Supporting Information

Atomic Structure of Boron Resolved using Machine Learning and Global Sampling

*Si-da Huang, Cheng Shang, Pei-Lin Kang, Zhi-Pan Liu**

Collaborative Innovation Center of Chemistry for Energy Material, Shanghai Key Laboratory of Molecular Catalysis and Innovative Materials, Key Laboratory of Computational Physical Science (Ministry of Education), Department of Chemistry, Fudan University, Shanghai 200433, China *email: *zpliu@fudan.edu.cn*

Table of Contents

1. SSW-NN methodology

1.1 HDNN architecture

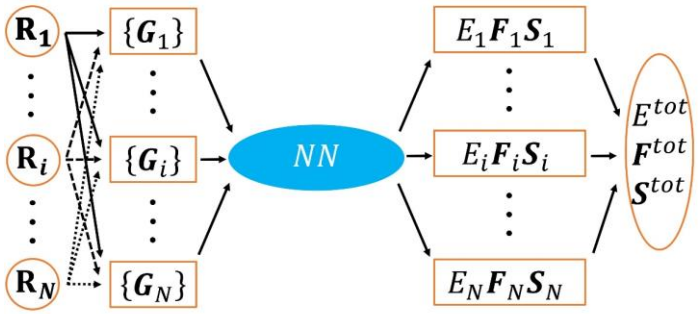


Fig. S1. Scheme for the HDNN architecture. The subscripts i and N represent atom indices and total number of atoms in a structure. The input of a NN is a set of structural descriptors $\{\mathbf{G}_i\}$ constructed from Cartesian coordinates $\{\mathbf{R}\}$ of a structure, and the outputs are the atomic properties $\{E_i, \boldsymbol{F}_i, \boldsymbol{S}_i\}$, i.e. energy, forces and stresses. The overall properties E^{tot}, $\boldsymbol{F}^{tot}$, and $\boldsymbol{S}^{tot}$, can be calculated from the individual atomic contributions.

In this work, we utilized the high dimensional neural network (HDNN) scheme to construct the NN[1, 2]. The NN architecture is schematically shown in Fig. S1. In Eq 1, the total energy E^{tot} can be decomposed and written as a linear combination of atomic energy E^i, which is the output of the standard neural network. The input nodes are a set of geometry-based structural descriptors, $\{\boldsymbol{G}_i\}$, and are very detailed discussed in main text.

$$E^{tot} = \sum_i E_i \,, (1)$$

The atomic force can be analytically derived according to Eq. 2, where the force component $F_{k,\alpha}$, α=x, y or z, acting on the atom k is the derivative of the total energy with respect to its coordinate $R_{k,\alpha}$. By combining with Eq. 1, the force component can be further related to the derivatives of the atomic energy with respect to j^{th} structural descriptors of atom i, $G_{j,i}$:

$$\mathrm{F_{k,\alpha}} = -\frac{\partial \mathrm{E^{tot}}}{\partial \mathrm{R_{k,\alpha}}} = -\sum_{\mathrm{i,j}} \frac{\partial \mathrm{E_i}}{\partial \mathrm{G_{j,i}}} \frac{\partial \mathrm{G_{j,i}}}{\partial \mathrm{R_{k,\alpha}}}, (2)$$

Similarly, the static stress tensor matrix element $\sigma_{\alpha\beta}$ can be analytically derived as:

$$\sigma_{\alpha\beta} = -\frac{1}{\mathrm{V}} \sum_{\mathrm{i,j,d}} \frac{(\mathbf{r_d})_\alpha (\mathbf{r_d})_\beta}{\mathrm{r_d}} \frac{\partial \mathrm{E_i}}{\partial \mathrm{G_{j,i}}} \frac{\partial \mathrm{G_{j,i}}}{\partial \mathrm{r_d}}, (3)$$

where $\mathbf{r}_d$ and r_d are the distance vector constituting of $G_{j,i}$ and its module, respectively, and V is the volume of the structure.

1.2 Constructing global dataset using SSW-NN

Undoubtedly, the dataset used for training the NN determines largely the quality of NN PES. Our previous work has shown that the stochastic surface walking (SSW) global optimzation[3, 4] can be used to fast generate a global dataset, which incorporates different structural patterns on the global PES. The details of SSW method can be found in the previous work. The SSW PES search is fully automated and does not need a priori knowledge on the system, such as the structure motif (e.g., bonding patterns, symmetry) of materials. The final obtained boron global dataset in this work is detailed in Table S1.

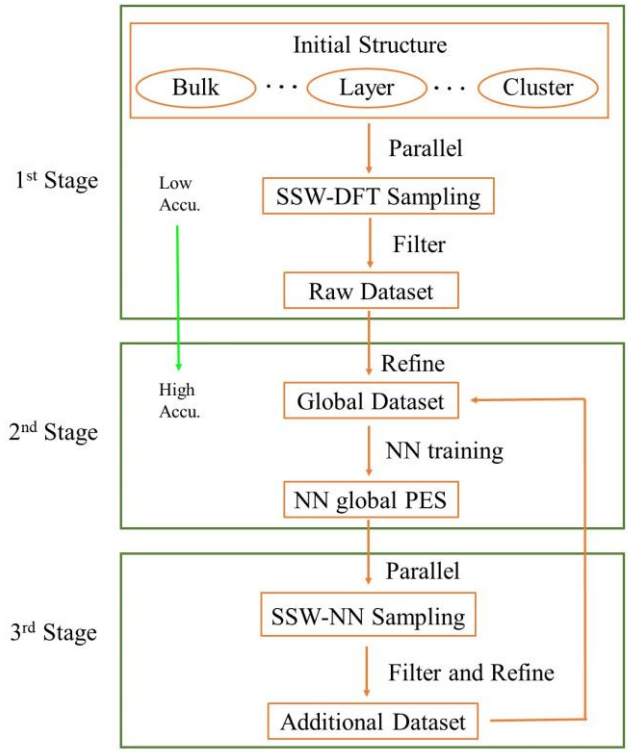
In brief, the SSW-NN method involves three stages for constructing the global dataset, as described in the following.

(i) **The first stage** constructs a raw dataset, which contains the most commmon atomic environment and serves as the training dataset for building an initial NN PES. This is done by performing first principles SSW global optimization in a massively parallel way. The first principles calculation is typically with low accuracy setups and restricted to small unit cells (typically below 20 atoms) to speed up the SSW search. By collecting and screening the structures from the SSW trajectories, a raw dataset is finally obtained.

(ii) **The second stage** trains an NN global PES. This is done by first refining the dataset using first principles calculation with high accuaracy setups, followed by the NN training on the accurate global dataset. The NN architecture applied in this stage utilizes a small set of structural descriptors and a small network size.

(iii) **The third stage** iteratively expands the global dataset. It targets to increase the predictive power of NN PES by incorporating more structural patterns into the dataset. This is done by carrying out SSW PES search using the NN PES obtained in the second stage, starting from a variety of initial structures. These initial structures are often randomly configurated and also include large systems with many atoms per unit cell (e.g. 100 atoms). The structures from all SSW

trajectories are collected and filtered to genereate the addtional dataset. This new dataset is then fed to the global dataset (back to stage 2) to start a new cycle of NN training.



Scheme. S1. Procedure for generating the training dataset using SSW global optimization. At the first stage, the SSW sampling is typically calculated with low accuracy first principle calculations. At the second stage, the global dataset is first refined using high accuracy setups, followed by the NN training on the accurate global dataset. At the third stage, an additional dataset is generated by SSW sampling utilizing NN PES obtained previously. This additional dataset is then fed into global dataset (back to stage 2) and start a new cycle of NN training.

Table S1. Structures in the first principles global dataset.

N_{atom}†	N_{bulk}‡	N_{layer}‡	$N_{cluster}$‡	N_{total}‡
12	32467	3037	8002	43506
14	51687	156	0	51843
28	7480	0	0	7480
40	469	15	21558	22042
13~52	8920	1441	0	10361
80	0	0	21333	21333
104~107	8858	0	0	8858
Total	109881	4649	50893	165423

†: the number of atoms per unit cell. The 13~52 entry has excluded those listed explicitly (12, 14, 28 and 40 atoms per cell).
‡: N_{bulk}, N_{layer}, $N_{cluster}$ are the number of structures belong to bulk, layer, and cluster types

1.3 Statistical assessment of structural descriptors for structure discrimination

The boron global dataset is an ideal test ground for structural descriptors. The dataset includes a vast amount of structures (165,423) ranging from bulk to clusters, where the number of boron atoms reaches to 5,334,657 in total. A critical task for us is to identify the most sensitive structural descriptors for distinguishing atomic environment.

For this purpose, we have to first generate a large number of distinct structural descriptors, which form a structural descriptor pool for selection. In total, 768 structural descriptors, belonging to the eight types (two Behler-type strcutural descriptors, BTSDs and six power-type strcutural descriptors, PTSDs), were generated by systematically adjusting the parameters in Eq. 3 to 11 (in the main test) with a fixed cutoff radius, i.e. 3.2 Å (the first and second nearest neighboring atomic environment for boron). There are 72, 64, 54, 45, 168, 180, 105, 80 for G^2, G^4, S^1, S^2, S^3, S^4, S^5 and S^6, respectively. For each structural descriptor, its value for each atom in the global dataset was calculated and scaled to (0, 1) according to the maximum and the minimum values. This finally generates a dataset matrix **X** with ($n \times p$) dimensionality, where the row (n) runs over all atoms and each column (p) gives the values from a particular structural descriptor in the dataset.

The principle component analysis (PCA)[5] statistics method was then utilized to compare statistically the structural descriptors. The PCA map each row vector $\boldsymbol{x}_{(i)}$ of dataset matrix **X** (i is the atom index) to a linearly uncorrelated new vector of principle component scores $\boldsymbol{t}_{(i)}$, given by the matrix transformation in Eq. S4, where the subscript L represents the possibility of dimensionality reduction by truncation to the first L loading vectors. To maximize variance, the k-th column of **W** with the element w_{jk} where j indexes the structural descriptor has to be the k-th eigenvector for the covariance matrix $\mathbf{X}^T\mathbf{X}$ of the dataset matrix **X**. It is useful to examine the weight w_{jk}^2 and the explained variance λ_k (Eq. S5) from PCA. The former represents the weight for the transformation from j-th structural descriptor to the k-th principal component; and the latter measures the dataset's variance that is projected onto the k-th principal component.

$$\boldsymbol{T_L} = \boldsymbol{XW_L} \quad \text{(S4)}$$

$$\lambda_k = \sum_i t_{ik}^2 \quad \text{(S5)}$$

$$S = \sum_k \lambda_k \times w_{jk}^2 \text{ , (S6)}$$

The above PCA were carried out in two steps. The structural descriptors belong to the same type were analyzed first by selecting L=1 to identify the one with the largest w_{j1}^2. The as-identified eight structural descriptors, one for each type, were again analyzed using PCA by selecting L=3 to compare their ability for structure discrimination. We found that the explained variance for the first three principle component are 80.9%, 12.6% and 3.0%, respectively, covering in total > 95% of the information from the structural descriptors. The transformation matrix element w_{jk}^2 from PCA in the first three principal components are listed in Table S2.

Table S2. The principal component analyses for different structural descriptors (SD) on the global dataset, showing the weights for the first three PCA components $wj1^2$ to w_{j3}^2 and the score (S) for each type of structural descriptors.

SD	Type†	n‡	w_{j1}^2/%	w_{j2}^2/%	w_{j3}^2/%	S
G^2	two	72	20.1	3.1	1.9	0.17
G^4	three	64	0.2	0.4	0.5	0.00
S^1	two	54	22.0	2.1	1.5	0.18
S^2	two	45	9.7	89.6	0.1	0.19
S^3	three	168	17.6	0.9	0.0	0.14
S^4	three	180	12.4	1.6	25.4	0.11
S^5	three	105	15.0	1.2	69.4	0.14
S^6	four	80	3.1	1.0	1.1	0.03

†: the type of the SD, i.e. two-, three- and four-body

‡: the total number of SDs in each type generated for screening out the one with the largest w_{j1}^2 at the first stage (see text).

In the first component, the w_{j1}^2 of two-body functions (G^2 and S^1, >20%) are generally larger than those of three-body functions (S^3, S^4 and S^5, ~15%), and both of them are much larger than that of four-body function (S^6, ~3%). This ordering is consistent with the common knowledge from classical force field, where the magnitude of the energy contribution follows stretching > bending > twisting. More importantly, for the second component, it is the S^2 (~90%) that dominates the weights,

and for the third component S^5 (~70%) prevails the other descriptors. These indicate that the S^2 and S^5 capture additional structural information, which can obviously be attributed to the incorporation of the spherical function in both cases.

Using Eq. S6, we can finally derive the score (*S*) for each structural descriptor by summing up their contributions in the first three principal components, which are also listed in Table S2. It tells that the two-body functions G^2 and S^1 achieve the similar scores, implying that they are indeed inter-replaceable (as suggested from Fig. 2 in main text). As for the three-body functions, the score for G_4 is considerably smaller than others (S^3, S^4 and S^5). This may not be surprising because the radial part of G^4 lacks of the ability to probe the atomic environment away from the atomic center.

In short, the PCA demonstrates that the new PTSDs outperform the BTSDs in the boron global dataset. In particular, it outlines the importance of S^2, S^4 and S^5 descriptors, which rank the top in two-body and three-body functions. Apparently, the incorporation of spherical harmonic function in S^2 and S^5 PTSDs enhances substantially the structures discrimination ability.

1.4 Performance of NN training with respect to the choice of structural descriptor and network size

Knowing the relative ranking of individual structural descriptors, we still need to identify the optimal set of structural descriptors that act together as the input to achieve the best performance in NN training. To do so, the network size, i.e. the number of training parameters in NN, must also be considered since it is well known that the fitting parameters can significantly affect the NN training speed and the predictive power.

While there are in principle infinite combinations by changing the structural descriptors and network size, we have experimented to setup two groups network, group N_A and group N_B, each with five different sets of structural descriptors. All the networks are fully-connected feedforward NN with two hidden layers. The two groups differ in the network parameters, as controlled by the number of neurons: group N_A has ~8500 network parameters and group N_B has ~25000 network parameters. The five sets of structural descriptors, denoted as set-1 to set-5, are as listed in Table S3, having 64, 90, 118, 130 and 152 structural descriptors. These ten different networks are therefore denoted as N_A-1 to N_A-5 for group N_A, and N_B-1 to N_B-5 for group N_B, and their details are summarized in Table S4. It is expected that the more types and the larger numbers of structural descriptors, as on going from set-1 to set-5, the better the atomic structure can be distinguished.

Table S3. The number of the structural descriptors (G^2 to S^6, in Eq. 3-11 in main text) in the set-1 to set-5.

set	G^2	G^4	S^1	S^2	S^3	S^4	S^5	S^6	Tot
1	0	6	23	9	6	20	0	0	64
2	8	0	29	9	22	22	0	0	90
3	6	11	29	10	28	25	0	9	118
4	6	11	29	22	28	25	0	9	130
5	6	11	29	22	28	25	22	9	152

Table S4. The network architecture in the standard notation of ten networks (Nt.) with the set 1 to 5 structural descriptors.

Nt.	architecture	Npara†	Nt.	architecture	Npara
N_A-1	64-65-65-1	8581	N_B-1	64-150-105-1	25711
N_A-2	90-61-51-1	8561	N_B-2	90-120-120-1	25561
N_A-3	118-50-48-1	8447	N_B-3	118-110-110-1	25411
N_A-4	130-47-47-1	8461	N_B-4	130-106-106-1	25335
N_A-5	152-43-42-1	8470	N_B-5	152-100-100-1	25501

†: the total number of parameters in the network.

To be more specific, the five sets of structural descriptors are mainly constituted by $S^1/S^2/S^3/S^4$ PTSDs, since they are more effective in distinguish boron structures demonstrated above. The G^2 and G^4 BTSDs are also selected to append to different sets: the set-1 with G^2 only, the set-2 with G4 only and the others with both G^2 and G^4. The four-body PTSD S^6 is added to the three largest sets, while the S^5 is only present in the set-5. It should be mentioned that at a given cutoff r_c, the S^5

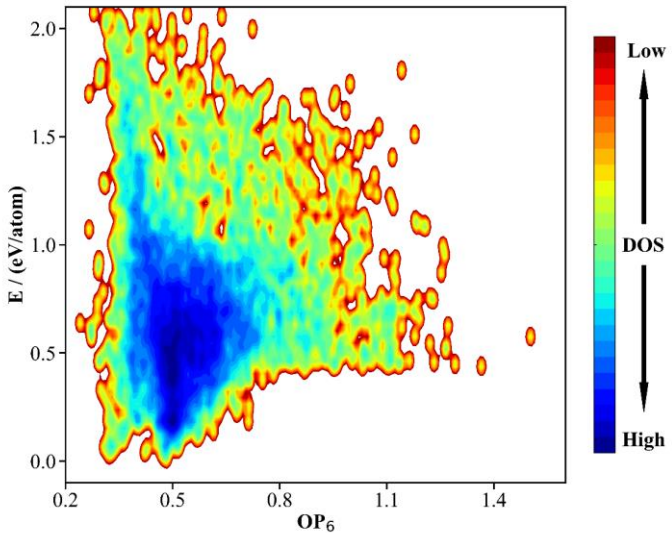


Fig. S2. The OP_6-E contour map of the selected dataset from global dataset for training purpose. OP_6 is the distance-weighted Steinhart order parameter in Eq. 1 (main text) with L=6 and the density of states (DOS) is indicated by color. The energy of α-B is set as energy zero.

and S^6 are typically the most expensive structural descriptors to compute and therefore they are only considered in the large sets. The parameters utilized for each structural descriptor have been tuned initially to ensure their coverage for different radius (up to 7 Å) and different angular distributions (one example of the parameters for structural descriptors in set-1 is shown in Table S5 to Table S9 in section 3).

For the purpose of the fast training of the ten different networks, we have generated a smaller but representative dataset. In total, 8,000 structures were taken randomly from the global dataset, which were then split into a training set (7,200 structures) and a test set (800 structures). Similarly, this dataset can be visualized, as shown in Fig. S2, by projected onto the distance-weighted Steinhart-type order parameter and the energy. Two figures, Fig. 1b (in main text) and Fig. S2, are indeed similar, indicating that this smaller dataset covers the whole global PES and is suitable for testing purpose. For future benchmarking in community, this 8000-dataset is also included in SI as seperated files, containing structural coordinates and energetics/forces/stresses data.

Our NN training procedure follows exactly that described in our previous work[2], where the performance function J_{tot} to measure the deviation of energy, force and stress, as defined in Eq. S7, is minimized using the L-BFGS algorithm[6, 7].

$$J_{tot} = J_E + \rho J_F + \tau J_\sigma = \frac{1}{2N}\sum_i \left(E_i^{NN} - E_i^{real}\right)^2 + \frac{\rho}{6N}\sum_{i,j,k}\left(F_{i,j,k}^{NN} - F_{i,j,k}^{real}\right)^2 + \frac{\tau}{18N}\sum_{i,\alpha,\beta}\left(\sigma_{i,\alpha,\beta}^{NN} - \sigma_{i,\alpha,\beta}^{real}\right)^2, (S7)$$

In the equation, i is structure indices, j is atom indices, $k = x$, y, or z, α, β are indices of stress matrix, E_i^{NN}, E_i^{real}, $F_{i,j,k}^{NN}$, $F_{i,j,k}^{real}$, $\sigma_{i,\alpha,\beta}^{NN}$, and $\sigma_{i,\alpha,\beta}^{real}$ are energy, forces, and stresses from NN and first-principles calculations, respectively. The adjustable parameters ρ and τ are set as constant, being 1000.0 and 4.0 in the training.

1.5 General rules for choosing the optimal network

Our main results for the NN training of the ten different networks are shown in Fig. S3. Fig. S3a and b plots the performance function J_{tot} of the training dataset against the training epoch for the small networks (N_A) and the large networks (N_B), respectively, and Fig. S3c plots the performance function of the test dataset against the training epoch for all networks. From the figures, we can obtain four general rules on the training of NN global PES, which are summarized as follows.

First, the fitting ability of the NN depends, not surprisingly, both on the structural descriptor and on the network size. The more complete the structural descriptor and the larger the network size is, the lower the performance function would be. This is in accordance with the general knowledge that the fitting capability and the network size are positively correlated in the fully-connected network. The lowest performance function for both group N_A and N_B occurs in the largest set-5, where the performance function reaches 43.6 for N_A -5 (purple line in Fig. S3a) and 31.7 for N_B-5 (purple line in Fig. S3b).

Second, with a given network size, the performance function for the training dataset will converge to some constant value when the structural descriptors approach to the complete. In the group N_A, the performance function converges to ~45 for N_A-

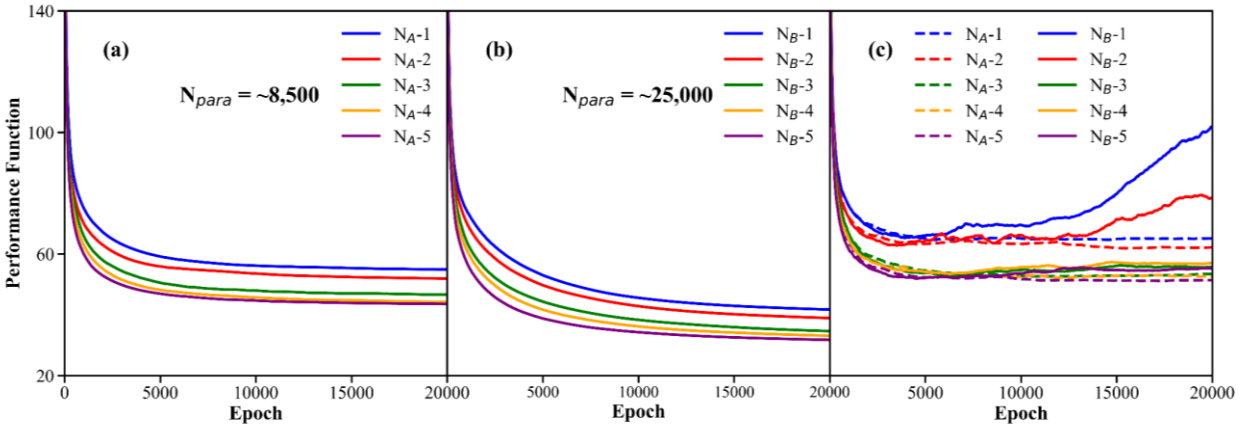


Fig. S3. The NN training, showing the evolution of performance function against training epoch. **(a)** Performance for the networks N_A-1 to N_A-5 in training dataset, N_{para} is the total number of parameters in the network; **(b)** Performance for the networks N_B-1 to N_B-5 in training dataset; **(c)** Performance of all networks in test dataset.

3 to N_A-5; in the group N_B, the performance function converges to ~32 for N_B-3 to N_B-5. This suggests the presence of a saturation limit for the structural descriptor set, when all structures in the dataset can be discriminated. Beyond the saturation limit, the further increase of the structural descriptors is effectively equivalent to the increase of the network size and has a less obvious effect on the overall performance.

Third, the performance function for the test dataset, even with the largest structural descriptor set and the largest network size, i.e. N_B-5, is always poorer compared to that for the training dataset. Importantly, this is unlike that in the training dataset, where the performance function always improves by continuously expanding the structural descriptor set or enlarging the network size. Fig. S3c shows that the performance function for the test dataset are very close for all large sets, N_A-3, N_A-4, N_A-5, N_B-3, N_B-4 and N_B-5. It suggests that the predictive ability of NN as measured by the performance function for the test dataset is hard to improve when the test dataset has no overlapping with the training dataset. In practice, one would have to incorporate as many as possible structural patterns in the training dataset to shrink the difference between training and test dataset.

Fourth, the performance function for the test dataset converges much rapidly with respect to the structural descriptor set and the network size. Fig. S3c shows that the performance function for the test dataset in all networks reaches to the plateau within 5000 training epochs, which is in contrast with the slow convergence of the performance function for the training set, typically beyond 10000 epochs. Excessively long training, on the other hand, could lead to the overfitting. As shown by the rapid increase in the performance function for N_B-1 and N_B-2 after 10000 epochs, these two networks with the small structural descriptor set but the large network size are obviously more vulnerable to the overfitting. By contrast, the performance function for N_B-3 to N_B-5 only increase slightly after long training, which implies that the overfitting can be overcome by using large structure descriptor set. In general, the large structure descriptor set and the small network size are preferable to avoid the overfitting.

1.6 DFT calculation setups

All DFT calculations are performed using the periodic plane wave method as implemented in the VASP package[8, 9]. The ionic core electrons are described using the projector augmented wave (PAW) pseudopotential[10]. The electron exchange and correlation effects are described by the GGA-PBE functional[11]. For the low accuracy calculations used in SSW sampling to collect raw data set, the kinetic energy cutoff for plance wave basis is 400 eV; the first Brillouin zone is sampled using the Monkhorst−Pack scheme[12] with a (3 × 3 × 3). For the high accuracy calculations used in single-point energy refinement calculations to build the global data set, these two key setups increase to 600 eV and (4 × 4 × 4), respectively.

References

1. J. Behler and M. Parrinello, *Phys. Rev. Lett.*, 2007, **98**, 146401.

2. S.-D. Huang, C. Shang, X.-J. Zhang and Z.-P. Liu, *Chem. Sci.*, 2017, **8**, 6327-6337.

3. C. Shang, X.-J. Zhang and Z.-P. Liu, *Phys. Chem. Chem. Phys.*, 2014, **16**, 17845-17856.

4. C. Shang and Z.-P. Liu, *J. Chem. Theory Comput.*, 2013, **9**, 1838-1845.

5. I. Jolliffe, *Wiley StatsRef: Statistics Reference Online*, 2014.

6. D. C. Liu and J. Nocedal, *Math. Program.*, 1989, **45**, 503-528.

7. J. Nocedal, *MATH. COMPUT.*, 1980, **35**, 773-782.

8. G. Kresse and J. Hafner, *Physical Review B*, 1993, **47**, 558.

9. G. Kresse and J. Furthmüller, *Computational Materials Science*, 1996, **6**, 15-50.

10. G. Kresse and D. Joubert, *Physical Review B*, 1999, **59**, 1758.

11. J. P. Perdew, K. Burke and M. Ernzerhof, *Physical review letters*, 1996, **77**, 3865.

12. H. J. Monkhorst and J. D. Pack, *Physical review B*, 1976, **13**, 5188.

2. Pair distribution function of α-B

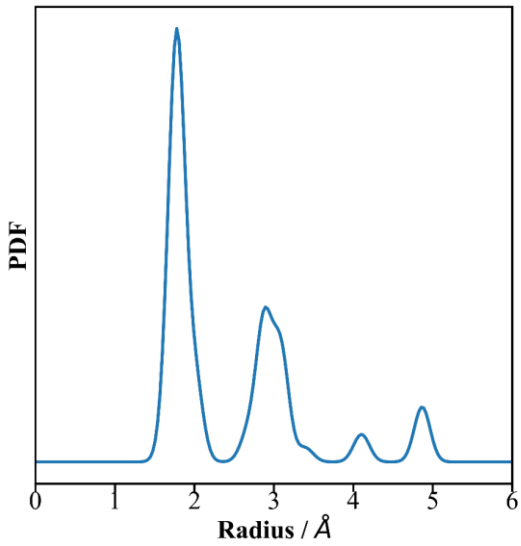


Fig. S4. Pair distribution function of α-B. The first shell B-B distance ranges from 1.67 to 2.01 Å.

3. Parameters for structural descriptors in set-1.

Table S5. Parameters of the angular structural descriptors G^4 (see Eq. 4 in main text) used to describe the atomic environment. r_c is cutoff radius, η, ζ and λ are adjustable parameters.

No.	r_c (Å)	η (Å^{-2})	λ	ζ	No.	r_c (Å)	η (Å^{-2})	λ	ζ
1	2.6	0.0500	2	1	4	3.7	0.0028	4	-1
2	2.6	0.1000	4	-1	5	6.0	0.0140	6	-1
3	3.7	0.0028	8	1	6	6.0	0.0140	4	1

Table S6. Parameters of the angular structural descriptors S^1 (see Eq. 6 in main text) used to describe the atomic environment. r_c and n are cutoff radius and adjustable parameter, respectively.

No.	r_c (Å)	n	No.	r_c (Å)	n
1	2.3	-1	13	2.9	24
2	3.3	-2	14	3.7	24
3	4.0	-2	15	4.0	24
4	6.5	-2	16	2.0	1
5	2.3	4	17	2.4	1
6	5.5	4	18	2.8	2
7	2.0	16	19	3.8	2
8	2.3	16	20	4.0	2
9	2.5	16	21	5.0	2
10	3.7	16	22	5.8	2
11	5.0	16	23	6.8	2
12	6.0	16	-	-	-

Table S7. Parameters of the radial structural descriptors S^2 (see Eq. 7 in main text) used to describe the atomic environment. L is adjustable parameter. Also see Table S2 caption for explanations.

No.	r_c (Å)	L	n	No.	r_c (Å)	L	n
1	2.2	2	6	6	3.5	-1	2
2	2.2	2	4	7	6.0	4	6
3	2.2	2	2	8	6.0	4	4
4	3.5	-1	6	9	6.0	4	2
5	3.5	-1	4	-	-		-

Table S8. Parameters of the angular structural descriptors S^3 (see Eq. 8 in main text) used to describe the atomic environment. r_c is the cutoff radius, n, m, ζ and λ are adjustable parameters.

No.	r_c (Å)	n	m	ζ	λ	No.	r_c (Å)	n	m	ζ	λ
1	2.2	-1	16	8	1	4	3.2	-2	16	8	-1
2	2.4	-2	16	8	-1	5	3.6	-2	16	8	-1
3	2.8	-2	16	8	1	6	4.0	-2	16	8	1

Table S9. Parameters of the angular structural descriptors S^4 (see Eq. 9 in main text) used to describe the atomic environment. p is the adjustable parameter. Also see Table S4 caption for explanations.

No.	r_c (Å)	n	m	p	ζ	λ	No.	r_c (Å)	n	m	p	ζ	λ
1	2.1	16	16	16	4	1	11	6.7	16	16	1	4	1
2	2.1	16	16	16	4	-1	12	6.7	16	16	1	4	-1
3	3.7	16	16	16	4	1	13	3.2	1	1	4	4	-1
4	3.7	16	16	16	4	-1	14	3.2	1	1	4	4	1
5	6.7	16	16	16	4	1	15	5.2	1	1	4	4	-1
6	6.7	16	16	16	4	-1	16	5.2	1	1	4	4	1
7	2.2	16	16	1	4	1	17	3.2	1	16	4	4	-1
8	2.3	16	16	1	4	-1	18	3.2	1	16	4	4	1
9	4.0	16	16	1	4	1	19	5.2	1	16	4	4	-1

4. Parameters for structural descriptors in final NN PES

Table S10. Parameters of the radial structural descriptors S^1 used to describe the atomic environment. Also see Table S2 caption for explanations.

No.	r_c (Å)	n	No.	r_c (Å)	n
1	1.9	0	21	3.4	16
2	2.2	0	22	3.9	2
3	1.8	16	23	4.4	2
4	2.1	16	24	4.9	2
5	1.6	8	25	5.4	2
6	1.9	8	26	5.9	2
7	2.2	8	27	6.4	2
8	1.6	2	28	3.9	8
9	1.9	2	29	4.4	8
10	2.2	2	30	4.9	8
11	2.6	2	31	5.4	8
12	2.9	2	32	5.9	8
13	3.4	2	33	6.4	8
14	2.2	-3	34	3.9	16
15	2.6	-3	35	4.4	16
16	2.9	-3	36	4.9	16
17	3.4	-3	37	5.4	16
18	2.2	16	38	5.9	16
19	2.6	16	39	6.4	16
20	2.9	16	-	-	-

Table S11. Parameters of the radial structural descriptors S^2 used to describe the atomic environment. Also see Table S3 caption for explanations.

No.	r_c (Å)	L	n	No.	r_c (Å)	L	n
1	1.7	2	6	19	4.1	2	2
2	1.7	8	6	20	4.6	2	6
3	1.8	2	4	21	5.1	2	2
4	2.1	2	4	22	5.7	2	6
5	2.4	2	4	23	6.2	2	2
6	2.7	2	4	24	6.7	2	6
7	3.0	2	4	25	1.8	8	2
8	3.6	2	4	26	2.1	8	6
9	4.1	2	4	27	2.4	8	2
10	4.6	2	4	28	2.7	8	6
11	5.1	2	4	29	3.0	8	2
12	5.7	2	4	30	3.6	8	6
13	1.8	2	2	31	4.1	8	2
14	2.1	2	6	32	4.6	8	6
15	2.4	2	2	33	5.1	8	2
16	2.7	2	6	34	5.7	8	6
17	3.0	2	2	35	6.2	8	2
18	3.6	2	6	36	6.7	8	6

Table S12. Parameters of the angular structural descriptors S^3 used to describe the atomic environment. Also see Table S4 caption for explanations.

No.	r_c (Å)	n	m	ζ	λ	No.	r_c (Å)	n	m	ζ	λ
1	1.8	2	8	8	-1	9	3.2	2	8	8	1
2	1.9	2	16	16	-1	10	3.2	2	8	8	-1
3	2.0	2	8	8	1	11	4.4	2	8	8	1
4	2.0	2	8	8	-1	12	4.4	2	8	8	-1
5	2.2	2	2	4	1	13	2.2	-3	8	8	1
6	2.2	2	2	4	-1	14	2.2	-3	8	8	-1
7	6.4	2	2	4	1	15	3.4	-3	8	8	1
8	6.4	2	2	4	-1	16	3.4	-3	8	8	-1

Table S13. Parameters of the angular structural descriptors S^4 used to describe the atomic environment. Also see Table S5 caption for explanations.

No.	r_c (Å)	n	m	p	ζ	λ	No.	r_c (Å)	n	m	p	ζ	λ
1	2.2	4	16	16	4	1	27	2.2	2	16	16	4	1
2	2.2	4	16	16	4	-1	28	2.2	2	16	16	4	-1
3	2.4	4	16	16	4	1	29	2.4	2	16	16	4	1
4	2.4	4	16	16	4	-1	30	2.4	2	16	16	4	-1
5	2.6	4	2	2	4	1	31	2.6	2	2	2	4	1
6	2.6	4	2	2	4	-1	32	2.6	2	2	2	4	-1
7	2.8	4	2	8	4	1	33	2.8	2	2	8	4	1
8	2.8	4	2	8	4	-1	34	2.8	2	2	8	4	-1
9	3.0	4	2	2	4	1	35	3.0	2	2	2	4	1
10	3.0	4	2	2	4	-1	36	3.0	2	2	2	4	-1
11	3.2	4	2	8	4	1	37	3.2	2	2	8	4	1
12	3.2	4	2	8	4	-1	38	3.2	2	2	8	4	-1
13	3.5	4	2	4	4	1	39	3.5	2	2	4	4	1
14	3.5	4	4	4	4	-1	40	3.5	2	4	4	4	-1
15	3.8	4	-3	4	4	1	41	3.8	2	-3	4	4	1
16	3.8	4	-3	4	4	-1	42	3.8	2	-3	4	4	-1
17	4.2	4	2	4	12	1	43	4.2	2	2	4	12	1
18	4.2	4	8	4	4	-1	44	4.2	2	8	4	4	-1
19	5.0	4	4	4	6	1	45	5.0	2	4	4	6	1
20	5.0	4	2	4	4	-1	46	5.0	2	2	4	4	-1
21	5.7	4	2	4	4	1	47	5.7	2	2	4	4	1
22	5.7	4	4	4	4	-1	48	5.7	2	4	4	4	-1
23	6.0	4	2	16	4	1	49	6.0	2	2	16	4	1
24	6.0	4	4	16	4	-1	50	6.0	2	4	16	4	-1
25	6.4	4	4	4	4	1	51	6.4	2	4	4	4	1
26	6.4	4	8	4	4	-1	52	6.4	2	8	4	4	-1

Table S14. Parameters of the angular structural descriptors S^5 (see Eq. 10 in main text) used to describe the atomic environment. *m* and *p* are adjustable parameters. Also see Table S3 caption for explanations.

No.	r_c (Å)	L	n	m	p	No.	r_c (Å)	L	n	m	p
1	2.2	4	2	2	2	10	2.6	6	2	8	8
2	2.2	4	2	2	2	11	3.2	2	2	2	2
3	2.4	4	2	2	2	12	3.2	6	2	2	2
4	2.4	4	2	2	2	13	4.8	2	2	2	2
5	2.2	2	2	2	2	14	4.8	6	2	2	2
6	2.2	6	2	2	2	15	3.2	2	2	4	8
7	2.4	2	2	2	2	16	3.2	6	2	4	8
8	2.4	6	2	2	2	17	4.8	2	2	4	8
9	2.6	2	2	8	8	18	4.8	6	2	4	8

Table S15. Parameters of the angular structural descriptors S^6 (see Eq. 11 in main text) used to describe the atomic environment. Also see Table S5 caption for explanations.

No.	r_c (Å)	n	m	p	ζ	λ	No.	r_c (Å)	n	m	p	ζ	λ
1	1.9	2	8	8	4	1	7	2.8	2	2	-2	4	1
2	1.9	2	8	8	4	-1	8	2.8	2	2	-2	4	-1
3	2.2	2	2	8	4	1	9	3.2	2	2	2	8	1
4	2.2	2	2	8	4	-1	10	3.2	2	2	2	8	-1
5	2.2	2	2	2	8	1	11	4.2	2	2	2	8	1
6	2.2	2	2	2	8	-1	12	4.2	2	2	2	8	-1

5. Phonon and free energy calculation

The phonon frequencies of the transition states are determined based on the numerical finite difference approach, employing the PHONOPY package[13]. For that, the size of the system was keep rhombohedral unit cell (1×1×1, 106/107-atom per cell) with the (8×8×8) Monkhorst-Pack mesh. With a displacement of ±0.01 Å of nonequivalent atoms, a set of displaced supercells was generated. For each minimum, the number of displaced cells is 636 (=106×6) for 106-atom cell and 642 (=107×6) for 107-atom cell. For each displaced cell, the DFT calculations were performed to obtain the force on each atom due to the displacements. These forces were carried back to the PHONOPY to calculate the phonon dispersion curves. Helmholtz free energy (F), entropy of vibration (S), and zero-point energy (ZPE) at finite temperature can be obtained from phonon spectra based on quasi-harmonic approximation. At a given temperature the lowest value of the free energy determines the stable phase. The free energy of the crystal is a sum of a ground state energy and the free energy contribution from the lattice vibrations. The first term is directly obtained from DFT calculation, i.e. at T=0 K. The second one is temperature dependent and in the harmonic approximation it is calculated from the phonon density of states using the following equation,

$$F_{harm} = rk_BT \int_0^{\infty} g(\bar{\omega})\, ln\left[2sinh\left(\frac{\hbar\bar{\omega}}{2k_BT}\right)\right] d\bar{\omega}$$

where r is the number of degree of freedom in a primitive unit cell, $\bar{\omega}$ denotes the phonon frequency, g($\bar{\omega}$) denotes the density of phonon states (DOS), $\hbar$ is the Planck constant, and k_B is the Boltzmann constant. In this approach the thermal expansion of crystal is neglected.

References

13. A. Togo, F. Oba and I. Tanaka, *Phys. Rev. B*, 2008, **78**, 134106.

6. Convergence of occupancy rates with respect to number of minima

The occupancy rates are calculated according to Eq.12-13 in the main text. To examine the numerical convergence in calculating POS occupancy, we have calculated the occupancy rates by using the top 100, 200 and 300 minima structures in the partition function summation. As shown in Table S16, the occupancy rates differ by no more than 0.1% for these results. Therefore, the numerical convergence for POS occupancy calculated from boron global PES is achieved with the 100~300 lowest energy structures.

Table S16. Occupancy rates for POSs from DFT and NN at different temperatures (1000, 1500 and 2000 K) and considered different number of structures (N_{str}).

	N_{str}	B13	B16	B17	B18	B19	B20	Res‡
1000K								
NN	100	68.3	29.5	15.0	15.0	1.5	1.2	0.5
NN	200	68.3	29.5	15.0	15.0	1.5	1.2	0.5
NN	300	68.3	29.4	15.0	15.0	1.5	1.2	0.5
DFT	100	70.1	26.8	13.2	13.2	1.6	2.4	0.6
DFT	200	70.1	26.8	13.2	13.2	1.6	2.4	0.6
DFT	300	70.1	26.8	13.2	13.2	1.6	2.4	0.6
1500K								
NN	100	71.9	23.7	11.5	11.5	4.7	2.2	2.1
NN	200	71.8	23.4	11.5	11.5	4.6	2.2	2.0
NN	300	71.8	23.3	11.6	11.5	4.6	2.3	2.0
DFT	100	73.2	22.8	10.1	10.1	4.6	2.8	2.3
DFT	200	73.2	22.8	10.1	10.1	4.6	2.8	2.3
DFT	300	73.2	22.8	10.1	10.1	4.6	2.8	2.3
2000K								
NN	100	74.4	20.7	8.9	8.9	7.0	2.3	3.2
NN	200	74.2	20.4	9.1	9.1	6.8	2.4	3.2
NN	300	74.1	20.2	9.1	9.1	6.8	2.4	3.2
DFT	100	75.0	20.8	8.2	8.2	6.7	2.7	3.7
DFT	200	75.0	20.7	8.2	8.1	6.7	2.7	3.7
DFT	300	75.0	20.7	8.1	8.1	6.7	2.7	3.7

‡: residual atoms with unknown location

7. Electronic analyses of β-I-15

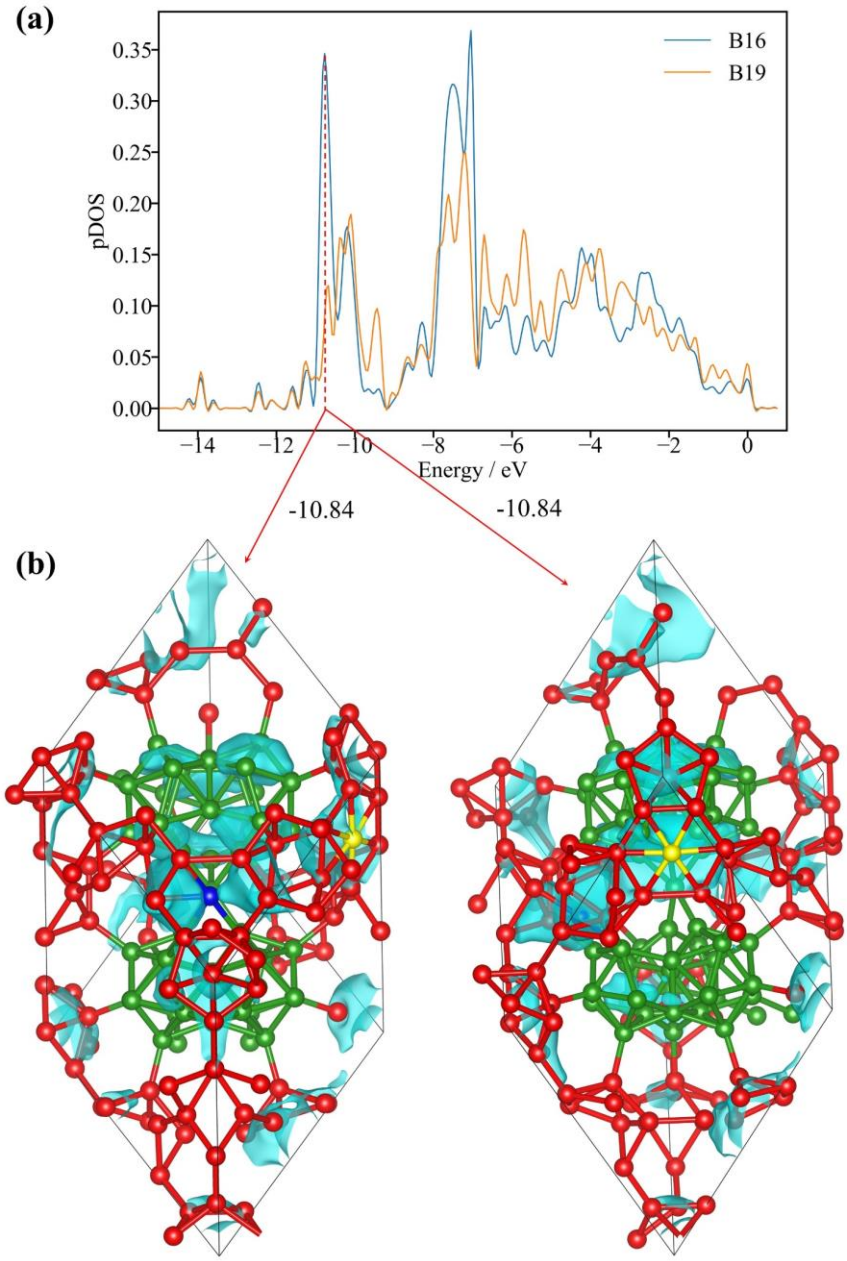


Fig. S5. (a) Projected density of states (pDOS) of the filling B16 (blue ball) and B19 (orange ball) *2p* orbital in β-I-15. The energy zero is set as valence band maximum (VBM). The center of B16 *2p* band occurs at -6.35eV below VBM, while that of B19 is at -6.02 eV below VBM. **(b)** The charge density (the square of wavefuntion) contour plot for the selected state at energy -10.84 (marked as red dotted line in **(a)**), *i.e.* the first major bonding peak for B16 and B19. Both the charge density are delocalized around the B16/B19 and also nearby atoms in B_{12} (red balls) or B_{28} (green balls) cages, which illustrates the the multi-center bonding for the doping B16/B19 atoms.