Supporting Information

Machine learning aided band gap prediction of semiconductors with low concentration doping

Yuqi Tang, Haiyuan Chen, Jianwei Wang* and Xiaobin Niu*

[a]School of Materials and Energy, University of Electronic Science and Technology of China, Chengdu 610054, P.R. China

*Corresponding author: jianwei_wang@uestc.edu.cn, xbniu@uestc.edu.cn

1. The details of the features used in present work

In present work, the constituent elements related features were omitted because the three elements in $GaAs_{1-x}N_x$ are fixed. We pay more attention to structure related features. There are three components in our features list: site encoding list, the maximum and the minimum distances between doping (nitrogen) atoms. The GaAs supercell has 64 atoms, thus there are 32 substitutable sites for nitrogen atoms. The substitutable sites are arranged sequentially according to their positions in the cartesian coordinate system, forming a list of 32 elements. The 32 elements are all composed of 0 and 1. When a substitutable site is replaced by a nitrogen atom, the corresponding value in the list changes from 0 to 1, otherwise all elements remain 0. For example, the site at coordinates (0.125, 0.125, 0.125) corresponds to the first element in the list, and when the site is replaced by a nitrogen atom, the list changes from [0 0 0 … 0] to [1 0 0 … 0]. In this way, the change of spatial configuration can be clearly reflected. Moreover, the distance between two defects will affect the electronic structure of doped systems. The maximum and the minimum values of the distance between two of all doping (nitrogen) atoms are taken as feature to complete our design of the features. The description of each part of the features is shown in Table S1.

Table S1 Description of each part of the features

Features	Description
Site encoding list	As shown in the GaAs primary cells in top of Fig. S1 (b), the four anion sites are labeled as 1, 2, 3 and 4, respectively. So, in the site encoding list, there are 4 elements and every element in the list has an index corresponding to the anion sites. For the bulk GaAs, the list is [0 0 0 0] because all the 4 As sites are not substituted by N atoms. If the third As site is replaced by N atom, then the list turns to [0 0 1 0].
The maximum distance between dopants (nitrogen atoms)	The maximum distance between dopants (nitrogen atoms). Special cases: For configurations with doping concentrations of 0 and 1/64, the maximum distance

	between doping atoms is 0 Å; For configurations with doping concentration of 2/64, the maximum and the minimum distances between doping atoms are equal.
The minimum distance between dopants (nitrogen atoms)	The minimum distance between dopants (nitrogen atoms). The special cases are similar to these for the maximum distance between dopants.

2. Features' design for n-atom supercells (n≥64)

As demonstrated in main manuscript, we designed a structure feature to capture the impurity configurations. In present machine learning (ML) models, the input features were designed without any constrains but were taken 64-atoms supercells as an example in our main text. A site encoding list, containing a set of 0 and 1, was designed as structure feature. The index of this list was encoded with possible sites. The feature of the site encoding list also has fractional values. In the following, we will demonstrate what the meaning of the fractional values in structure feature.

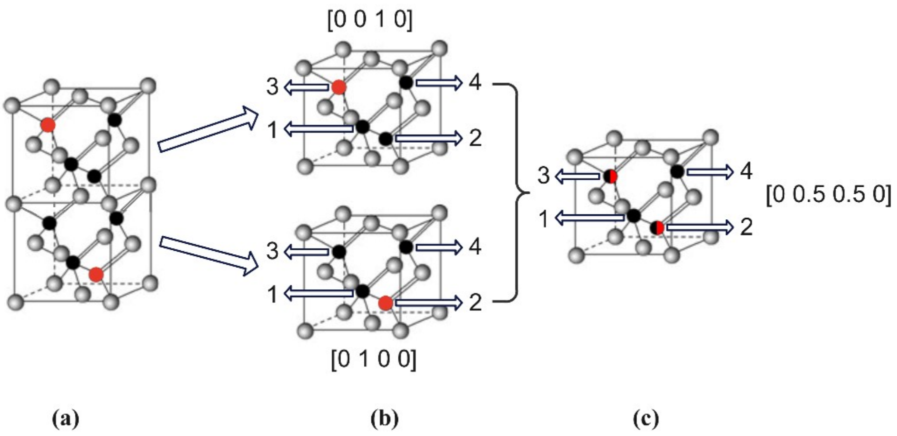

Fig. S1 Conversion of features input form from GaAs supercell of the 16-atom system into GaAs primary cell.

The 16-atom supercell can be recognized as twice of 8-atom unit cell. The first structure feature for 8-atoms unit cell is easily transformed to that for 16-atom supercell, which is shown in Table S2. The site 1, 2, 3, and 4 in 8-atom unit cell (up) and the site 1', 2', 3', and 4' in 8-atom unit cell (down) are collected and renewed as 1,2,3,4,5,6,7, and 8 in 16-atom supercell. As shown in Fig. S1(b), the site encoding list for up 8-atom unit cell and down 8-atom unit cell are [0 0 1 0] and [0 1 0 0], respectively. With the site numbers in up and down 8-atom unit cell (the index of site encoding list) are combined, the site encoding list combine together as [0 0 1 0 0 1 0 0], which is the site encoding list for 16-atom supercell. This is the transform in the forward direction as demonstrated in Table S2.

Table S2 The transform of site encoding list in the forward direction

Site encoding list	0	0	1	0	0	1	0	0

16-atom supercell	1	2	3	4	5	6	7	8
Site encoding list	0	0	1	0	0	1	0	0
8-atom unit cell	1	2	3	4	1’	2’	3’	4’

In the backward direction, if we have a feature of site encoding list [0 0 1 0 0 1 0 0] for 16-atom supercell, how can we get the corresponding feature of the site encoding list for 8-atom unit cell? The procedures are demonstrated as following.

Table S3 The transform of site encoding list in the backward direction

Site encoding list	0	0	1	0	0	1	0	0
16-atom supercell	1	2	3	4	5	6	7	8
Site encoding list	0	0	0.5	0	0	0.5	0	0
8-atom unit cell	1	2	3	4	1’	2’	3’	4’
Site encoding list	0	0.5	0.5	0				
Merging 8-atom unit cell	1	2	3	4				

First, the 16-atom supercell can be thought as twice of 8-atom unit cell. We can get site encoding lists for the two 8-atom unit cell: [0 0 1 0] and [0 1 0 0].

Second, to keep the concentration fixed at 1/16, the values in the site encoding list for 8-atom unit cell should be divided by a factor 2. Here the factor 2 comes from the ratio between atomic numbers of 16-atom supercell and that of 8-atom unit cell. So we get two renewed site encoding lists [0 0 0.5 0] and [0 0.5 0 0] for two 8-atom unit cells. The factor is to keep the doping concentration fixed at 1/16.

Third, the renewed site encoding lists for two 8-atom unit cell ([0 0 0.5 0] and [0 0.5 0 0]) are merged as one site encoding list ([0 0.5 0.5 0]) for 8-atom unit cell just by adding them directly like vectors.

All the procedures are demonstrated in Table S3.

The other two feature, the maximum and minimum distance between the dopants in the 16-atom supercell are used. Similarly, the features input of the 128-atom supercell can be converted to the features input of 64-atom supercell by this way.