Correcting the systematic error of the density functional theory calculation: the alternate combination approach of genetic algorithm and neural network*

Wang Ting-Ting(王婷婷)[a)b)], Li Wen-Long(李文龙)[a)], Chen Zhang-Hui(陈章辉)[b)†], and Miao Ling(缪 灵)[a)]

[a)] *Department of Electronic Science and Technology, Huazhong University of Science and Technology, Wuhan* 430074, *China*
[b)] *State Key Laboratory for Superlattices and Microstructures, Institute of Semiconductors, Chinese Academy of Sciences, Beijing* 100083, *China*

(Received 3 December 2009; revised manuscript received 31 December 2009)

The alternate combinational approach of genetic algorithm and neural network (AGANN) has been presented to correct the systematic error of the density functional theory (DFT) calculation. It treats the DFT as a black box and models the error through external statistical information. As a demonstration, the AGANN method has been applied in the correction of the lattice energies from the DFT calculation for 72 metal halides and hydrides. Through the AGANN correction, the mean absolute value of the relative errors of the calculated lattice energies to the experimental values decreases from 4.93% to 1.20% in the testing set. For comparison, the neural network approach reduces the mean value to 2.56%. And for the common combinational approach of genetic algorithm and neural network, the value drops to 2.15%. The multiple linear regression method almost has no correction effect here.

Keywords: density functional theory, neural network, genetic algorithm, alternate combination
PACC: 6430, 8730

1. Introduction

Density functional theory (DFT) has become a leading tool to predict the electronic properties of many-body systems, and is widely used in condensed-matter physics, computational physics, computational chemistry and molecular biology.[1−3] In spite of great success, the theory has some inherent disadvantages and its calculation results are often not accurate enough.[4−6] This is due to the approximation methods adopted in the calculation process. The main problem of DFT originates in the fact that the exact form for exchange–correlation functional is not known yet, and some popular approximate forms such as local-density approximation and generalized gradient approximation will bring deviation to the calculation.[4−6] At the same time, DFT often uses pseudo-potential methods to reduce computational burden at the cost of accuracy. The value of energy cutoff, the sampling of k-points and other numerical parameters used in the calculation also bring deviation to the result.

However, the deviation between calculations and experiments is systematic error. The calculated results for different structures usually show the same tendency as their experimental values. If we can find the law of the deviation, then the DFT calculated results can be corrected through the law for new structures. Figure 1 shows the process of such corrected methods. Firstly DFT is used to calculate the properties of many compounds whose experimental properties are already known. Then some data mining methods such as linear regression or neural network are adopted to find the relationship between calculation values and experimental ones. Such a relationship then can be used to correct the calculation results of new compounds. It can be found that this method is an external correction and post-processing strategy. Compared with some internal correction, this method treats the DFT as a black box and models the error through external statistical information. When the error's statistical law of one kind of system is established, its corrected speed will be higher than internal correction. However, it may be difficult to analyse the physical meaning of such an external correction, although physical insight is not necessary for correction. At the same time, it is a great challenge for data mining methods to discover the error's law from few data.

*Project supported by the National Basic Research Program of China (973 Program) (Grant No. G2009CB929300) and the National Natural Science Foundation of China (Grant No. 60521001 and 60925016).
†Corresponding author. E-mail: zhanghuichen88@gmail.com
© 2010 Chinese Physical Society and IOP Publishing Ltd

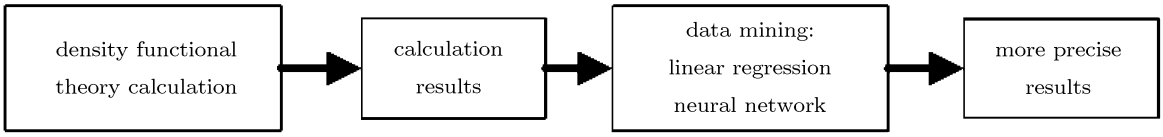


Fig. 1. The flow chart of the correction process.

There have been some researchers paying attention to such an external corrected strategy. For example, Hu *et al.*[7−9] used neural network (NN) to establish the relationship between the calculated heat of formation and experimental data of 180 organic molecules. Duan *et al.*[10,11] proposed a linear regression method to correct the prediction of heat of formation from Hartree–Fock/DFT calculation. NN, as one of artificial intelligence methods, is a computational model whose architecture essentially mimics the knowledge acquisition and organisational skills of human brain.[12,13] It has strong mapping ability. But the common NN generates initial synaptic weights and thresholds randomly and adopts error back-propagation (BP) algorithm, which usually leads to local optimum and poor performance. Meanwhile, genetic algorithm (GA) has strong ability of searching for the global optimum.[14] In consideration of these facts, a combinational approach of genetic algorithm and neural network (GANN) was proposed to improve the accuracy of DFT calculation for the absorption energies of 150 organic molecules by Li and co-workers.[15] The combinational approach firstly adopted GA to optimise the initial values of the NN weights, and then used the BP algorithm to further train the network. Compared with the common NN, the combinational approach can obtain better corrected results.

However, the common GANN method still has some weakness for adopting GA only one time to optimise the network's weights and thresholds. In order to find the coarse value of the global optimum in one time, GA needs large population, which will result in heavy computational burden.[16−19] Besides, GA with large population is easy to run into local optimum and has difficulty in obtaining the global optimum in one time. In consideration of these disadvantages, this paper presents an alternate combinational approach of genetic algorithm and neural network (AGANN). This alternate approach adopts GA with small population to optimise the NN firstly, and then all the corresponding network architectures of the GA population are further trained by BP algorithm. Afterwards, these network architectures are imported to GA for global optimisation a second time, and then these architectures optimised by GA are transmitted to BP algorithm again, and the cycle is repeated until obtaining the global optimum. Compared with the common GANN, the alternate combinational approach adopts GA with small population, and transmits network architectures between GA and BP repeatedly.

The following parts will give the AGANN method step by step, which is then used to correct the systematic error of the lattice energies of the DFT calculation for 72 metal halides and hydrides. The corrected results are also compared with the corrections using the NN, the common GANN, and the multiple linear regression method.

2. Correction methods

2.1. The NN

An NN of three-layer architecture is adopted in this study.[12] It consists of an input layer containing input neurons of physical descriptors, a hidden layer with a certain number of hidden neurons, and an output layer which outputs the lattice energies. Figure 2 shows the network's architecture. The input physical descriptors and the number of hidden neurons are to be determined. The physical descriptors are selected according to their correlation to lattice energies. The calculated lattice energy $E_{\rm c}$ contains the largest information and is obviously the first descriptor. The number of cation electrons $N_{\rm c}$ and the number of anion electrons $N_{\rm a}$ are the primary factors that affect lattice energies, and thus are selected as the second and the third descriptor respectively. The chemical valence of the cation $V_{\rm c}$ reflects the strength and number of the bonds, and is the fourth descriptor. These four descriptors form the input layer. The number of hidden neurons needs to be identified by experiments. A great number of experiments show that the hidden layer with 8 neurons yields the best results. Hence the 4-8-1 architecture is used for the NN.

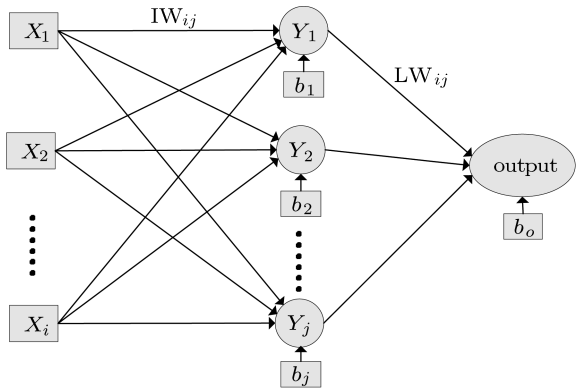


Fig. 2. The architecture of the NN. IW$_{ij}$ and LW$_{ij}$ indicate the link weights; b_1, b_2 ... b_j indicate the thresholds of the hidden neurons; b_o indicates the threshold of the output neuron.

The log-sigmoid function is selected as the transfer function of the hidden neurons and output neuron. And the weights and thresholds of the network are trained by the BP algorithm. To obtain better convergent speed, the input values from descriptors and the expected output values are normalised to 0–1.

2.2. The GANN

In the training process of the common NN, initial weights and thresholds are determined randomly, and different initial network parameters may lead to different final convergent results. These randomness and uncertainty will enlarge the difficulty of network training. At the same time, the common neural network adopts BP algorithm, which usually brings about local optimum. In view of the above fact, GA can be used to generate the coarse values of the global optimum of NN weights and thresholds. The combinational approach of GA and NN[15] is shown as follows:

Step 1 Create a large number of initial NN. Parameters of each network are generated randomly.

Step 2 Encode all of the networks to the population of GA. Each network corresponds to a chromosome (individual) of GA. Each parameter of a network corresponds to a gene of a chromosome.

Step 3 Use GA to evolve the population until the generation times reaches the maximum. The evolving of the population is implemented by a number of genetic operators such as fitness evaluation, selection, crossover, mutation.

Step 4 Select the best solution of the population and decode its corresponding chromosome to an NN network.

Step 5 Use BP algorithm to further train the NN until arriving at the optimum.

On the whole, GA is used in searching for the coarse value of the global optimum while BP is used to identify the accurate global optimum from the coarse value. Experiments show that the program yields the best result when the genetic parameters are set as follows: the number of the population's chromosomes is 100; the preserving probability of the fittest individuals is 0.1; the crossover probability is 0.8; the mutation probability is 0.05; the number of possibly mutated points is 10.

2.3. The AGANN

According to what has been shown above, there is only once to transmit the network architecture between GA and BP. In order to find the coarse global optimum in one single time, GA needs a large population, which will result in heavy computational burden. Besides, GA with a large population is easy to get into local optimum and is difficult to jump out. In view of this fact, an alternate combination approach of GA and NN is presented. It adopts GA with small population and transmits network architectures between GA and BP repeatedly. The main steps are shown as follows.

Step 1 Create a small number of initial neural networks. Parameters of every network are generated randomly.

Step 2 Encode all of the networks to the population of GA. Each network corresponds to a chromosome of GA. Each parameter of a network corresponds to a gene of a chromosome.

Step 3 Use GA to evolve the population until the generation times reaches the preset value. The evolving of the population is implemented by a number of genetic operators such as fitness evaluation, selection, crossover, mutation.

Step 4 Decode all the chromosomes to NNs.

Step 5 Use BP algorithm to further train all the NN until arriving at the preset iteration epochs.

Step 6 If the best architecture in all the networks arrives at the desired optimal value, end the program. If not, turn to step 2.

As shown above, GA is firstly used to find the coarse values of a number of local optimums while BP is then used to find the accurate values of these local optimums. But these accurate optimums are local, so GA is used again to jump out of these optimums to find more optimums, and BP is used to identify the accurate values of these new optimums, and repeat the cycle until obtaining the global optimum. On the

whole, the function of GA is to find the coarse values of local optimums and jump out of accurate local optimums, while the use of BP is to search for the accurate values of local optimums.

It can be found that the convergent speed and performance of the overall process are mainly influenced by the number of chromosomes, the preset generation times of GA, and the preset iteration epochs of BP. These values should neither be too large nor too small. For the number of chromosomes, a too large one needs long convergent time while a too small one cannot ensure the global optimum. For the preset generation times of GA, if its amount is too large, the evolution of the population will be too deep in every GA optimisation and the chromosomes tend to be completely the same, which will bring about difficulty in generating more different local optimums. But the program cannot find the local optimum with a too small amount. Similarly, a too large value of the preset iteration epochs of BP need more computational time while a too small one cannot satisfy the desire. These values need to be identified by many experiments. Experiments show that the program yields the best results when the number of chromosomes is 10, the preset generation times of GA is 2 and the preset iteration epochs of BP is 10. Other parameters are the same as the GANN method.

2.4. Multiple linear regression method

For comparison, this study also adopts multiple linear regression (MLR) method[7,8] to correct the result of DFT calculation. The model of MLR method is shown as follows:

$$E_{\rm e} = E_{\rm c} + a_1 N_{\rm c} + a_2 N_a + a_3 V_{\rm c} + a_4, \qquad (1)$$

where $E_{\rm e}$ is the experimental value of the lattice energy while $E_{\rm c}$ is the calculated value; $N_{\rm c}$, $N_{\rm a}$, and $\rm V_c$ refers to the number of cation electrons, the number of anion electrons and the chemical valence of the cation respectively; a_1, a_2, a_3 and a_4 are the coefficients.

3. Results and discussions

The 72 metal halides and hydrides are listed in Table 1. The DFT method is firstly used to calculate their lattice energies, which are then corrected by AGANN. The DFT method here is based on the generalised gradient approximation[20] and Perdew–Wang 91 exchange–correlation functional,[21] and adopts the plane wave pseudo-potential method.[22−24] For comparison, NN, GANN, and MLR methods are also applied in the correction respectively. All of the 72 experimental data[25] are randomly divided into a training set (60 crystals) and a testing set (12 crystals). Table 1 lists the experimental data and the corrected results of the four methods.

Table 1. The experimental lattice energies (kJ/mol) of 72 metal halides and hydrides, and the relative errors (%) of the DFT calculated values to the experimental values. And the relative errors (%) of the values corrected respectively by four kinds of methods to the experimental values.

No.	crystals	Expt.[a]	DFT[b]	DFT-NN[c]	DFT-GANN[d]	DFT-AGANN[e]	DFT-MLR[f]
1	CsBr	647	–0.8502	1.7207	0.9559	1.3195	0.1482
2	KBr	691	3.5162	0.8572	1.3525	0.4711	–5.6909
3	LiBr	820	4.0357	1.2184	0.6514	–0.0827	–7.5213
4	NaBr	754	4.5673	1.5011	0.8133	–0.2158	–5.9359
5	RbBr	668	3.1719	–4.6512	0.6895	–0.7711	–1.1066
6	CsCl	670	0.7089	2.8892	–2.7600	–2.1657	7.4739
7	KCl	720	3.7283	0.0671	–1.1209	0.1391	0.2902
8	LiCl	864	4.5196	–0.8954	–1.3724	0.2082	–1.9505
9	NaCl	790	5.2424	0.0820	–0.0353	0.4820	0.1376
10	RbCl	695	3.6145	–2.5627	–1.0617	–0.8100	5.0944
11	CsF	759	6.5575	–0.4035	0.6547	1.1152	14.8051
12[t]	KF	829	7.4773	–2.0330	1.1148	0.3233	6.5749
13	LiF	1049	6.9636	0.0331	–0.6748	0.0486	3.2813
14	NaF	930	8.1051	–0.0360	1.4964	–0.8979	5.6262
15	RbF	795	7.4241	0.4291	1.2578	1.2964	10.8906

Table 1. (*Continued*).

No.	crystals	Expt.[a]	DFT[b]	DFT-NN[c]	DFT-GANN[d]	DFT-AGANN[e]	DFT-MLR[f]
16	CsH	653	–12.1520	0.1869	0.7716	–0.4504	0.0797
17	CuH	1254	–4.4780	–0.0108	–0.4236	–0.0113	–2.1447
18	KH	713	–8.0142	–0.1982	–0.5843	0.8174	–6.6407
19	LiH	918	–4.2908	0.0275	0.7181	–0.3666	–6.6169
20[t]	NaH	807	–4.8609	0.0004	2.0271	1.5966	–5.5772
21[t]	RbH	684	–8.8977	3.1200	2.4840	0.8978	–2.3431
22	CsI	613	–2.0637	0.4736	0.4178	0.2844	–7.3502
23	KI	650	1.5148	–0.3344	–0.7561	0.8875	–14.2524
24[t]	LiI	764	3.3492	0.0872	3.1719	0.6690	–14.1421
25	NaI	705	3.4967	0.6062	–0.5703	–0.4384	–13.2494
26	RbI	632	3.3832	–1.2025	–0.6813	–0.9285	–7.2887
27[t]	$BeBr_2$	2914	3.3355	–4.0306	1.0698	–0.7034	–1.1637
28	$CoBr_2$	2643	9.6474	0.6831	0.6838	0.5139	6.3809
29	$FeBr_2$	2577	6.7056	–0.2394	–0.0562	–0.3039	3.2799
30	$HgBr_2$	2639	–10.3092	0.2907	–0.0139	0.0377	–9.6711
31	$MgBr_2$	2451	2.0857	–0.3581	0.1081	–0.0194	–2.6281
32	$NiBr_2$	2721	10.8231	0.1680	0.4305	0.3462	7.7218
33	$PdBr_2$	2751	6.3923	–1.3003	0.1203	0.0507	4.5985
34	$SnBr_2$	2256	1.8292	0.3240	0.1869	–0.2719	–0.0130
35	$CuCl_2$	2824	6.6797	–1.3063	–1.4559	–1.6920	5.1367
36	$NiCl_2$	2786	10.5180	1.5163	1.4975	1.3044	8.8841
37	$SrCl_2$	2170	1.8684	0.0714	–0.1845	0.1960	0.6678
38	CoF_2	3042	9.2009	–0.0040	–0.5925	–0.0927	8.2083
39	SrF_2	2513	4.7066	0.0316	0.3716	0.0441	4.3572
40	BaH_2	2133	–5.9989	0.0159	–0.0282	0.0980	–3.9579
41	BeH_2	3306	–2.2326	–0.0836	–0.0634	0.0315	–3.9778
42	CaH_2	2406	–5.8561	0.0136	0.0399	–0.1564	–6.9595
43[t]	MgH_2	2718	–3.2990	1.8315	–1.2214	0.2008	–4.8486
44	NdH_2	2394	–6.4105	0.5432	0.0327	0.0576	–4.2668
45	PrH_2	2405	–6.1681	–0.9487	–0.0143	–0.1358	–4.1151
46[t]	ScH_2	2744	–0.9909	2.6704	0.7466	1.2491	–1.8874
47[t]	SmH_2	2389	–8.8906	4.8342	–2.2343	1.2962	–6.5794
48	YH_2	2733	–4.1749	0.4435	–0.1348	0.1526	–3.7929
49	CaI_2	2087	–0.1189	–0.0566	0.0283	0.1701	–6.7708
50	CoI_2	2561	10.1197	0.7070	0.2058	0.3831	5.2310
51	FeI_2	2491	6.8579	–0.0311	–0.4348	–0.2378	1.7537
52	MgI_2	2340	2.4107	–0.1506	–0.0719	0.0212	–4.1875
53	NiI_2	2637	10.9631	–0.2792	–0.3777	–0.4016	6.2891
54	SnI_2	2206	1.5591	–0.0002	–0.3124	0.0024	–2.0867
55	$AlBr_3$	5360	2.0383	0.2764	–0.0022	0.0016	–0.7951
56	$AsBr_3$	5365	1.4477	–0.8840	–0.0241	–0.0120	–0.6575
57	$GaBr_3$	5569	2.1410	–0.2598	–0.0202	0.0418	0.0431
58	$InBr_3$	5117	–0.2641	–0.0234	–0.0027	0.0003	–1.8626
59	$RuBr_3$	5232	4.7831	1.0229	–0.0010	0.0004	3.0337
60	$ZrBr_3$	4758	2.4244	0.0040	–0.0036	0.0513	0.3371
61	$AlCl_3$	5513	2.2642	–0.0032	–0.0393	0.0221	0.2144
62	$BiCl_3$	4707	–0.1419	–0.1500	0.0018	0.0026	0.3522
63	YCl_3	4524	2.9483	0.1101	–0.0015	–0.0450	1.5691
64[t]	CrF_3	6065	7.8039	4.1738	3.3774	3.4214	6.5785
65	AlH_3	5969	–2.2705	–0.0092	0.0318	–0.0204	–3.5850
66[t]	YH_3	4910	–4.7667	–5.6294	4.5022	–2.7093	–5.3339
67[t]	AsI_3	5295	0.7603	–0.6243	–0.2894	–0.0243	–2.1067
68	GaI_3	5496	1.3292	–0.0491	0.0297	–0.0610	–1.5038
69	$SnBr_4$	8852	–0.3954	0.0054	0.0092	–0.0038	–1.7300
70	$SnCl_4$	8930	0.3600	–0.0140	–0.0143	–0.0091	–0.5277
71	$ZrCl_4$	8144	3.5913	0.0096	–0.0017	0.0186	2.3790
72[t]	TiI_4	8918	4.6974	1.6507	–3.6133	–1.3268	2.3258

[t]The crystals belong to the testing set.
[a]The experimental data taken from Ref. [12].
[b]The relative errors of the DFT calculation.
[c]The relative errors of the DFT calculation and NN correction.
[d]The relative errors of the DFT calculation and GANN correction.
[e]The relative errors of the DFT calculation and AGANN correction.
[f]The relative errors of the DFT calculation and MLR correction.

As listed in Table 1, the lattice energies of the DFT calculation have a certain deviation from experimental values. But lattice energies of most of the crystals all get improved and closer to the experimental values after being corrected by the method of NN, GANN or AGANN. Especially by the AGANN method, the value of nearly every kind of crystals gets improved greatly. However, the MLR method turns out to be ineffective in this correction, and the deviation between calculation and experiment does not get improved although changed after the correction. These can also be found from Table 2, which clearly shows the mean absolute values of the relative errors for the DFT calculation and the four correction methods. As seen from Table 2, the AGANN method has the best correction result. The mean absolute value of the relative errors of the training set decreases from the previous 4.54% to 0.35% after the AGANN correction, while for the testing sets, the mean value decreases from 4.93% to 1.20%. The GANN method has the second best effect, and the mean absolute values of the training and testing sets are reduced to 0.46% and 2.15% respectively. For the NN correction, the mean values of the training and testing sets are 0.55% and 2.56% respectively. The MLR method almost has no correction effect here, and there are no great changes after the correction for both the mean values of the training set and the testing set. This may be attributed to the complex relationship between the DFT calculated results and the experimental ones. The relationship shall be nonlinear and much more complicated, so the linear regression method shall be ineffective.

Table 2. The mean absolute values of the relative errors (%) for the DFT calculation and four correction methods.

	DFT	DFT-NN	DFT-GANN	DFT-AGANN	DFT-MLR
training set	4.54	0.55	0.46	0.35	4.35
testing set	4.93	2.56	2.15	1.20	4.96
overall	4.61	0.88	0.74	0.49	4.45

Figure 3 shows the calculated lattice energies of the DFT, DFT-NN, DFT-GANN and DFT-AGANN method versus the experimental values. Figure 3(a) is for the calculated lattice energies using the pure DFT method; Fig. 3(b) is for the values obtained using the method of the DFT with NN correction; Fig. 3(c) is for the values obtained using the method of the DFT with GANN correction, and Fig. 3(d) is for the values obtained using the method of the DFT with AGANN correction. The results of the MLR method are not shown in Fig. 3 for its invalidation in the correction. Compared with Fig. 3(a), the calculated results of all the other three figures are closer to the experimental values for both the training and testing sets. And from the methods of DFT-NN, DFT-GANN to DFT-AGANN, the corrected values are more and more close to the experimental ones. The result of the AGANN approach is the best. However, as also shown in Table 1, there are several crystals whose relative errors are still quite high in the testing set after the AGANN correction. This may result from the fact that such a kind of samples has a fewer amount of data and their characteristics cannot be extracted by NN method. If more experimental data are available, the correction effect can be further improved.

In the above discussion, it can be found that the corrected result of the AGANN method is better than that of the NN, GANN, or MLR method. Compared with the GANN method, the mean absolute value of the relative errors in the AGANN method decreases about one half for the testing set. This indicates that the alternate combinational approach of GA and NN indeed can serve as a better correction tool than the common combinational approach. Table 3 lists the weights and thresholds of the optimal network in the AGANN method. These weights and thresholds can be directly used in the correction of DFT calculation for new metal halides and hydrides.

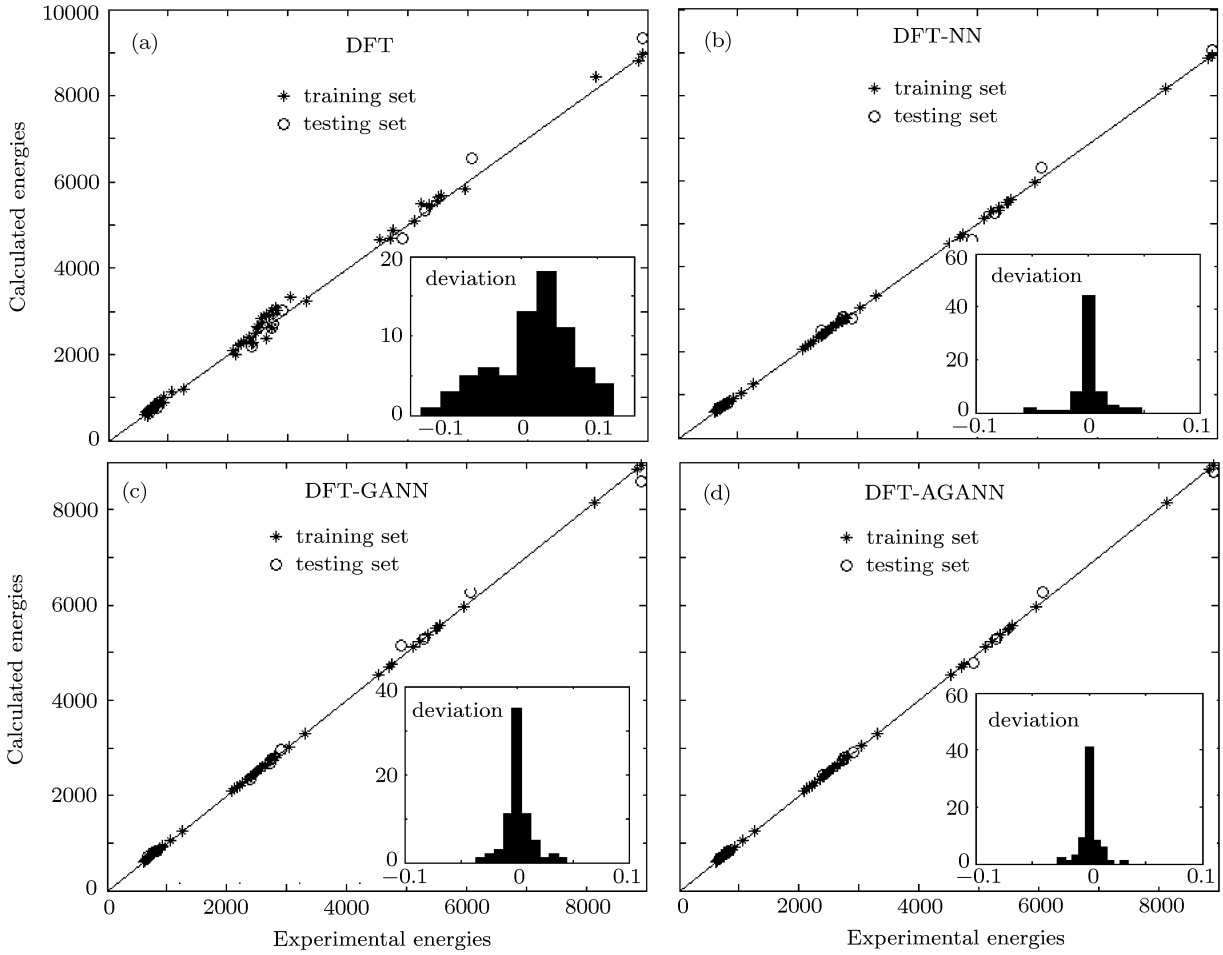


Fig. 3. The calculated lattice energies versus the experimental lattice energies. The horizontal coordinate of each figure is the experimental lattice energy, and the vertical coordinate is the calculated value through the method of DFT, DFT-NN, DFT-GANN, or DFT-AGANN. Their values are in units of kJ·mol^{-1}. The diagonal line is where the vertical and horizontal values are equal. Crosses (×) denote the training set while circles (o) the testing set. Insets are the distribution histograms of the overall relative errors.

Table 3. The weights and thresholds of the optimal network in the AGANN method.

weights	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	
E_c	58.17	–6.43	3.28	–13.63	16.86	16.21	–3.18	–4.15	
N_c	–33.76	71.99	–0.06	–3.13	6.23	5.94	–3.06	0.33	
N_a	2.72	8.49	0.09	30.51	–1.08	–1.01	1.47	–0.27	
V_c	–21.70	27.22	0.61	4.75	–24.42	–23.48	–0.59	1.44	
E_e	0.02	0.05	0.88	–0.04	2.47	–2.59	–0.18	–2.93	
	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	E_e
thresholds	–4.95	–62.00	–2.95	2.00	1.12	1.08	0.33	–2.53	0.39

The data of the E_c, N_c, N_a, V_c rows are the weights between the input layer and the hidden layer, and the E_e row involves the weights between the hidden layer and the output layer. The data of the last row are the thresholds of the hidden neurons and the output neuron.

4. Summaries

This paper presents an external correction for DFT calculation of 72 metal halides and hydrides' lattice energies. An alternate combinational approach of GA and NN is proposed on the basis of analysing the disadvantages of the common combinational approach

in detail. This alternate strategy first adopts GA to find the coarse values of a number of local optimums, and then BP is used to find the accurate values of these local optimums. Afterwards, GA is used to jump out of these optimums and to find more optimums, and BP is again used to identify the accurate values of these new optimums, and the cycle will be repeated until the global optimum is obtained. Compared with the common GANN, the alternate combination approach transmits network architectures between GA and BP repeatedly and it only needs GA with a small population. But the common GANN transmits network architectures between GA and BP only one time and needs a GA with large population. The experiment of metal halides and hydrides shows that the AGANN has the best correction effect among the four methods of NN, GANN, AGANN and MLR. Through the AGANN correction, the mean absolute values of the relative errors of the training set and the testing set decrease from the previous 4.54% and 4.93% to 0.35% and 1.20% respectively. The two mean values are reduced to 0.46% and 2.15% respectively by the GANN correction. Besides, although the alternate method has to apply GA and BP many times, its convergent time is not longer than the common GANN. This is because GA adopts small population here, and every GA optimisation and every BP optimisation are much shallower compared with the common GA. These indicate that the alternate strategy is indeed better than the common combinational approach.

The AGANN described above is a simple model, and there is still much room for improvement. For example, we can adopt adaptive generation times of GA and iteration epochs of BP, which can change with the optimisation state. And chaos operator can be used to generate initial population of GA to ensure that the small population has enough ergodicity and the chromosomes have enough difference between each other. These measures can further improve the overall convergent speed and performance.

The external correction of DFT calculation presented here can also be used to evaluate other properties such as heat of formation, optical absorption energies, bonding energy and reaction barrier of organic molecules and the band gap of semiconductors. Although this method treats the DFT as a black box and models the error through external statistical information, it may be helpful for gaining physical insight if the input physical descriptors correlate closely to exchange–correlation functional. More calculation data and experimental data will also help improve the precision of the correction.

References

[1] Hohenberg P and Kohn W 1964 *Phys. Rev.* **136** B864
[2] Kohn W and Sham L J 1965 *Phys. Rev.* **140** A1133
[3] Kohn W 1999 *Rev. Mod. Phys.* **71** 1253
[4] Bechstedt F and Del Sole R 1988 *Phys. Rev.* B **38** 7710
[5] Castleton C W M, Höglund A and Mirbt S 2006 *Phys. Rev.* B **73** 35215
[6] John P P and Yue W 1992 *Phys. Rev.* B **45** 13244
[7] Hu L H, Wang X J, Wong L H and Chen G H 2003 *J. Chem. Phys.* **119** 11501
[8] Wang X J, Wong L H, Hu L H, Chan C Y, Su Z M and Chen G H 2004 *J. Chem. Phys.* A **108** 8514
[9] Duan X M, Li Z H, Song G L, Wang W N, Chen G H and Fan K N 2005 *Chem. Phys. Lett.* 410 **125**
[10] Duan X M, Song G L and Li Z H 2004 *J. Chem. Phys.* **121** 7086
[11] Duan X M, Li Z H, Hu H R, Song G L, Wang W N, Chen G H and Fan K N 2005 *Chem. Phys. Lett.* 409 **315**
[12] Widrow B and Lehr M A 1990 *Proc. IEEE* **9** 1415
[13] Li W, Chen J F and Wang T 2009 *Chin. Phys.* B **18** 2441
[14] Lance D C 1995 *Practical Handbook of Genetic Algorithms* (Boca Raton Florida USA: CRC Press)
[15] Li H, Shi L L, Zhang M and Su Z M 2007 *J. Chem. Phys.* **126** 144101
[16] Cook D F, Ragsdale C T and Major R L 2000 *Eng. Appl. Artif. Intell.* **13** 391
[17] Goldberg D E and Holland J H 1988 *Machine Learning* **3** 95
[18] Yu A L 2008 *Chin. Phys.* B **17** 878
[19] Zhang S, Liu H X, Gao D T and Du S D 2003 *Chin. Phys.* **12** 594
[20] John P P, Kieron B and Matthias E 1996 *Phys. Rev. Lett.* **77** 3865
[21] Perdew J P and Wang Y 1992 *Phys. Rev.* B **45** 13244
[22] Kresse G and Hafner J 1993 *Phys. Rev.* B **47** 558
[23] Kresse G and Furthmuller J 1996 *Phys. Rev.* B **54** 11169
[24] David V 1990 *Phys. Rev.* B **41** 7892
[25] David R L 2006 *Handbook of Chemistry and Physics* 87th edition (Boca Raton Florida USA: CRC Press)