SCIENTIFIC REPORTS

OPEN

An extended cluster expansion for ground states of heterofullerenes

Yun-Hua Cheng[1], Ji-Hai Liao[1], Yu-Jun Zhao[1,2] & Xiao-Bao Yang[1,2]

It is challenging to determine the ground states of heterofullerenes due to the numerous isomers. Taking the $C_{60-n}B_n$ heterofullerenes ($1 \leq n \leq 4$) as an example, our first-principles calculations with the isomer enumeration present the most stable structure of $C_{57}B_3$, which is energetically favored by 0.73 eV than the reported counterpart. It was difficult to conduct the enumeration for the isomers with n beyond 4 because of the expensive first-principle calculations. Here, we propose a nomenclature to enhance structural recognition and adopt an extended cluster expansion to describe the structural stabilities, in which the energies of the heterofullerenes with various concentrations are predicted by linear combination of the multi-body interactions. Unlike the conventional cluster expansion, the interaction parameters are derived from the enumeration of $C_{60-n}B_n$ ($n = 1\sim4$), where there are only 4 coefficients to be fitted as a function of composition for the consideration of local bonding. The cross-validation scores are 1~2 meV per atom for both $C_{55}B_5$ and $C_{54}B_6$, ensuring the ground states obtained from our model are in line with the first-principles results. With the help of the structural recognition, the extended cluster expansion could be further applied to other binary systems as an effective complement to the first-principle calculations.

Since the discovery in 1985[1], the C_{60} fullerene has attracted great attentions due to various potential applications with its unique structure-dependent properties. The cage of C_{60} fullerene, with the size which is large enough to be observed by transmission electron microscopy or scanning probe methods[2,3], is likely to keep stable when built into molecular circuits as semiconductor materials[4–7]. Doping has been adopted as the conceivable way to alter its charge distribution and then tune the optical, electronic and magnetic properties in the solid state[8–10], including the way of adding exohedral, endohedral and substitutional atoms[11]. As neighbors to carbon in the Periodic Table, boron and nitrogen with similar atomic radius are the popular choices as the heteroatoms for the substitution of one or more of the carbon atoms[12–14]. The $C_{60-n}B_n$ heterofullerenes with $1 \leq n \leq 6$ were produced by Laser vaporization of a graphite pellet containing boron nitride powder[15], which indicated that boron doped C_{60} cage still appeared to be particularly stable. During the synthesis and characterization of $C_{59}N$[16], $C_{59}N$ in the vapor phase was found existing in monomer form as a molecular free radical[17], where the single $C_{59}N$ heterofullerene molecule could be used as a new molecular rectifier in a double-barrier tunnel junction via the single electron tunneling effect[18].

Theoretically, Kurita *et al.* found that the molecular structures of $C_{59}B$ and $C_{59}N$ maintained the cage of C_{60}, which was distorted due to a large-size dopant such as Sulfur[19]. In addition to the calculation of $C_{59}B$ by the first-principles method[20], the ground state geometries of $C_{60-n}N_n$ and $C_{60-n}B_n$ for $2 \leq n \leq 8$ were screened using semiempirical MNDO, AM1, PM3, and *ab initio* methods[21]. As $C_{48}B_{12}$ and $C_{48}N_{12}$ are promising components for molecular rectifiers[22], Garg *et al.*[23] reported a detailed study of structural, electronic and vibrational properties of B-doped heterofullerenes ($C_{60-n}B_n$, for $n = 1\sim12$) based on *ab initio* calculations, concluding that the maximum number of boron atoms in a pentagon/hexagon ring was one/two. In general, it is difficult to determine the ground state of heterofullerenes due to two main obstacles: (i) only small amount of isomers for given compositions were considered; (ii) the optimized heterofullerenes largely depend on the initial geometry of numerous possible isomers[24]. Hence, there is still a lack of theoretical studies to search the energetically-preferred structures of heterofullerenes.

In this paper, we perform a systematic investigation of $C_{60-n}B_n$ ($n = 1\sim6$) based on the first-principles calculations with the congruence check, in which the structure recognition is achieved by a uniform numbering scheme for C_{60}. Furthermore, an extended cluster expansion is proposed to estimate the total energies with all the possible

[1]Department of Physics, South China University of Technology, Guangzhou, 510640, People's Republic of China. [2]Key Laboratory of Advanced Energy Storage Materials of Guangdong Province, South China University of Technology, Guangzhou, 510640, P. R. China. Correspondence and requests for materials should be addressed to X.-B.Y. (email: scxbyang@scut.edu.cn)
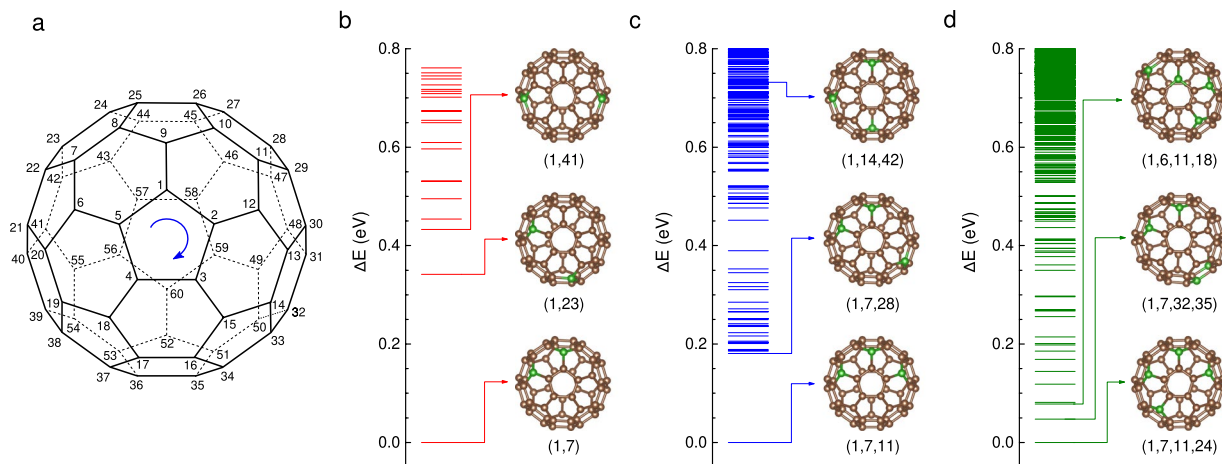
Figure 1. C_{60} fullerene cage with the systematic numbering scheme[25] and the isomers energy relative to the lowest energy for $C_{60-n}B_n$ ($2 \leq n \leq 4$). (**a**) Systematic numbering scheme. Any atom in the C_{60} fullerene (the coordinates are listed in Supplementary Table S1) has its unique sequence number (SN). The arrow indicates the direction of the numbering commencement. (**b**) The energies of $C_{58}B_2$ isomers. The lowest, the 2nd and 3rd lowest energetic structure of $C_{58}B_2$, denoted by (1, 7), (1, 23) and (1, 41), respectively are shown. (**c**) The energies of $C_{57}B_3$ isomers. The cage with the lowest and the 2nd lowest energy denoted by (1, 7, 11) and (1, 7, 28), respectively are shown. The cage denoted by (1, 14, 42) is the previously reported[23] ground state of $C_{57}B_3$ which has a higher energy of 0.73 eV than the global minimal state (1, 7, 11). (**d**) The energies of $C_{56}B_4$ isomers. The global minimal energetic structure, the 2nd and 3rd lowest energetic cage, denoted by (1, 7, 11, 24), (1, 7, 32, 35) and (1, 6, 11, 18), respectively, are shown. In (**b**), (**c**) and (**d**), we only show the relative energies lower than 0.8 eV.

SI	E (eV)	SI	E (eV)	SI	E (eV)	SI	E (eV)
(1, 7)	0.00	(1, 3)	0.53	(1, 32)	0.67	(1, 24)	0.74
(1, 23)	0.34	(1, 16)	0.59	(1, 57)	0.70	(1, 15)	0.74
(1, 41)	0.43	(1, 33)	0.61	(1, 9)	0.71	(1, 14)	0.75
(1, 50)	0.45	(1, 56)	0.65	(1, 35)	0.71	(1, 34)	0.76
(1, 52)	0.49	(1, 49)	0.65	(1, 31)	0.71	(1, 2)	1.19
(1, 60)	0.53	(1, 13)	0.67	(1, 6)	0.72		

Table 1. The 23 isomers of $C_{58}B_2$. The structural indexes (SIs) of $C_{58}B_2$ isomers are sorted in the order of ascending energy (columns titled by E) relative to that of the isomer denoted by (1, 7), which is the minimal energetic structure of the $C_{58}B_2$ heterofullerenes.

pair, three-body, and four-body interactions derived from the enumeration of $C_{60-n}B_n$ ($n = 1\sim4$), however, there are only four coefficients to be fitted for the consideration of composition. We determine the ground state structures of $C_{55}B_5$ and $C_{54}B_6$ as confirmed by the first-principles calculations, indicating the possible application to other alloy systems.

Structure recognition

We adopt the systematic numbering scheme recommended by the IUPAC[25] to identify the vertices of the C_{60} cage, as shown in Fig. 1a. Any atom in the C_{60} fullerene cage (the coordinates are listed in Supplementary Table S1) has its unique sequence number (SN). An isomer of $C_{60-n}B_n$ heterofullerenes is denoted by an index consisting of the ascending ordered SNs of the substituted vertices, *i.e.* (σ_1, σ_2, ..., σ_n) which is called structural index (SI). According to the congruence check, the total number of the $C_{60-n}B_n$ ($n \leq 4$) heterofullerenes isomers is 4,517, which is only about 1% of the corresponding combination number (see Supplementary Table S2). All the total energies of these candidates are obtained by the first-principles calculations (details in Supplementary Information), where all the calculated structures are fully relaxed without any symmetry constraint. The ground state structures with the energy profiles are shown in Fig. 1. There are 23 different isomers (shown in Table 1) for $C_{58}B_2$, which is in agreement with earlier calculations[23,26,27]. As shown in Fig. 1b, the global minimum structure of $C_{58}B_2$ is the cage with two boron atoms at the opposite sites of a hexagon ring. The ground states of $C_{57}B_3$ and $C_{56}B_4$, shown in Fig. 1c and d, respectively, have the similar pattern that all boron atoms are at the opposite vertices of the hexagon rings adjacent to each other, where there are no more than two boron atoms on a hexagon ring. The ground state structures of $C_{58}B_2$ and $C_{56}B_4$ are in agreement with those previous calculations[21,23]. However, the total energy of the most stable $C_{57}B_3$ is 0.73 eV lower than that of the one proposed by Garg *et al*.[23], which ranks 126th according to our enumeration of the 303 isomers as shown in Fig. 1c.

Buckminsterfullerene, *i.e.*, the only isomer fulfilling the isolated pentagon rule (IPR)[28], is a spherical molecule with 60 carbon atoms at vertices, containing 32 faces including 20 hexagons and 12 pentagons where no pentagon shares a vertex[29]. Treating C_{60} as a semiregular polyhedron and considering all the symmetry operations represented by the 120 symmetry matrices (SMs), we obtain a 60×120 matrix called numbering matrix (NM) (available in the Supplementary Dataset file), in which the nth row lists the coincident atoms for the nth atom under all the symmetry operations and the nth column contains the corresponding coincident atoms for all the 60 atoms under the operation of the nth SM.

Herein, based on the NM of C_{60}, we propose a nomenclature for the $C_{60-n}B_n$ heterofullerenes. The flow chart of our structure recognition method is shown in Supplementary Fig. S1. The detailed information of the structure recognition method is available in the Supplementary Information. We have deduced all the SIs of the inequivalent structures of $C_{60-n}B_n$ heterofullerenes for $2 \leq n \leq 10$, the numbers of the SIs are listed in Supplementary Table S2, which is in good agreement with the previous results[24,30,31]. The inequivalent structures can also be singled out by our recently developed structure recognition method[32]. Note that our nomenclature is derived from the symmetry operation matrices, which can be obtained according to the coordinates of the system with the corresponding symmetry operations. Thus, this nomenclature can be extended for the C_{60} with non-IPR isomers, as well as the larger fullerenes e.g. C_{70} and C_{82}.

Extended cluster expansion method

Combined with the isomer enumeration and the first-principles calculations, we have determined the ground state structures of $C_{60-n}B_n$ heterofullerenes with $2 \leq n \leq 4$. For those heterofullerenes with higher boron concentration, we can enumerate all the isomers by the recognition method discussed above. However, it will be over expensive to search the ground state structures with the first-principles calculations, because there are 45,718, 418,470, 3,220,218, 21,330,558, 123,204,921 and 628,330,629 isomers for $C_{55}B_5$, $C_{54}B_6$, $C_{53}B_7$, $C_{52}B_8$, $C_{51}B_9$ and $C_{50}B_{10}$, respectively. Analogue to the conventional cluster expansion(CE)[33], an extended cluster expansion (ExCE) method is discussed aiming at this problem in the following.

As is known, the CE method is an efficient tool for studying structural properties of any binary structures over a wide range of concentrations[34–38], parameterizing the total energy for any given configuration of A_xB_{1-x} $(0 \leq x \leq 1)$ to avoid the expensive cost of the first-principles calculations. The enthalpy of formation for a certain configuration $\overrightarrow{s}$ is described exactly by a set of multi-body interaction parameters J_i combined as the form of an Ising-like Hamiltonian[37], which is often approximated as a polynomial function of occupation variables,

$$\Delta H_{CE}(\overrightarrow{s}) = \sum_{\alpha} m_\alpha J_\alpha \prod_{i \in s'} s_i$$

(1)

where the summation is over all the non-equivalent clusters, a set of sites i denoted by α, and the average is taken over all the clusters α that are equivalent to α by symmetry. The coefficients are defined as effective cluster interaction (ECI) parameters, and m_α is the number of the clusters equivalent to α.

In general, the total energies of given configurations are described by the combine of single-atom contributions, pair interactions and multi-body interactions, which are expected to gradually converge as more interactions are considered. However, fitting with larger number of parameters will be also time expensive. To balance the accuracy and efficiency of CE method, the number of effective multisite interactions can be greatly reduced[39], while the combinations of possible effective interactions result in another global optimization. Herein, we attempt to derive the multi-body interactions taking account of the total impurity concentration and fit the total energy with fewer parameters for higher accuracy.

An isomer of $C_{60-n}B_n$ heterofullerenes, taking the carbon atoms as the background, can be viewed as a cluster of boron atoms denoted by $(\sigma_1, \sigma_2, \cdots, \sigma_n)$. We suppose the total energy be attributed to all the subclusters, which are enumerated for fitting the total energy. Firstly, there is only one isomer for $C_{59}B$, therefore the energy difference of $C_{59}B$ relative to C_{60} is $E_1 = E_{tot} - E_0 = 3.39$ eV, where E_{tot} and E_0 are the total energy of $C_{59}B$ and C_{60}, respectively. The energy difference E_1 is responsible for the reaction heat when one C atom is substituted by one boron atom, which can be considered as the single-atom contribution in the expansion. For a $C_{58}B_2$ isomer whose subclusters are two equivalent boron single dopants, the environment of each boron atom is different from that of $C_{59}B$, thus the energies can be expressed as $E_0 + 2c_1E_1$ with the coefficient c_1 as a function of boron concentration. Based on all the energies of $C_{58}B_2$ isomers, we fit this coefficient c_1 to be 0.980 with the average deviation of 0.208 eV. Similar to the CE method, the fitting quality is determined by the cross-validation (CV) score[39],

$$CV = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(E_i^{DFT} - \hat{E}_i)^2}$$

(2)

where E_i^{DFT} and $\hat{E}_i$ denote the DFT calculated and predicted energy of a particular structure i. The deviation of E_i^{DFT} from $\hat{E}_i$ is taken as the interactive energy of the corresponding boron cluster. For $C_{58}B_2$, The 23 fitting deviations are the B-B interactions in the 23 isomers, respectively.

Similarly, the total energy of a $C_{57}B_3$ isomer denoted by $\overrightarrow{\sigma}$, is contributed by three single dopants and three pair interactions. For example, the isomer of $C_{57}B_3$ denoted by $(1, 7, 11)$ can be expanded as 6 clusters including 3 singles and 3 pairs. The initial SIs of the subclusters are listed in Fig. 2a along with their smallest SIs by our recognition method. Apart from the singles, the isomer denoted by $(1, 7, 11)$ has the 3 subclusters denoted by $(1, 7)$, $(1, 7)$ and $(1, 24)$. Hence we express the total energy as $E_0 + 3c_1E_1 + c_2\sum_\alpha E_2^\alpha$, where E_2^α denotes the B-B interaction for any boron pair as a subcluster of $\overrightarrow{\sigma}$. The coefficients c_1 and c_2 are fitted to be 0.972 and 0.782 respectively, and the average deviation is 0.097 eV. The 303 fitting deviations are further taken as the 303 3-body interactive energies, respectively. For a $C_{56}B_4$ heterofullerene, the total energy is contributed by 4 single dopants, 6 pair
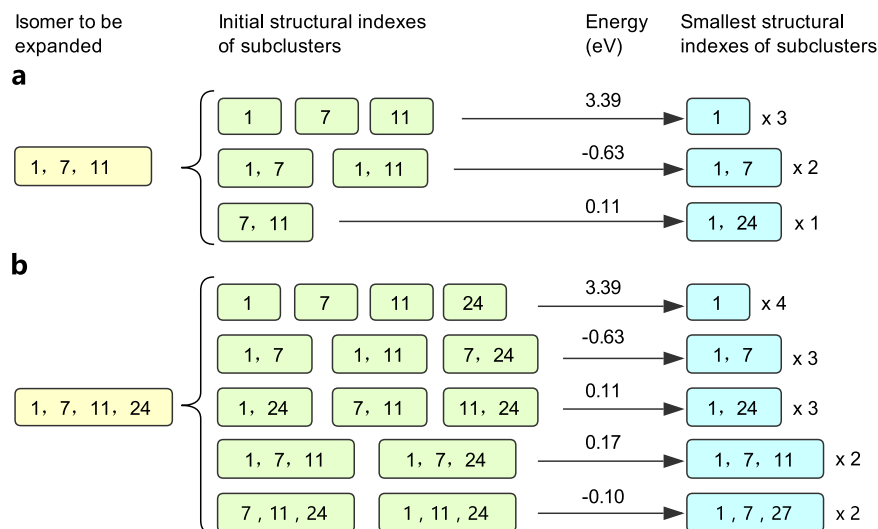
3

Figure 2. Examples for the cluster expansion. (**a**) Expansion of (1, 7, 11). (**b**) Expansion of (1, 7, 11, 24). The initial structural indexes (SIs) are obtained simply from all possible combinations for the numbers from the SI of the isomer to be expanded. The smallest SIs are obtained by our structure recognition method. The energy values shown are the fitting errors which are supposed to be the energetic contributions of the corresponding subclusters. The numbers after multiple signs indicate the numbers of equivalent subclusters in cluster expansion.

$C_{60-n}B_n$	Fitting cutoff	c_1	c_2	c_3	c_4	CV (eV)
$C_{58}B_2$	1	0.980				0.208
$C_{57}B_3$	2	0.972	0.782			0.097
$C_{56}B_4$	3	0.966	0.690	0.640		0.065
$C_{55}B_5$	2	0.928	0.200			0.106
$C_{55}B_5$	3	0.959	0.594	0.416		0.064
$C_{55}B_5$	4	0.960	0.585	0.429	0.132	0.060
$C_{54}B_6$	2	0.919	0.160			0.148
$C_{54}B_6$	3	0.963	0.592	0.397		0.125
$C_{54}B_6$	4	0.963	0.587	0.399	0.031	0.124

Table 2. Cross-validation (CV) score versus fitting cutoff. Fitting cutoff refers to the number of atoms in the largest subcluster used for the fitting. We perform three kinds of fittings for $C_{55}B_5$ and $C_{54}B_6$ in which the largest subcluster are pair, triplet and quadruplet, respectively.

interactions, and 4 triplet interactions. For example, the subclusters from the expansion of the isomer of $C_{56}B_4$ denoted by (1, 7, 11, 24) are listed in Fig. 2b. Analogously, we obtain the 4,190 4-body interactions when the energies of all $C_{56}B_4$ isomers are fitted by E_0, $3E_1$, the 2-body and 3-body interactions, in which the average fitting deviation is 0.065 eV and the coefficients c_1, c_2 and c_3 are 0.966, 0.690 and 0.640, respectively.

As shown above, the coefficients reflect the boron atom's concentration and the fitting deviations are attributed to the multi-body interactions. The fitting coefficients and CV scores for $C_{58}B_2$, $C_{57}B_3$ and $C_{56}B_4$ are listed in Table 2. It shows that the introducing of multi-body interactions will improve the accuracy of cluster expansion and the interactions will decrease as boron dopants increasing. For example, the coefficients c_1 is from 0.980 to 0.966, and c_2 is from 0.782 to 0.690. Fig. 3 shows the statistical distributions for the 2-body, 3-body and 4-body interactions. Compared to the 3-body and 4-body interactions, the 2-body interactions distribute in a wider range. The 4-body interactions exhibit the similar characteristics of normal distribution around zero point. It can be inferred that the interactions of 2-body and 3-body are much more important than that of 4-body or other multi-body interactions, hence the fitting will also reach a rather good convergence even if only 2-body 3-body interactions are considered.

Herein, we propose an extended cluster expansion for the $C_{60-n}B_n$ heterofullerene, where the energy of isomer denoted by $\overrightarrow{\sigma}$ is expressed as

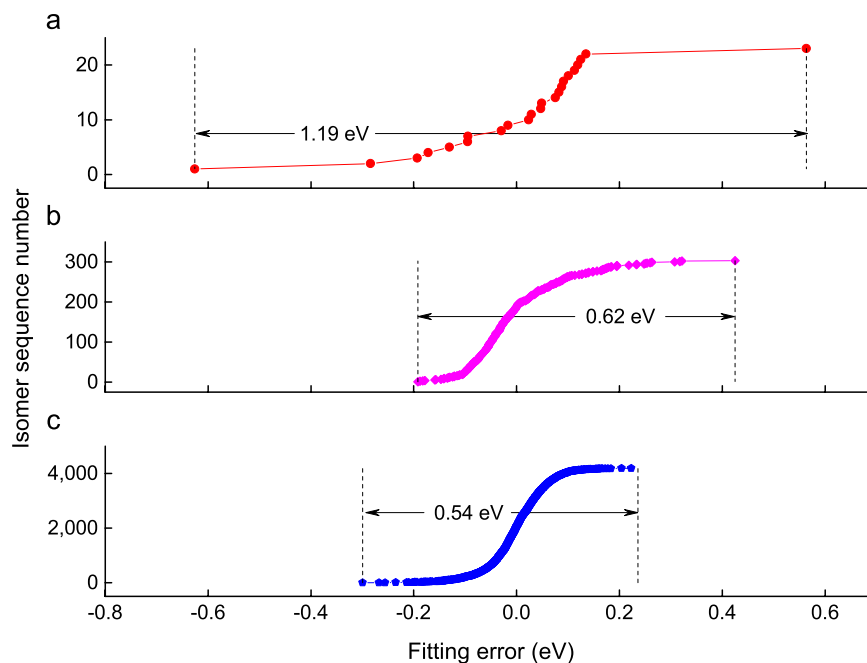$$E_{DFT}(\overrightarrow{\sigma}) = \hat{E}(\overrightarrow{\sigma}) + E_n \tag{3}$$

Figure 3. Distribution of the fitting errors of $C_{58}B_2$, $C_{57}B_3$ and $C_{56}B_4$. (**a**) The energetic contribution of boron pairs. (**b**) The interactive energy of boron triplets. (**c**) The interactive energy of boron quadruplets. The y-axis represents the sequence numbers for isomers when sorted by their fitting errors in ascending order. The ranges of the x-axis are the same so that we can easily compare the distribution ranges of the fitting errors of different interactions marked by figures between pairs of arrows.

where $E_{DFT}(\overrightarrow{\sigma})$ and $\hat{E}(\overrightarrow{\sigma})$ are the DFT calculated energy and the predicted energy, respectively, and E_n denotes the fitting deviation which is supposed to be responsible for the n-body interaction of the boron cluster denoted by $\overrightarrow{\sigma}$. The predicted energy is as follows,

$$\hat{E}(\overrightarrow{\sigma}) = E_0 + \sum_{i=1}^{n-1} c_i E_i \tag{4}$$

where the summation runs over all possible sizes of the subclusters of $\overrightarrow{\sigma}$. The first term E_0 represents the energy of C_{60}, and E_i denotes the total effective energy contribution from all the clusters with i heteroatoms, as is expressed as below

$$E_i = \sum_{\alpha} E_i^{\alpha} \tag{5}$$

where the summation is over all those subclusters consisting of i heteroatoms, $i.e.\, 1 \le \alpha \le C_n^i$. Different from the conventional CE, the multi-body interactive energies E_i apart from E_n, should be multiplied by different combination coefficients c_i before they make contribution to the total energy of the $C_{60-n}B_n$ heterofullerene cages, where the coefficients are obtained by fitting the DFT-calculated energies of the selected $C_{60-n}B_n$ heterofullerenes with those E_i for $i < n$. To balance the accuracy and computation cost, we set 4 as the maximum value of the summing index in Eq. (4) for the cages of $C_{60-n}B_n$ where $n \ge 5$.

We show the flow chart of our method in Fig. 4 and make a detailed description for the process in searching the ground state structures for the $C_{55}B_5$ cage.

(1) Generate the SIs of all the $C_{55}B_5$ isomers (45,718 in all), list all the subclusters of these SIs and calculate the total energies of the isomers by Eq. (3), where the combination coefficients for the interaction of singles, pairs and triplets are from the fitting of the energies of $C_{56}B_4$, and the coefficient for quadruplet-body interaction is 1.

(2) Choose the 100 minimum energetic structures and calculate their total energies (saved in E_i^{DFT}) using the first-principles calculations.

(3) Retain the corresponding coefficients with the total energies from the first-principles calculations. Use the coefficients to calculate the total energies of all the isomers by Eq. (3).

(4) Apart from those selected before, select the 100 minimum energetic structures and calculate their total energies (appended to E_i^{DFT}) using DFT.

(5) Fit the energies of the structures, which has been selected until now, by Eq. (3), and update the corresponding coefficients.

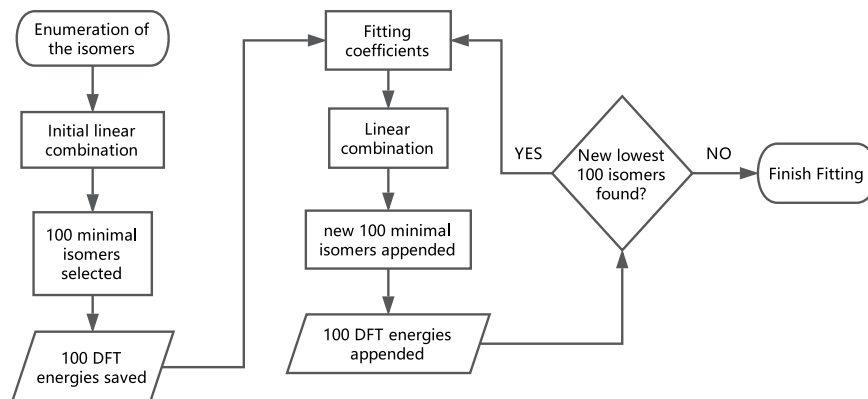(6) Use the coefficients to calculate the total energies of all the isomers.

Figure 4. Flowchart for the extended cluster expansion (ExCE) method. The selection of the coefficients for the initial linear combination have a certain arbitrariness, which may bring about the different 100 lowest energy isomers for the first fitting iteration but not affect the final convergence and the searching for the putative ground state isomer. Here, we set 100 as the number of the new added isomers for each iteration, which may differ for different iteration and alloy systems.

(7) Check whether the latest DFT calculation brings about new structures whose energies are among the minimum 100 ones in E_i^{DFT} If it does, repeat the steps 4 to 7, otherwise, break the process.

The fitting ultimately reaches a rather reasonable convergence after several hundreds of structures with the lowest predicted energy are calculated. Supplementary Table S3 and Table S4 shows the variations of coefficients for $C_{55}B_5$ and $C_{54}B_6$, respectively, as a function of the number of isomers calculated by the first-principles method. Note that the coefficients of c_2 and c_3 are around 0.6 and 0.4 respectively, while the coefficient c_4 is approaching zero. Similar to the conventional CE, the energy of interatomic bonds is usually dominated by short-range inter-actions[39]. On the other hand, enormous interactions would be introduced when we use the $C_{60-n}B_n$ with higher boron concentration for cluster expansion, which will be expensive in computational cost. Note that any other binary systems can be similarly searched by the extended cluster expansion, where the appropriate cutoff of the size of the subcluster should be carefully made to balance of the accuracy and computation cost. The nomenclature and extended cluster expansion can be also applied for the ternary systems, where different atoms are distinguished from the candidates found in the binary systems.

Application to $C_{55}B_5$ and $C_{54}B_6$

According to the structure recognition, there are 45,718 and 418,470 inequivalent structures for $C_{55}B_5$ and $C_{54}B_6$, respectively. Using the method discussed above and following those steps, we have made a prediction for the ground state structures of $C_{55}B_5$ and $C_{54}B_6$, where the energy profiles are shown in Fig. 5 (detailed in Supplementary Table S5). The optimized fitting coefficients of $C_{55}B_5$ are adopted for the initial combination coefficients for $C_{54}B_6$. As the fitting steps move forward, most of the energies of the newly added isomers in each fitting iteration gradually increase. However, the 100th lowest energy of the fitting iteration decreases rapidly and eventually converges, after 6/8 fitting iterations for $C_{55}B_5$/$C_{54}B_6$. The minimum energetic isomers for both $C_{55}B_5$ and $C_{54}B_6$ emerge in the first iteration. The optimized fitting coefficients were obtained for $C_{55}B_5$/$C_{54}B_6$ after the energies of the selected 600/800 isomers are calculated and fitted, for which the results are listed in Table 2. The CV score of the final fitting is 0.064/0.124 eV for $C_{55}B_5$/$C_{54}B_6$, and the largest deviation of total energy between ExCE method and DFT calculations is 0.192/0.403 eV for $C_{55}B_5$/$C_{54}B_6$, indicating that the fitting energies are in good agreement with the DFT calculations. For both $C_{55}B_5$ and $C_{54}B_6$, the coefficients c_1 are close to 1, implying that single boron atom does make an important contribution despite of the concentration. The coefficient c_4 is much smaller and the quadruplet interactions play a trivial role in the ExCE calculations, since the fitting will be in good accuracy when the pair and triplet interactions are considered in the energy predications of $C_{60-n}B_n$ for n $\geq$ 5. This is consistent with the above assumption that the energy of interatomic bonds should be usually dominated by short-range interactions.

For $C_{55}B_5$, the ExCE energy versus the computational energy is shown in Fig. 6a. The putative ground state is (1, 7, 11, 24, 27). The five heteroatoms are located at the 5 apposite sites of 5 hexagon rings and make up of a pentagon which encloses a carbonic pentagon ring, with the similar pattern of the ground state of $C_{60-n}B_n$ for $2 \leq n \leq 4$. The next preferred positions for boron atoms are (1, 7, 11, 32, 35), with a total energy of 0.32 eV higher. It was reported in ref.[23] that the minimal energetic structure for $C_{55}B_5$ was (1, 7, 18, 51, 59). In contrast, this structure is found to be 253rd in the stability ranking and higher in energy by 0.68 eV than our minimal energetic structure.

For $C_{54}B_6$, the putative lowest energy is from (1, 6, 11, 18, 24, 27) which is shown in Fig. 6b. In this isomer, 4 boron atoms are at the consecutive opposite sites and the other two are at the isolated opposite sites. The next most favorable configuration is (1, 7, 11, 16, 24, 36). Chen et al.[21] reported that the global minimum structure was (1, 6, 9, 12, 15, 18), but from our result, this structure has a higher energy of 0.34 eV relative to our minimal energy and ranks 100th in our ascending order list of the total energies. Garg et al.[23] predicted the minimal energetic cage for $C_{54}B_6$ was (1, 6, 11, 18, 27, 31), now in our calculation, this structure rates the 340st in the ranking of stability and is less stable by 0.52 eV with respect to our minimal energy structure.
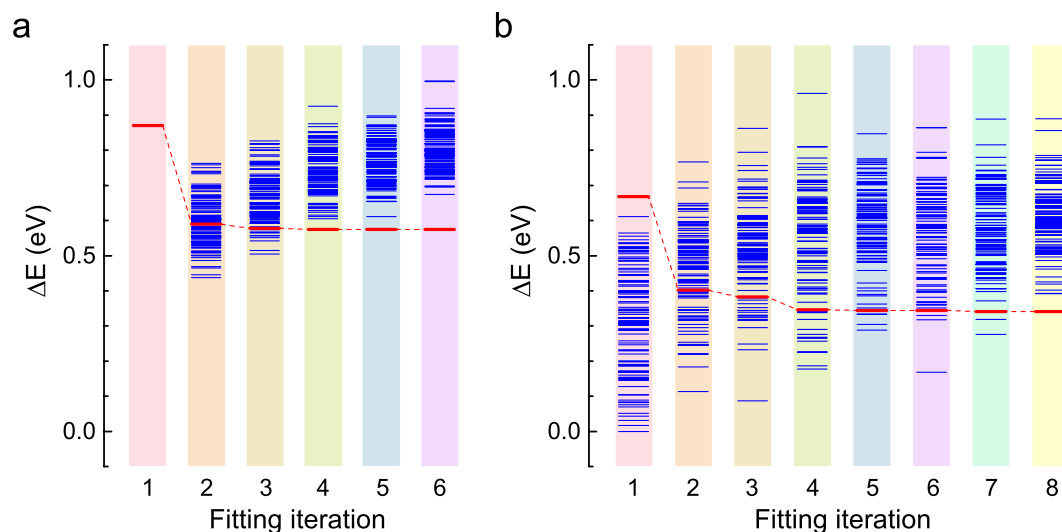
Figure 5. Calculated energies of the isomers in the fitting process. For each fitting iteration, the new found 100 $C_{55}B_5$ and $C_{54}B_6$ isomers with the lowest energies are shown in (**a**) and (**b**), respectively. With the fitting iteration moving forward, the numbers of the fitted isomers have an increase of 100 per iteration, reaching 600 and 800 finally in (**a**) and (**b**), respectively. In (**a**) and (**b**), all the energies are relative to the corresponding minimum ones. Only the energies of those newly added isomers for each iteration are shown. The solid line linked by the dashed lines represents the 100th lowest energy of the current fitting iteration.
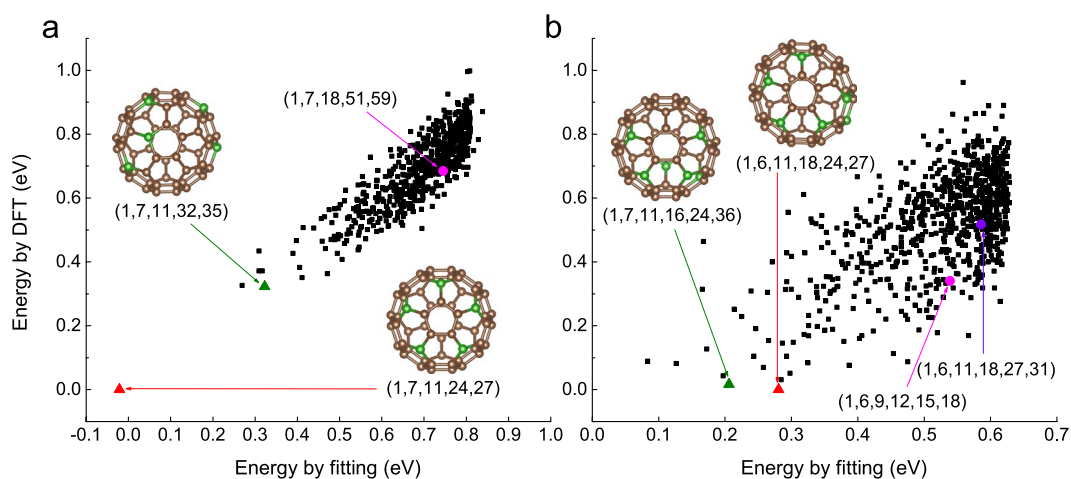
Figure 6. Energy predicted by ExCE versus DFT calculated energy. (**a**) The predicted and DFT calculated energies for $C_{55}B_5$ isomers. The putative ground state and the isomer with the 2nd lowest energy for $C_{55}B_5$ are denoted by (1, 7, 11, 24, 27) and (1, 7, 11, 32, 35), respectively. The ground state of $C_{55}B_5$ reported previously[23] is marked as (1, 7, 18, 51, 59). (**b**) The predicted and the DFT calculated energies for $C_{54}B_6$ isomers. The ground state and the isomer with the 2nd lowest energy for $C_{54}B_6$ are denoted by (1, 6, 11, 18, 24, 27) and (1, 7, 11, 16, 24, 36), respectively. The two points marked as (1, 6, 9,12, 15, 18) and (1, 6 11, 18, 27, 31) represent the reported ground states by Chen[21] and Garg[23], respectively.

Summary

We have developed a nomenclature to enhance structural recognition and adopted an extended cluster expansion to describe the structural stabilities, which is good agreement with the results from the first-principles calculations. Unlike the conventional cluster expansion, the interaction parameters are derived from the enumeration of $C_{60-n}B_n$ ($n = 1 \sim 4$), where there are only 4 coefficients to be fitted for the composition consideration. Notably, we have found the stable isomers of $C_{57}B_3$, $C_{55}B_5$, and $C_{54}B_6$, which are energetically favored by at least 0.3 eV than the reported counterparts. With the symmetry operation matrices, the nomenclature can be applied for other binary/ternary systems, where the ground state structures are searched with the extended cluster expansion. Thus, our finding will be an effective complement to the first-principles calculations in materials science.

References

1. Kroto, H. W., Heath, J. R., O'Brien, S. C., Curl, R. F. & Smalley, R. E. C_{60}: Buckminsterfullerene. *Nature* **318**, 162–163 (1985).
2. Joachim, C., Gimzewski, J. K., Schlittler, R. R. & Chavy, C. Electronic Transparence of a Single C_{60} Molecule. *Physical review letters* **74**, 2102–2105 (1995).
3. Sakurai, T. *et al.* Scanning tunneling microscopy study of fullerenes. *Progress in Surface Science* **51**, 263–408 (1996).
4. Haddock, J. N., Zhang, X., Domercq, B. & Kippelen, B. Fullerene based n-type organic thin-film transistors. *Organic Electronics* **6**, 182–187 (2005).
5. Wöbkenberg, P. H. *et al.* High mobility n-channel organic field-effect transistors based on soluble C_{60} and C_{70} fullerene derivatives. *Synthetic Metals* **158**, 468–472 (2008).
6. Yakuphanoglu, F. Electrical conductivity, optical and metal–semiconductor contact properties of organic semiconductor based on MEH-PPV/fullerene blend. *Journal of Physics and Chemistry of Solids* **69**, 949–954 (2008).
7. Moriarty, P. J. Fullerene adsorption on semiconductor surfaces. *Surface Science Reports* **65**, 175–227 (2010).
8. Ray, C. *et al.* Synthesis and Structure of Silicon-doped Heterofullerenes. *Physical review letters* **80**, 5365–5368 (1998).
9. Hirsch, A. & Nuber, B. Nitrogen Heterofullerenes. *Accounts of Chemical Research* **32**, 795–804 (1999).
10. László, F. & László, M. Electronic properties of doped fullerenes. *Reports on Progress in Physics* **64**, 649 (2001).
11. Pichler, T. *et al.* On-ball doping of fullerenes: The electronic structure of $C_{59}N$ dimers from experiment and theory. *Physical review letters* **78**, 4249–4252 (1997).
12. Hummelen, J. C., Knight, B., Pavlovich, J., Gonzalez, R. & Wudl, F. Isolation of the heterofullerene C_{59N} as its dimer $(C5_{9N}2)$. *Science* **269**, 1554–1556 (1995).
13. Hultman, L. *et al.* Cross-linked nano-onions of carbon nitride in the solid phase: Existence of a novel $C_{48}N_{12}$ aza-fullerene. *Physical review letters* **87**, art. no.-225503 (2001).
14. Schultz, D., Droppa, R., Alvarez, F. & dos Santos, M. C. Stability of small carbon-nitride heterofullerenes. *Physical review letters* **90** (2003).
15. Guo, T., Jin, C. & Smalley, R. E. Doping bucky: formation and properties of boron-doped buckminsterfullerene. *The Journal of Physical Chemistry* **95**, 4948–4950 (1991).
16. Yu, R. *et al.* Simultaneous Synthesis of Carbon Nanotubes and Nitrogen-Doped Fullerenes in Nitrogen Atmosphere. *The Journal of Physical Chemistry* **99**, 1818–1819 (1995).
17. Butcher, M. J. *et al.* $C_{59}N$ monomers: Stabilization through immobilization. *Physical review letters* **83**, 3478–3481 (1999).
18. Jin, Z. *et al.* Single $C_{59}N$ molecule as a molecular rectifier. *Physical review letters* **95**, 045502/045501–045504 (2005).
19. Kurita, N., Kobayashi, K., Kumahora, H., Tago, K. & Ozawa, K. Molecular structures, binding energies and electronic properties of dopyballs $C_{59}X$ (X = B, N and S). *Chemical Physics Letters* **198**, 95–99 (1992).
20. Miyamoto, Y., Hamada, N., Oshiyama, A. & Saito, S. Electronic structures of solid BC_{59}. *Physical Review B* **46**, 1749–1753 (1992).
21. Chen, Z. F., Zhao, X. Z. & Tang, A. C. Theoretical studies of the substitution patterns in heterofullerenes $C_{60-x}N_x$ and $C_{60-x}B_x$ (x = 2-8). *Journal of Physical Chemistry A* **103**, 10961–10968 (1999).
22. Xie, R. H. *et al.* Tailorable acceptor $C_{60-n}B_n$ and donor $C_{60-m}N_m$ pairs for molecular electronics. *Physical review letters* **90** (2003).
23. Garg, I., Sharma, H., Dharamvir, K. & Jindal, V. K. Substitutional Patterns in Boron Doped Heterofullerenes $C_{60-n}B_n$ (n = 1-12). *Journal of Computational and Theoretical Nanoscience* **8**, 642–655 (2011).
24. Shinsaku, F. Socceranes Derivatives of Given Symmetries. *Bulletin of the Chemical Society of Japan* **64**, 3215–3223 (1991).
25. Cozzi, F., Powell, W. H. & Thilgen, C. Numbering of fullerenes - (IUPAC Recommendations 2005). *Pure and Applied Chemistry* **77**, 843–923 (2005).
26. Hedberg, K. *et al.* Bond lengths in free molecules of buckminsterfullerene, C_{60}, from gas-phase electron diffraction. *Science (New York, NY)* **254**, 410–412 (1991).
27. Qi, J., Zhu, H., Zheng, M. & Hu, X. Theoretical studies on characterization of heterofullerene $C_{58}B_2$ isomers by X-ray spectroscopy. *RSC Advances* **6**, 96752–96761 (2016).
28. Kroto, H. W. The stability of the fullerenes Cn, with n = 24, 28, 32, 36, 50, 60 and 70. *Nature* **329**, 529–531 (1987).
29. Thilgen, C. & Diederich, F. Structural Aspects of Fullerene Chemistry-A Journey through Fullerene Chirality. *Chemical Reviews* **106**, 5049–5135 (2006).
30. Balasubramanian, K. Enumeration of chiral and positional isomers of substituted fullerene cages (C_{20}-C_{70}). *The Journal of Physical Chemistry* **97**, 6990–6998 (1993).
31. Babic, D., Doslic, T., Klein, D. J & Misra, A. Kekulenoid addition patterns for fullerenes and some lower homologs. *Bulletin of the Chemical Society of Japan* **77**, 2003–2010 (2004).
32. Li, X.-T., Yang, X.-B. & Zhao, Y.-J. Geometrical eigen-subspace framework based molecular conformation representation for efficient structure recognition and comparison. *The Journal of Chemical Physics* **146**, 154108 (2017).
33. Sanchez, J. M., Ducastelle, F. & Gratias, D. Generalized cluster description of multicomponent systems. *Physica A* **128a**, 334–350 (1984).
34. Sluiter, M. H. F. & Kawazoe, Y. Cluster expansion method for adsorption: Application to hydrogen chemisorption on graphene. *Physical Review B* **68** (2003).
35. Hart, G. L. W., Blum, V., Walorski, M. J. & Zunger, A. Evolutionary approach for determining first-principles hamiltonians. *Nature Materials* **4**, 391–394 (2005).
36. Seko, A., Yuge, K., Oba, F., Kuwabara, A. & Tanaka, I. Prediction of ground-state structures and order-disorder phase transitions in II-III spinel oxides: A combined cluster-expansion method and first-principles study. *Physical Review B* **73** (2006).
37. Muzyk, M., Nguyen-Manh, D., Kurzydlowski, K. J., Baluc, N. L. & Dudarev, S. L. Phase stability, point defects, and elastic properties of W-V and W-Ta alloys. *Physical Review B* **84** (2011).
38. Nahas, S., Ghosh, B., Bhowmick, S. & Agarwal, A. First-principles cluster expansion study of functionalization of black phosphorene via fluorination and oxidation. *Physical Review B* **93** (2016).
39. van de Walle, A. & Ceder, G. Automating first-principles phase diagram calculations. *Journal of Phase Equilibria* **23**, 348 (2002).

Acknowledgements

Author Contributions

Y.C. and X.Y. designed the research. Y.C. did the computation. Y.C., J.L. and X.Y. analyzed the data and discussed the results. Y.C., Y.Z. and X.Y. wrote the manuscript. All the authors have reviewed and finalized the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-017-16469-0.

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.