 eV, respectively. For the GANN correction, the rms deviations of the training and testing sets are reduced to 0.14 and 0.24 eV (see Table III), respectively. The GANN approach improved DFT calculation results in both the training set and the testing set separately. But there exist some data that the relative errors are bigger in the testing set. The reason lies that this kind of sample data is fewer in the training set and the features cannot be extracted in the training procedure of neural networks. The prediction accuracy of this approach can be further improved as more and better experimental data are available. The consistency between the training and testing sets implied that the GANN results could indeed predict the absorption energy with higher accuracy than BPN.

The TDDFT/B3LYP/6-31G(d) calculations are carried out to evaluate the absorption energies of the 150 organic molecules, and their overall resulting rms deviation from the

TABLE III. rms deviation of TDDFT/B3LYP/6-31G(d), DFT-BPN, and DFT-GANN corrections (in eV).

	TDDFT/B3LYP/6-31G(d)	BPN	GANN
Training set	0.48	0.20	0.14
Testing set	0.41	0.28	0.24
Overall	0.47	0.22	0.16

TABLE IV. Optimized values of synaptic weights Wx_{ij} and Wy_{j}..

Weights	y_1	y_2	y_3	y_4	y_5
Wx_{1j}	−0.6038	−2.1719	−0.5519	−0.3159	−0.3137
Wx_{2j}	1.4869	−3.5033	−3.927	−0.6994	−0.7107
Wx_{3j}	0.3134	−0.3081	−1.1594	1.0523	1.029
Wx_{4j}	0.1430	2.5483	2.5185	−0.0792	−0.0556
Wx_{5j}	−1.2349	5.9349	6.8976	−1.6979	−1.6171
Wx_{6j}	11.602	−18.2867	2.8302	2.9062	3.0625
Wx_{7j}	−10.1388	20.9544	−3.0306	−2.9728	−3.1000
Wx_{8j}	−0.0937	−0.6150	−0.7567	0.6438	0.6275
Wx_{9j}	0.1659	−1.8207	−1.9974	1.2436	1.2049
Bias	1.8475	3.1329	1.3302	−0.2798	−0.2594
Wy_j	−1.5316	−0.7562	−0.8168	−79.6885	81.0453

experimental data is 0.47 eV. Upon the traditional BP neural network correction approach, the rms deviation of the calculated absorption energies of 150 organic molecules is reduced from 0.47 to 0.22 eV for the TDDFT/B3LYP/6-31G(d) calculation. With the GANN correction, the rms deviation is reduced from 0.47 to 0.16 eV (see Table III).

In our GANN approach, the BP algorithm is used to further train the neural networks after the GA stopped. Only the GA, the rms deviation result is 0.31 eV for the 150 organic molecules using the optimal initial synaptic weights of neural networks. The rms deviation is reduced to 0.16 eV using the optimal final synaptic weights by the further training of the BP algorithm. The improvement shows that the GANN approach combining the good qualities of the GA and the BP algorithm is effective at solving the problem.

In the procedure of GA running, we find that the best individual has no obvious improvement and the generation is approximately 100 by many experiments. So the maximum number of GA generations is set at 100 and size of the population is $N=150$. The final values of the weights Wx_{ij} and Wy_j optimized by the GANN approach are shown in Table IV. We use these weights to evaluate the absorption energies.

IV. CONCLUSION

To summarize, we have developed a promising new GANN approach to improve the results of first-principles quantum mechanical calculations. In this GANN approach, GA is adopted in searching the optimal initial synaptic weights for neural networks of prespecified topology, while BP is employed in further training the neural networks to find the optimal final synaptic weights. It is employed to reduce the errors of calculated absorption energy of 150 molecules. This combined GANN correction approach avoids being trapped at local minima of the traditional BPN approach, thus leading to improved DFT calculation results as compared to those of BPN. The rms deviation of TDDFT calculated absorption of our 150 molecules is reduced from 0.47 to 0.16 eV. Simulation results and comparisons demonstrate the feasibility and effectiveness of the approach. The accuracy of this approach can be further improved as more and better experimental data are available. The larger the experimental database is, the more accurately the GANN approach predicts. In principle, the accuracy of our approach is limited only by the precision and size of the experimental database.

Besides the absorption energy, our GANN approach can be generalized to calculate other properties such as heat of formation, ionization energy, dissociation energy, etc. The GANN approach combines the good qualities of the GA and the BP neural networks and is shown to be effective at solving the problem. This combined GANN approach may be employed practically as predictive tools in materials research and design.

ACKNOWLEDGMENTS

Support from the Hong Kong Research Grant Council (HKU 7013/05P) is gratefully acknowledged. The authors acknowledge the financial support from the National Natural Science Foundation of China (Project Nos. 20373009 and 20573016) and the Specialized Research Fund for the Doctoral Program in Higher Education Institutions of the Ministry of Education of China (No. 20030183063) for financial support. Science Foundation for Young Teachers of Northeast Normal University (No. 20061004) is also greatly appreciated.

[1] H. F. Schaefer III, *Methods of Electronic Structure Theory* (Plenum, New York, 1977), and references therein.

[2] R. G. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules* (Oxford University Press, New York, 1989), and references therein.

[3] C. J. Cramer, *Essentials of Computational Chemistry: Theories and Models* (Wiley, West Sussex, England, 2002).

[4] K. K. Irikura and D. J. Frurip, *Computational Thermochemistry: Prediction and Estimation of Molecular Thermodynamics* (American Chemical Society, Washington, DC, 1998).

[5] L. H. Hu, X. J. Wang, L. H. Wong, and G. H. Chen, J. Chem. Phys. **119**, 11501 (2003).

[6] X. J. Wang, L. H. Wong, L. H. Hu, C. Y. Chan, Z. M. Su, and G. H. Chen, J. Phys. Chem. A **108**, 8514 (2004).

[7] X. J. Wang, L. H. Hu, L. H. Wong, and G. H. Chen, Mol. Simul. **30**, 9 (2004).

[8] X. Zheng, L. H. Hu, X. J. Wang, and G. H. Chen, Chem. Phys. Lett. **390**, 186 (2004).

[9] G. R. Hutchinson, M. A. Ratner, and T. J. Marks, J. Phys. Chem. A **106**, 10596 (2002).

[10] J. H. Holland, *Adaptation in Natural and Artificial Systems* (The University of Michigan Press, Ann Arbor, 1975).

[11] J. J. David and J. F. Frenzel, IEEE Expert **8**, 26 (1993).

[12] J. D. Schaffer, in *Proceedings of the Workshop on Combinations of Genetic Algorithms and Neural Networks*, edited by D. Whitley (The IEEE Computer Society, Los Alamitos, CA, 1992).

[13] D. Floreano and J. Urzelai, Neural Networks **13**, 431 (2000).

[14] R. Hecht-Nielsen, *Neurocomputing* (Addison-Welsley, Reading, MA, 1990).

[15] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, Nature (London) **323**, 533 (1986).

[16] Z. Michalewicz, *Genetic* Algorithms+Data Structures=Evolution *Programs* (Springer-Verlag, New York, 1992).

[17] J. A. Joines and C. R. Houck, Proceedings of the IEEE Conference On Evolutionary Computation, Orlando, FL, 1994 (IEEE), p. 579.

[18] L. Davis, *The Handbook of Genetic Algorithms* (Van Nostrand Reinhold, New York, 1991).