


**Unified theory of direct or indirect band-gap nature of conventional semiconductors**Lin-Ding Yuan,<sup>1,2</sup> Hui-Xiong Deng,<sup>1,2</sup> Shu-Shen Li,<sup>1,2,3</sup> Su-Huai Wei,<sup>4,\*</sup> and Jun-Wei Luo<sup>1,2,3,†</sup><sup>1</sup>State Key Laboratory of Superlattices and Microstructures, Institute of Semiconductors, Chinese Academy of Sciences, Beijing 100083, China<sup>2</sup>College of Materials Science and Opto-Electronic Technology, University of Chinese Academy of Sciences, Beijing 100049, China<sup>3</sup>Beijing Academy of Quantum Information Sciences, Beijing 100193, China<sup>4</sup>Beijing Computational Science Research Center, Beijing 100193, China (Received 28 February 2018; revised manuscript received 23 November 2018; published 26 December 2018)

Although the direct or indirect nature of the band-gap transition is an essential parameter of semiconductors for optoelectronic applications, the reasons for why some of the conventional semiconductors have direct or indirect band gaps remains ambiguous. In this paper, we reveal that the existence of the occupied cation  $d$  bands is a prime element in determining the directness of the band gap of semiconductors through the  $s$ - $d$  and  $p$ - $d$  couplings, which push the conduction band energy levels at the  $X$  and  $L$  valley up, but leave the  $\Gamma$ -valley conduction state unchanged. This unified theory unambiguously explains why diamond, Si, Ge, and Al-containing group III-V semiconductors, which do not have active occupied  $d$  bands, have indirect band gaps, and the remaining common semiconductors, except GaP, have direct band gaps. Besides  $s$ - $d$  and  $p$ - $d$  couplings, bond length and electronegativity of anions are two remaining factors regulating the energy ordering of the  $\Gamma$ ,  $X$ , and  $L$  valleys of the conduction band, and are responsible for the anomalous band-gap behaviors in GaN, GaP, and GaAs that have direct, indirect, and direct band gaps, respectively, despite the fact that N, P, and As are in ascending order of the atomic number. This understanding will shed light on the design of direct band-gap light-emitting materials.

DOI: [10.1103/PhysRevB.98.245203](https://doi.org/10.1103/PhysRevB.98.245203)**I. INTRODUCTION**

Whether a semiconductor has a direct or indirect band gap is of fundamental importance to its optoelectronic applications [1,2]. If the conduction band minimum (CBM) occurs at the same point in  $k$  space as the valence band maximum (VBM), which is usually at the center ( $\Gamma$  point) of the Brillouin zone for conventional semiconductors, then the energy gap is referred to as a direct band gap, otherwise as an indirect band gap [2]. If a semiconductor has a direct band gap and the electric dipole transition from VBM to CBM is allowed, the electron-hole pairs will recombine radiatively with a high probability. As a result, high-quality direct band-gap semiconductors, such as GaAs and InP, are used to make highly efficient light emitters. They are essential materials for lasers, light-emitting diodes (LEDs), and other photonic devices [3,4], whereas, in indirect band-gap semiconductors, such as Si and Ge, optical transitions across an indirect band gap are not allowed, and, thus, these materials are not efficient light emitters. Despite the paramount importance of the direct or indirect nature of the band-gap transition for conventional semiconductors, which have been extensively studied in the past seven decades, the understanding of the formation of their direct or indirect band gaps remains ambiguous.

In spite of the extensive utilizing of sophisticated but opaque computational approaches like the semiempirical

pseudopotential method and first-principles density functional theory to correctly reproduce the experimentally measured band structures for semiconductors, the simple nearest-neighbor tight-binding (TB) theory is more straightforward to gain insight into the formation of the band structures because of its intuitive simplicity [5]. However, this simple TB model fails to reproduce some important band structure features, such as the band-gap nature for indirect semiconductors [2,6]. Although the introduction of additional unphysical parameters can cure the flaw of the simple  $sp^3$  TB model [7–10], it loses the advantage of its intuitive simplicity and thus is unlikely to uncover the origin of the direct and indirect band-gap natures of semiconductors. The poor understanding impedes the design of new direct band-gap light-emitting materials. Specifically, Si is ubiquitous in the electronics industry but is unsuitable for optoelectronic applications because it has an indirect band gap. In the past five decades, numerous ideas have been offered but failed to transform Si into an efficient light emitter [11,12] utilizing various modalities of symmetry reduction, including the use of porous silicon [13,14]; invoking alloy-induced luminescence [15,16]; and the method of low-symmetry allotropes of silicon [17–19], and diamondlike (III-V)-Si alloys [20,21]. Lack of fundamental understanding of the mechanism controlling the indirect band-gap nature of Si might be the main reason for the difficulty of developing Si-based direct band-gap materials.

Here, we reveal that the occupied cation  $d$  bands, which were neglected in previous models of the TB approach [5–10], play a prime role in forming the direct band gap of semiconductors via the  $s$ - $d$  and  $p$ - $d$  couplings. These couplings repel

\*suhuaiwei@csrc.ac.cn

†jwluo@semi.ac.cn

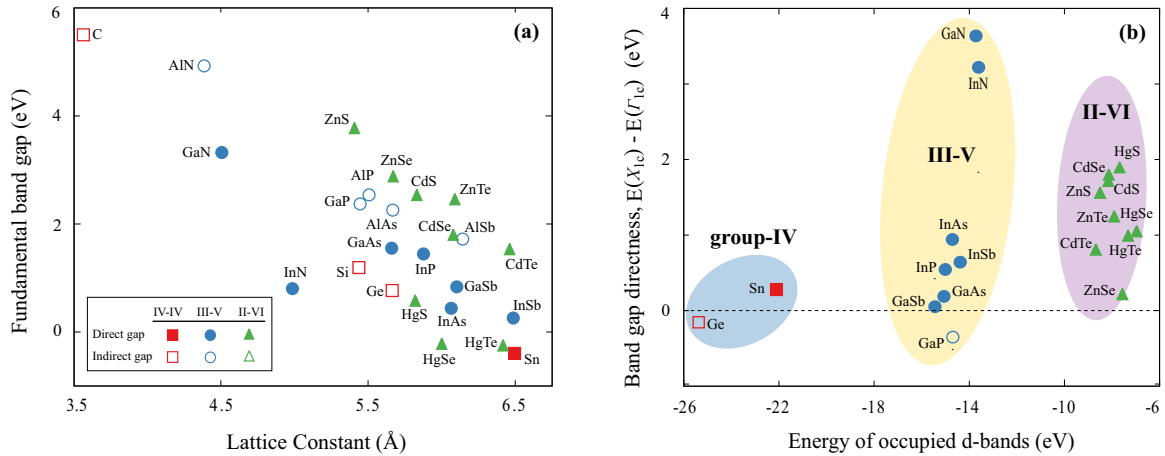


FIG. 1. (a) Fundamental band gaps of group II-VI, group III-V, and group IV semiconductors as a function of their lattice constants  $a$ . Filled symbols represent direct band-gap semiconductors and open symbols indirect band-gap semiconductors. On the other hand, the squares indicate the group IV semiconductors, the circles group III-V semiconductors, and the triangles group II-VI semiconductors. (b) The energy difference between levels of the CB  $X$  valley and  $\Gamma$  valley as a function of the energy of the cation  $d$  bands (relative to the VBM) for semiconductors consisting of cations having occupied  $d$  orbitals.

the conduction band energy levels of the  $X$  and  $L$  valleys up but leave the  $\Gamma$  valley intact. From group IV through group III-V to group II-VI semiconductors, the occupied cation  $d$  orbitals become closer in energy to the anion  $s$  and  $p$  orbitals, leading the  $s$ - $d$  and  $p$ - $d$  coupling to be strongest in group II-VI semiconductors, and hence all their band gaps are direct. Either the lack of, or the low-lying position of, the occupied  $d$  orbitals in cations of diamond, Si, Ge, and Al-containing group III-V semiconductors is responsible for their nature of indirect band gap.

## II. COMPUTATIONAL APPROACH

In this study, electronic structures are calculated utilizing density functional theory (DFT) [22–24] based first-principles methods within the general gradient approximation (GGA) [25] implemented in the Vienna *ab initio* simulation package (VASP) [26]. In the DFT calculations of conventional semiconductors, The projector augmented wave (PAW) pseudopotential [27,28] and Perdew, Burke, and Ernzerhof (PBE) functional [25] are employed with a plane-wave expansion up to 400 eV and a  $\Gamma$ -centered  $8 \times 8 \times 8$  Monkhorst-Pack [29]  $k$  mesh for the Brillouin zone sampling. Without shifting the qualitative results, we use the experiment lattice constants [30–32] for our models. To accurately reproduce the experimental results of band structures of conventional semiconductors, a modified Becke and Johnson exchange potential (mBJ) method [33] is adopted to improve the GGA description of band structures, which yields a comparable accuracy to  $GW$  methods but is computationally expensive compared with standard DFT calculations. To investigate the role of occupied  $d$  orbitals, we explicitly treat the occupied  $d$  shell of elements by considering them as valence states in the calculations. The obtained band gaps [see Fig. 1(a)] are in excellent agreement with experimental results. We employ the GGA +  $U$  method [34] to artificially tune the energy position of occupied  $d$  orbitals in semiconductors. Within this method, the energy level of the semicore  $d$  shell of cations can be tuned

by the on-site Coulomb interaction parameter  $U$ . This enables us to straightforwardly investigate the effect of cation  $d$  levels on the band edge states and further the band-gap properties.

## III. RESULTS AND DISCUSSION

### A. All direct band-gap semiconductors possessing occupied cation $d$ orbitals

We at first examine the nature of band gaps of all conventional group IV elemental, and group III-V and group II-VI compound semiconductors, which are the semiconductor of practical interest for information technology [2,4,35]. Figure 1 shows that all group II-VI compound semiconductors and the majority of group III-V compound semiconductors, except Al-containing compounds and GaP, have a direct band gap, whereas all group IV elemental materials, except gray Sn, are indirect band gap [35,36]. We note that the cations of all group II elements (Zn, Cd, and Hg), group III elements Ga and In, and group IV elements Ge and Sn contain occupied  $d$  orbitals, which, however, are absent in the remaining group III element Al and group IV elements C and Si. Because cation elements of all direct band-gap semiconductors encompass occupied  $d$  orbitals, whereas all semiconductors made of cations without occupied  $d$  orbitals have indirect band gaps, it strongly suggests that the occupied cationic  $d$  orbitals play a central role in determining the directness of the band gap for conventional semiconductors. However, the filled  $d$  shells, if they exist, are often treated as core or semicore shells and are usually neglected in the description of the band structures for conventional semiconductors in early studies [6–11]. To uncover the role of the cationic  $d$  orbitals in the formation of bands, we carried out the theoretical analysis of the band structures of diamond elements and zinc-blende (ZB) compounds relying on the perturbation theory along with the symmetry analysis. For simplicity, here we study only the band structures of compounds in the zinc-blende structure, even for GaN and ZnO, which prefer to be in the hexagonal wurtzite (WZ) structure under ambient conditions. The relationship

TABLE I. The point group of the wave vector at the  $\Gamma$ ,  $X$ , and  $L$  points in the zinc-blende structure and the corresponding irreducible representations of atomic  $s$ ,  $p$ , and  $d$  orbitals as well as semiconductor conduction (CBE) and valence (VBE) band edges under these point groups.

| $k$ Point | $G(k)$   | CBE ( $k$ )    | VBE ( $k$ )   | $s$        | $p$              | $d$                                    |
|-----------|----------|----------------|---------------|------------|------------------|--|
| $\Gamma$  | $T_d$    | $\Gamma_1$     | $\Gamma_{15}$ | $\Gamma_1$ | $\Gamma_{15}$    | $\Gamma_{15} \oplus \Gamma_{12}$       |
| $X$       | $D_{2d}$ | $X_1$ or $X_3$ | $X_5$         | $X_1$      | $X_3 \oplus X_5$ | $X_1 \oplus X_2 \oplus X_3 \oplus X_5$ |
| $L$       | $C_{3v}$ | $L_1$          | $L_3$         | $L_1$      | $L_1 \oplus L_3$ | $L_1 \oplus L_3 \oplus L_3$            |

between the band gaps of ZB and WZ structures are well studied [37]. For example, if the band gap is direct in the ZB structure, such as for GaN and ZnO, it is also direct in the WZ structure; if the band gap is indirect in the ZB structure, such as for AlN, it could still be direct in the WZ structure. Because the ZB  $X$  and  $L$  points fold to the same  $U$  points in the WZ structure, the eigenvalues at the  $U$  points in the WZ structure are the average of the ZB states at  $X$  and  $L$ . Such average raises the conduction band valley at the  $U$  points to higher than that at  $\Gamma$  and thus makes WZ AlN direct band gap [37].

### B. Symmetry enforced $s$ - $d$ and $p$ - $d$ coupling

The relative energy positions between the  $\Gamma$  valley,  $X$  valley, and  $L$  valley in the lowest conduction band determine the direct or indirect nature of the band gap in ZB semiconductors since the VBM occurs exclusively at the  $\Gamma$  point in all group II-VI, group III-V, and group IV semiconductors. As given in Table I [38], in the zinc-blende structure, the *CB edge at the  $\Gamma$  point* ( $\Gamma$  valley) transforms according to the  $\Gamma_1$  irreducible representation, the *VB edge* transforms according to the  $\Gamma_{15}$  irreducible representation, and the atomic  $d$  orbitals belong to the  $\Gamma_{15}$  and  $\Gamma_{12}$ , respectively. Since the  $d$  orbitals have the same  $\Gamma_{15}$  irreducible representation as the  $p$ -like VB edge state, the coupling between  $p$  and  $d$  orbitals at the  $\Gamma$  point could be quite significant. However, the  $s$ - $d$  coupling is forbidden because the atomic  $d$  orbitals have no common irreducible representations with the  $s$ -like CB  $\Gamma$ -valley state. Therefore, the existence of the occupied  $d$  orbitals will have a significant influence on the formation of the band gap by pushing the VBM up and leaving the CB  $\Gamma$  valley intact. The *CB edge at the  $X$  point* ( $X$  valley) transforms according to the  $X_1$  (or  $X_3$ ) irreducible representation of the  $D_{2d}$  wave-vector group, and is mainly derived from atomic  $s$  and  $p$  orbitals, whereas five  $d$  orbitals belong to the  $X_1$ ,  $X_2$ ,  $X_3$ , and  $X_5$  irreducible representations, respectively. Therefore, at the  $X$  point, the  $d$  orbital state can couple to the CB  $X$  valley. The mostly  *$s$ -like CB edge at the  $L$  point* ( $L$  valley) has the  $L_1$  representation of the  $C_{3v}$  wave-vector group, whereas the five  $d$  orbitals belong to the  $L_1$  and two  $L_3$  representations, respectively. Same as the  $X$  point, the  $d$  orbital state can couple to the CB  $L$  valley. Subsequently, the existence of the occupied  $d$  orbitals will repel the CB  $X$  valley and  $L$  valley up due to the symmetry allowed  $s$ - $d$  coupling and  $p$ - $d$  coupling at the  $X$  and  $L$  points, but will not affect the CB  $\Gamma$  valley owing to its lack of atomic  $p$  orbitals plus symmetry forbidden  $s$ - $d$  coupling at the  $\Gamma$  point (see Appendix C for diamond structure). Such  $s$ - $d$  coupling at the  $X$  and  $L$  points is evidenced by the fact that,

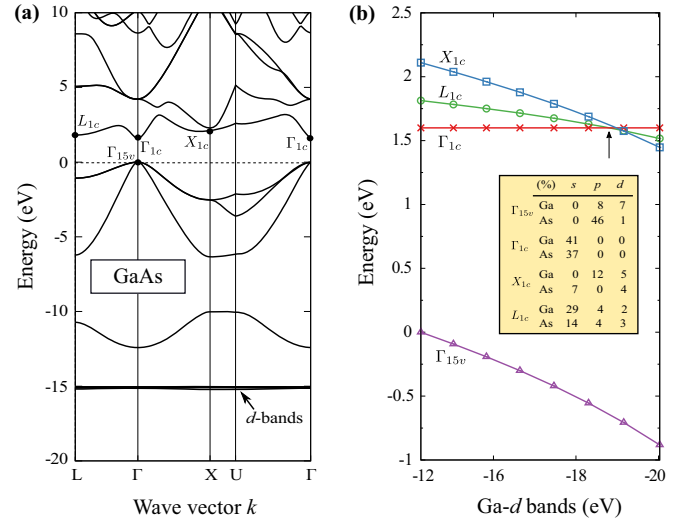


FIG. 2. Band structure of GaAs and the varying of its band edge levels as a function of the energy position of the Ga 3d bands. (a) The band structure of GaAs calculated using the mBJ-GGA approach. (b) The varying of the band edge levels of the CB  $\Gamma$ ,  $X$ , and  $L$  valleys and VBM as pulling the Ga 3d bands down through increasing the applied on-site Coulomb  $U$  on the Ga 3d orbitals relying on the mBJ-GGA approach. The vertical arrow indicates the transition between direct and indirect band gap because of the crossing between the  $\Gamma$  valley and the  $L$  valley. Inset of (b) summarizes the atomic orbital components of GaAs band edges at  $U = 0$  projected to spheres around each atom.

due to the low-lying  $s$  orbital energy of nitrogen, the  $s$ - $d$  coupling away from the  $\Gamma$  point is so strong in GaN and InN that leads to two  $s$ -like peaks observed in their photoemission near the bottom of the valence band [39]. This  $s$ - $d$  coupling not only has significant influence on the band structure near the bottom of the valence band but should also largely affect the conduction band at both the  $X$  and  $L$  points. Unfortunately, the latter has yet to be discovered even though a remarkable amount of the  $d$  character was found 30 years ago in the CB  $X$  and  $L$  valleys in conventional semiconductors [40].

### C. Effect of $s$ - $d$ and $p$ - $d$ coupling on conduction band edges

To illustrate the effect of the occupied  $d$  orbitals on the nature of the band gap, we examined the energy level shifting of the CB  $\Gamma$ ,  $X$ , and  $L$  valleys caused by the  $s$ - $d$  and  $p$ - $d$  coupling due to the existence of the occupied  $d$  shells, which were neglected previously, as schematically shown in Fig. 2(a). We took GaAs as the prototype to demonstrate the suggested unique role of the occupied cation  $d$  orbitals (the anion  $d$  orbitals are so low in energy that it is negligible compared with the cation  $d$  orbitals). Figure 2(b) shows the first-principles calculated GaAs band structure using the density functional theory (DFT) based on the modified Becke-Johnson (mBJ) exchange potential in combination with the generalized gradient approximation (GGA) correlation (mBJ-GGA). One can see that the Ga 3d bands with narrow bandwidths occur about 15 eV below the VBM. The interaction between  $d$  bands and the  $p$ -like VBM via  $p$ - $d$  coupling leads to a significant amount of the  $d$  character in the VBM [inset of Fig. 2(b)]. Since the

TABLE II. The theoretical predicted and experimentally measured energy positions of the occupied cation  $d$  bands in conventional semiconductors. The averaged energy position of the cation  $d$  bands is referred to the VBM for each semiconductor. The experiment results are deduced from the XPS data [44] with an accuracy within 1 eV.

| Group | Experiment or theory | Cation $d$ bands (eV) |       |       |       |       |       |
|-------|----------------------|-----------------------|-------|-------|-------|-------|-------|
| IV    |                      | Ge                    | Sn    |       |       |       |       |
|       | Theor.               | 24.67                 | 21.57 |       |       |       |       |
|       | Expt.                | 30                    | –     |       |       |       |       |
| III-V |                      | GaP                   | GaAs  | GaSb  | InP   | InAs  | InSb  |
|       | Theor.               | 14.51                 | 14.78 | 15.04 | 14.12 | 14.34 | 14.57 |
|       | Expt.                | 19                    | 19    | 18    | 17    | 17    | 17    |
| II-VI |                      | ZnS                   | ZnSe  | ZnTe  | CdS   | CdSe  | CdTe  |
|       | Theor.               | 7.33                  | 7.39  | 7.65  | 8.35  | 8.62  | 8.70  |
|       | Expt.                | 10                    | 10    | 11    | 11    | 12    | 12    |

$s$ - $d$  and  $p$ - $d$  couplings are allowed at the  $X$  and  $L$  points, the Ga  $3d$  bands repel the  $X$  and  $L$  valleys up in a significant amount of energy as evidenced by the incorporation of the finite  $d$  component in both the  $X_{1c}$  and  $L_{1c}$  states, in addition to expected dominated  $s$  and  $p$  components, whereas the inset of Fig. 2(b) shows the vanishing of the  $p$  and  $d$  characters in the  $\Gamma_{1c}$  state, confirming it to be purely an antibonding state of Ga  $4s$  and As  $4s$  and immune to the existence of the Ga  $3d$  bands.

To examine the effect of  $s$ - $d$  and  $p$ - $d$  coupling on the band edges, we artificially pull the Ga  $3d$  bands down to modify the  $s$ - $d$  and  $p$ - $d$  hybridizations. An adjustable Coulomb  $U$  acting on Ga  $3d$  orbitals is used as an effective knob to tune the energy position of Ga  $d$  levels by introducing Hubbard-type interactions into the DFT (DFT +  $U$ ). This method has been widely used to correct the underestimated DFT band gaps by

pulling the VBM down in energy through a  $p$ - $d$  coupling and leaving the CBM at  $\Gamma$  intact [41,42]. Here, we applied this method to investigate the impact of the Ga  $3d$  on the energies of the CB  $\Gamma$ ,  $X$ , and  $L$  valleys. Figure 2(b) shows the change of the energy positions of the  $X$  and  $L$  valleys, and VBM as varying the energy position of the Ga  $3d$  bands, regarding the  $\Gamma$  valley as an ideal reference level since, to the lowest order, it is free from the change of the  $d$  bands. As we pull the Ga  $3d$  bands down, the  $X$  and  $L$  valleys also move down in energy but in different rates, and, finally, the GaAs band gap becomes indirect, demonstrating the importance of the energy position of the occupied cation  $d$  bands in determining the nature of band gap unambiguously.

The energy position of the cation  $d$  bands relative to the (anion  $p$  dominated) VBM in conventional semiconductors depends mainly on the energy separation between the outermost anion  $p$  shell and the cation  $d$  shell, which decreases in the sequence of group IV, III-V, and II-VI semiconductors as shown in Fig. 3. Table II gives energy positions of the cation  $d$  bands relative to the VBM predicted by the mBJ-GGA calculations in comparison with experimental data [30–32,43] for those conventional semiconductors possessing occupied cation  $d$  orbitals. They are about  $-10$  eV for group II-VI compounds,  $-18$  eV for group III-V compounds, and  $-25$  eV for group IV elemental semiconductors, respectively. Going from group IV through group III-V to group II-VI semiconductors, the shallower cation  $d$  bands lead to stronger  $s$ - $d$  and  $p$ - $d$  couplings and hence larger repulsion of the  $X$  and  $L$  valleys by the low-lying occupied cation  $d$  bands. We compare the band structure for Ge, GaAs, and ZnSe with very similar lattice constants around  $5.65$  Å (shown in Fig. 4) and for Si, GaP, and ZnS with similar lattice constants around  $5.43$  Å (Fig. 5). In compared with GaAs, the enhanced  $s$ - $d$  and  $p$ - $d$  couplings in the group II-VI semiconductors repel the  $X$  and  $L$  valleys more and lead to the band gap being more direct, whereas in group IV Ge, the low-lying Ge  $4d$  bands result in weak  $s$ - $d$  and  $p$ - $d$  couplings so that its band gap becomes indirect. In the group IV diamond and Si and Al-based group III-V semiconductors, because the cation  $d$  orbitals are above the CB edges rather than in the occupied valence bands, the  $s$ - $d$  and  $p$ - $d$  couplings push the CB  $X$  and  $L$  valleys down instead of repelling them up, and subsequently make their band gap indirect.

Chemical trends of atomic energy levels

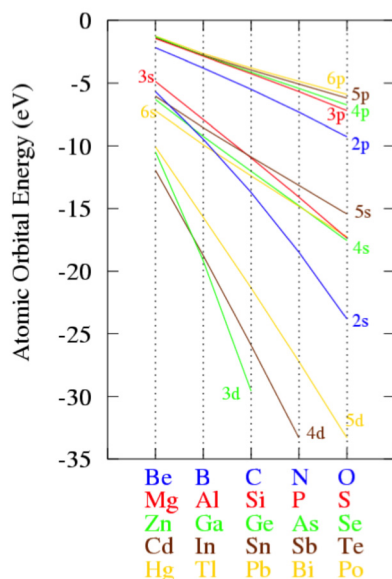


FIG. 3. Chemical trends of atomic energy levels predicted by using the local density approximation (LDA).

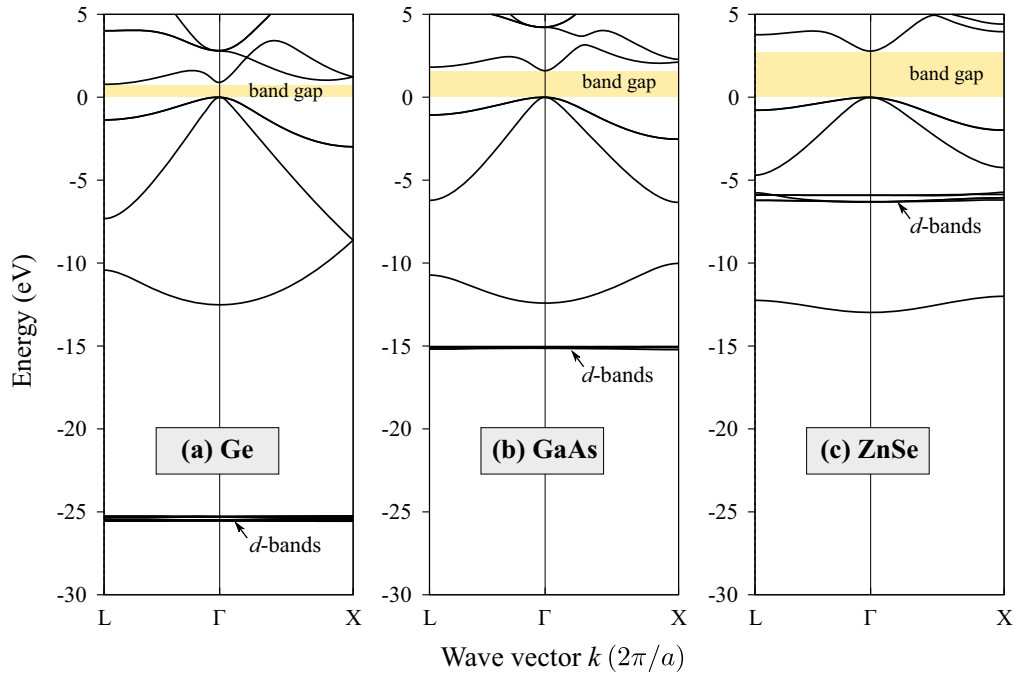


FIG. 4. Band structure of (a) Ge, (b) GaAs, and (c) ZnSe predicted by the first-principles mBJ-GGA approach without considering the spin-orbit interaction. Yellow area indicates the band gap. All energies are relative to the valence band maximum (VBM), and are set to zero. The lattice constants of Ge, GaAs, and ZnSe are very similar.

**D. Effect of remaining two factors**

Given that Ga possesses occupied 3*d* orbitals, the GaN and GaAs band gaps are, as expected, direct. However, GaP sitting in the middle between them is an indirect gap semiconductor as shown in Fig. 6. This abnormal band-gap behavior

indicates that besides the primary *s-d* and *p-d* couplings, other factors are also playing roles in determining the order of the  $\Gamma$ ,  $X$ , and  $L$  valleys in the lowest CB. We notice that the lattice constants of zinc-blende GaN, GaP, and GaAs are 4.531, 5.451, and 5.653 Å, respectively [36]. The bond length of GaP is  $\sim 3.6\%$  smaller than that of GaAs. Figure 6 shows

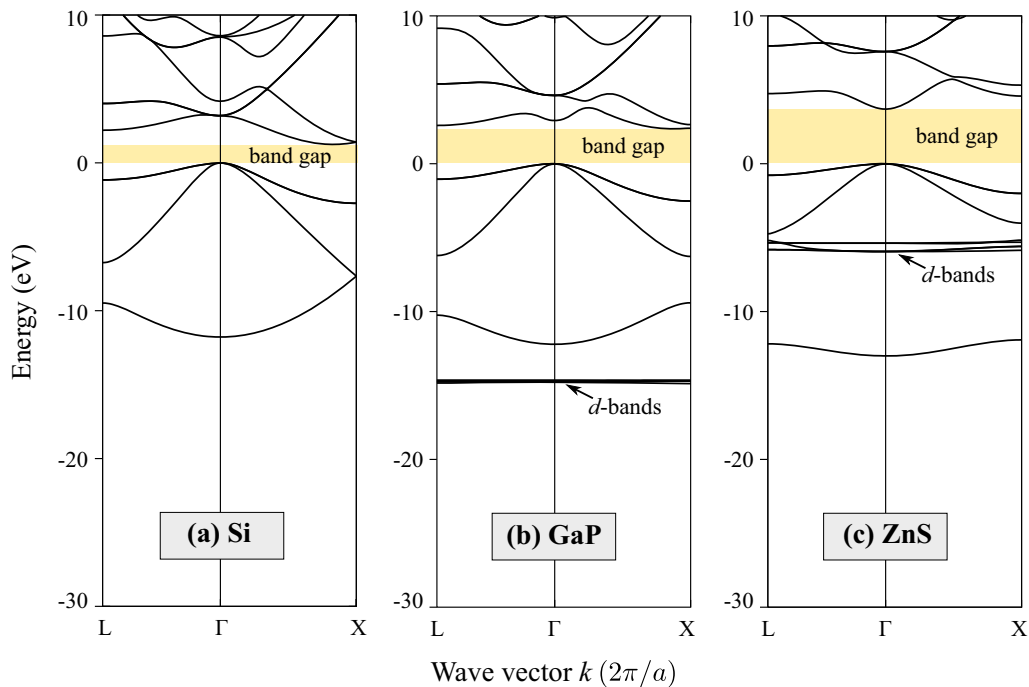


FIG. 5. Band structure of (a) Si, (b) GaP, and (c) ZnS predicted by the mBJ-GGA approach without considering the spin-orbit interaction. Yellow area indicates the band gap. All energies are relative to the valence band maximum (VBM), and are set to zero. The lattice constants of Si, GaP, and ZnS are very similar.



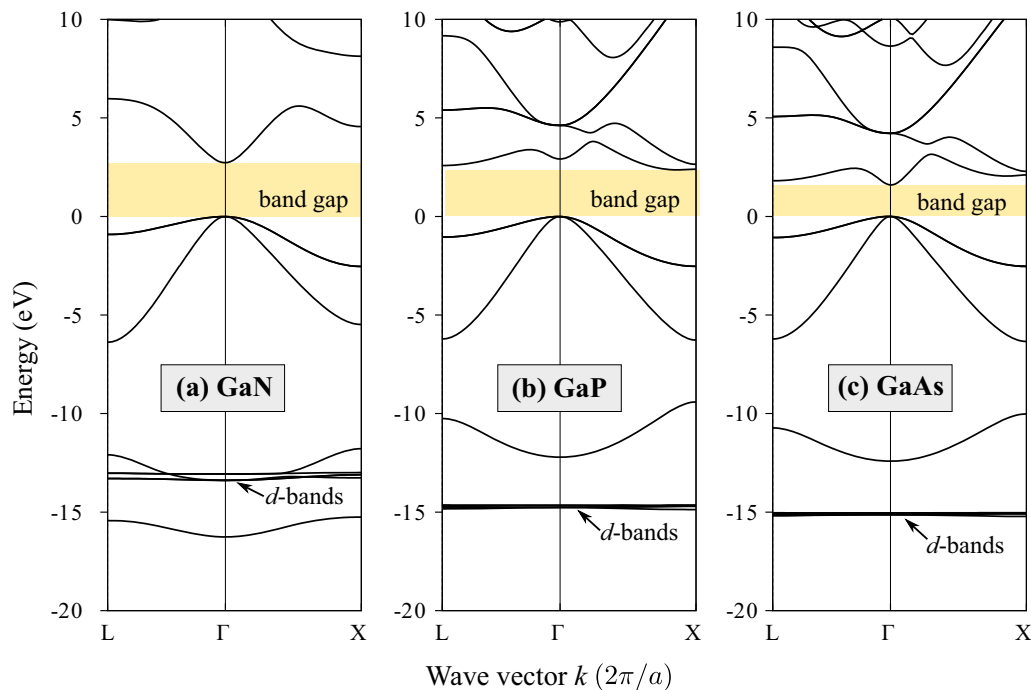


FIG. 6. Electronic structure of (a) GaN, (b) GaP, and (c) GaAs predicted by the mBJ-GGA approach without considering the spin-orbit interaction. Yellow area indicates the band gap. All energies are relative to the valence band maximum (VBM), and are set to zero.

that, in both GaP and GaAs, increasing the lattice constant (or expanding the volume) will raise the energy level of the  $X$  valley and lower the  $\Gamma$ -valley energy substantially. The  $L$  valley follows the  $\Gamma$  valley but at a much smaller rate and, thus, often sits in between the  $\Gamma$  and  $X$  valleys in conventional semiconductors. This phenomenon is owing to the  $X$  valley having a positive deformation potential and the  $\Gamma$  valley having a larger magnitude of negative deformation potential than that of the  $L$  valley, although both possess negative deformation potentials. If we stretch the lattice of GaP to equal

to that of GaAs, the GaP band gap would become direct. On the other hand, GaAs undergoes a direct-to-indirect band-gap transition as we compress the lattice of GaAs toward that of GaP. These demonstrate that besides  $s$ - $d$  and  $s$ - $p$  couplings, the bond length also plays a role in determining the nature of the band gap. Semiconductors having a larger lattice prefer to become more direct in the band gap (Fig. 7).

Following the above discussion, we would expect the band gap of GaN to be indirect since the bond length of GaN is much smaller than GaP and GaAs. However, GaN is a

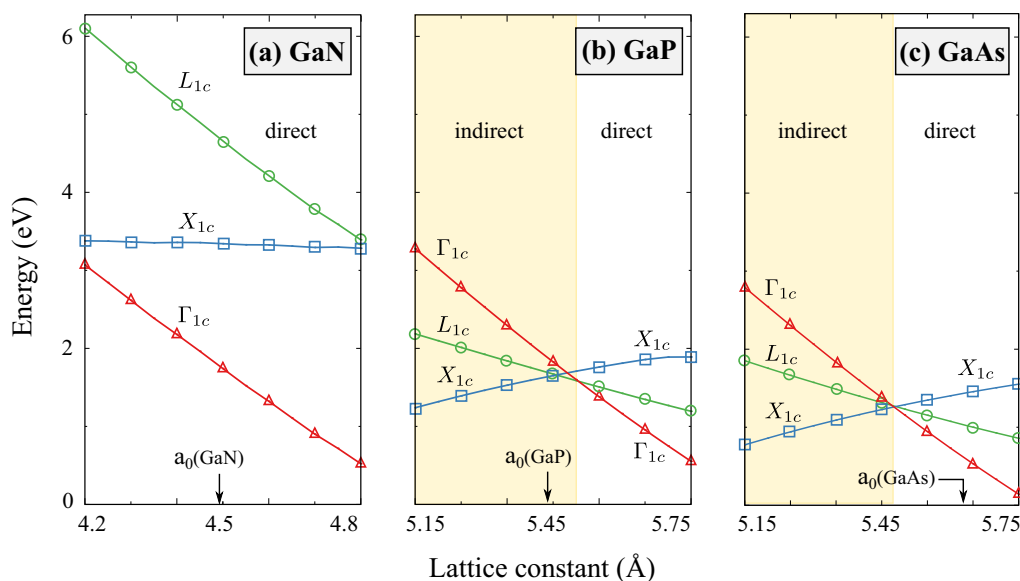


FIG. 7. Energies of the CB  $\Gamma$ ,  $X$ , and  $L$  valleys relative to the VBM as a function of the change of the lattice constant in (a) GaN, (b) GaP, and (c) GaAs. Arrows indicate the experimentally measured lattice constant  $a_0$  for each compound. The yellow areas mark the band gap becoming indirect.

classical direct gap semiconductor. Figure 7 shows that the deformation potentials of GaN are substantially different from that of GaP and GaAs; the energy level of the  $X$  valley is insensitive to the varying of the lattice constant with an inconsiderable negative deformation potential and the  $\Gamma$  valley drops at the same rate as that of the  $L$  valley as the lattice constant increases. AlN and InN share this exotic behavior as GaN [45]. This unusual behavior of the group III nitrides is due to nitrogen being among the most electronegative elements and much more electronegative than P and As elements. Figure 3 shows that the energy level of the N  $2s$  is 5.94 and 5.71 eV lower than that of the P  $3s$  and As  $4s$ , respectively. The low-lying N  $2s$  orbital is far away from the Ga  $3s$  orbital leading to a weak  $s$ - $s$  coupling in GaN according to the TB model [5] (see Appendix A for details), although it has a much smaller bond length than that of GaP and GaAs. Weak  $s$ - $s$  coupling results in the energy level of the  $\Gamma$  valley, which is the antibonding state of cation  $s$  and anion  $s$  orbitals, being slightly above the Ga  $3s$  level, and even lower than that of GaP and GaAs. Although the low-lying position of the N  $2s$  orbital also reduces the  $s$ - $p$  coupling and thus lowers the energy level of the  $X$  valley, the reduction in energy of the  $X$  valley is much less than the  $\Gamma$  valley due to the  $X$  valley having the lower bound limited by the atomic Ga  $3p$  level [5]. Consequently, GaN becomes a direct band-gap semiconductor [see Fig. 1(b)]. Thus, more electronegativity in the anions will also make the band gap of the semiconductors more direct, which is most significant in semiconductors containing O or N. Our analysis above indicates that despite the size of the atoms and electronegativity of the anions being able to play some roles in determining the directness of the band gap (especially for the boundary Ga compounds which have relatively deep  $3d$  orbitals), the coupling of the occupied cation  $d$  bands and unoccupied  $s$ ,  $p$  orbitals plays the prime role in determining the band-gap nature as manifested by the indirect gap of ZB AlN.

#### IV. SUMMARY

In summary, we have presented a unified theory for understanding the direct or indirect nature of the band gap in group II-VI, group III-V, and group IV semiconductors unambiguously. We found that the occupied cation  $d$  bands play a prime role in forming the direct or indirect band gap of conventional semiconductors via the  $s$ - $d$  and  $p$ - $d$  coupling with the states of the CB  $X$  and  $L$  valleys, which remarkably pushes their energy levels up, but leaves the  $\Gamma$  valley unchanged. From group IV through group III-V to group II-VI semiconductors, the occupied cation  $d$  orbitals become closer in energy to the anion  $s$  and  $p$  orbitals, leading the  $s$ - $d$  and  $p$ - $d$  coupling to be most active in group II-VI semiconductors, and hence all their band gaps are direct. Either the lack of, or the low-lying position of, the occupied  $d$  orbitals in cations of diamond, Si, Ge, and Al-containing group III-V semiconductors explains their nature of indirect band gap. This understanding will shed light on the design of direct band-gap light-emitting materials.

#### ACKNOWLEDGMENTS

J.-W.L. was supported by the National Natural Science Foundation of China (NSFC) under Grants No. 61474116 and No. 61888102 and the National Young 1000 Talents Plan. S.-H.W. was supported by the NSFC under Grants No. 51672023, No. 11634003, and No. U1530401, and the National Key Research and Development Program of China under Grant No. 2016YFB0700700.

#### APPENDIX A: SIMPLE NEAREST-NEIGHBOR $sp^3$ TB MODEL

In terms of the nearest-neighbor tight-binding approach, the formulas of the eigenvalues at the symmetry points  $\Gamma$ ,  $X$ , and  $L$  are given as follows [5,6]:

$$E(\Gamma_1) = \frac{\varepsilon_{s^+} + \varepsilon_{s^-}}{2} \pm \sqrt{\left(\frac{\varepsilon_{s^+} - \varepsilon_{s^-}}{2}\right)^2 + V_{ss}}, \quad (\text{A1})$$

$$E(\Gamma_{15}) = \frac{\varepsilon_{p^+} + \varepsilon_{p^-}}{2} \pm \sqrt{\left(\frac{\varepsilon_{p^+} - \varepsilon_{p^-}}{2}\right)^2 + \left(\frac{4}{3}V_{pp\sigma} + \frac{8}{3}V_{pp\pi}\right)^2}, \quad (\text{A2})$$

$$E(X_1) = \frac{\varepsilon_{s^+} + \varepsilon_{p'^-}}{2} \pm \sqrt{\left(\frac{\varepsilon_{s^+} - \varepsilon_{p'^-}}{2}\right)^2 + \frac{16}{3}V_{sp\sigma}^2}, \quad (\text{A3})$$

$$E(X_3) = \frac{\varepsilon_{s^-} + \varepsilon_{p'^+}}{2} \pm \sqrt{\left(\frac{\varepsilon_{s^-} - \varepsilon_{p'^+}}{2}\right)^2 + \frac{16}{3}V_{sp\sigma}^2}, \quad (\text{A4})$$

$$E(X_5) = \frac{\varepsilon_{p^+} + \varepsilon_{p^-}}{2} \pm \sqrt{\left(\frac{\varepsilon_{p^+} - \varepsilon_{p^-}}{2}\right)^2 + \left(\frac{4}{3}V_{pp\sigma} - \frac{4}{3}V_{pp\pi}\right)^2}, \quad (\text{A5})$$

$$E(L_3) = \frac{\varepsilon_{p'^+} + \varepsilon_{p'^-}}{2} \pm \sqrt{\left(\frac{\varepsilon_{p'^+} - \varepsilon_{p'^-}}{2}\right)^2 + \left(\frac{4}{3}V_{pp\sigma} + \frac{2}{3}V_{pp\pi}\right)^2}, \quad (\text{A6})$$

where  $V_{xx}$ ,  $V_{ss}$ , and  $V_{sp}$  are the coupling matrix element of the interaction Hamiltonian between the atomic orbitals and are usually referred to as the overlap parameters.

$$V_{ll'm} = \eta_{ll'm} \frac{\hbar^2}{md^2}, \quad (A7)$$

$$\frac{V_{s^*p\sigma}^2}{\varepsilon_p - \varepsilon_{s^*}} = \lambda_{sp\sigma} \frac{\hbar^2}{md^2},$$

and

$$\varepsilon_p - \varepsilon_s = \frac{9\pi^2}{16} \frac{\hbar^2}{md^2}, \quad (A8)$$

with

$$\begin{aligned} \eta_{ss\sigma} &= -9\pi^2/64, \\ \eta_{sp\sigma} &= 3\sqrt{3}\pi^2/64, \\ \eta_{pp\sigma} &= 3\pi^2/16, \\ \eta_{pp\pi} &= -3\pi^2/32, \\ \lambda_{sp\sigma} &= -27\pi^2/256. \end{aligned} \quad (A9)$$

#### APPENDIX B: SIMPLE TIGHT-BINDING THEORY FAILS TO TELL THE ORIGIN OF INDIRECT OR DIRECT NATURE IN BAND GAP

Although semiempirical and first-principles methods are often used to calculate and reproduce the experimentally measured band structures for semiconductors, the simple nearest-neighbor tight-binding (TB) theory based on the empirical bond orbital model pioneered by Harrison [5] is more straightforward to gain insight into the formation of the band structures because of its intuitive simplicity. It is well known that if only a minimal  $sp^3$  basis (one  $s$  orbital and three  $p$  orbitals for each atom) is used [6], this simple  $sp^3$  TB model, where the various tight-binding matrix elements are fitted to experiment or to more elaborate calculations, does not reproduce some important band structure features. For instance, the simple  $sp^3$  model can yield accurate valence band structures but fail in producing good conduction bands. For example, it predicts a rise in energy going from  $\Gamma$  to  $X$  whereas in the real case it can decrease. To remedy this deficiency, the addition of higher-lying *unoccupied* atomic states [7–10] is often used. In this case, atomic-site  $s$ - or  $d$ -like orbitals and bond-site “ $sp$ ” orbitals were introduced ad hoc as peripheral *higher-lying* atomic states on top of valence atomic  $s$  and  $p$  states, but neglecting the existence of low-lying occupied atomic  $d$  orbitals in many elements such as Ga, In, Zn, and Cd, to provide a good description of the conduction bands without modifying the valence bands much [7–10]. Such ad hoc high-energy excited  $d$  orbitals and/or  $s$  orbital [8] push the lowest conduction band down at the  $X$  and  $L$  points to correct the conduction band structure by adjusting the ad hoc  $s$  or  $d$  orbital energies and corresponding tight-binding matrixes

[7–9]. The introduction of such unphysical parameters can cure the flaw of the simple  $sp^3$  TB model to reproduce the conduction band structures of some semiconductors, but, in our opinion, fails to uncover the origin of the direct and indirect band-gap natures of these semiconductors.

#### APPENDIX C: SYMMETRY CONSIDERATIONS OF ZINC-BLENDE SEMICONDUCTORS

The electronic wave functions of a crystal at the point  $\mathbf{k}$  in reciprocal space are labeled by the irreducible representations of the symmetry group operations appropriate for  $\mathbf{k}$ , which is known as the group of the wave vector  $\mathbf{k}$ . The group of the  $\mathbf{k}$  vector at the  $\Gamma$  point or zone center is always the same as the point group of the crystal that is  $T_d$  for the zinc-blende structure [2,38], whereas the group of the  $\mathbf{k}$  vector at the  $L$  point is reduced to  $C_{3v}$  and at the  $X$  point to  $D_{2d}$  in the zinc-blende structure [2]. It is well known that in the zinc-blende semiconductors the VBM, which is always at the  $\Gamma$  point and composed mostly of anion  $p$  orbitals, belongs to the  $\Gamma_{15}$  irreducible representation of the  $T_d$  group. Here, we use single group representation since the spin-orbit coupling does not affect the directness of the band gap very much. The electronic state of the lowest CB at the  $\Gamma$  point belongs to  $\Gamma_1$  and is dominated by cation  $s$  orbitals. At the  $X$  point, it could be either the  $X_1$  state, which is a hybridization of cation  $s$  and anion  $p$  orbitals or the  $X_3$  state, which is a hybridization of cation  $p$  and anion  $s$  orbitals (here, we use the anion site as the origin [46]; for simplicity, we will use the  $X_1$  state in our discussion below). At the  $L$  point, it is the  $L_1$  state arising from the admixture of cation  $s$  and anion  $p$  orbitals as well as cation  $p$  and anion  $s$  orbitals.

At the  $\Gamma$  point, the atomic  $d$  orbitals belong to  $\Gamma_{15}$  and  $\Gamma_{12}$ , respectively, and thus have the same  $\Gamma_{15}$  irreducible representation as the  $p$ -like VB edge state; the coupling between  $p$  and  $d$  orbitals at the  $\Gamma$  point could be quite significant. However, the  $s$ - $d$  coupling is forbidden because the atomic  $d$  orbitals have no common irreducible representations with the  $s$ -like CB  $\Gamma$ -valley state. Therefore, the existence of the occupied  $d$  orbitals will have a significant influence on the formation of the band gap by pushing the VBM up and leaving the CB  $\Gamma$ -valley intact. At the  $X$  point, five  $d$  orbitals belong to the  $X_1$ ,  $X_2$ ,  $X_3$ , and  $X_5$  irreducible representations, respectively, and, therefore, can couple to the CB  $X$  valley. At the  $L$  point, five  $d$  orbitals belong to  $L_1$  and two  $L_3$ , respectively, and same as the  $X$  point, the  $d$  orbital state can couple to the CB  $L$  valley. It should be noted that for the diamond structure of the group IV elements, the group symmetry is  $O_h$ , different from the  $T_d$  point group of the zinc-blende structure. This could affect the coupling between the  $p$  and  $d$  orbitals. However, since these group IV elements in general have no occupied  $d$  orbitals—or if they have occupied  $d$  orbitals, they are deep—it would not affect our analysis if we treat them as materials with the reduced  $T_d$  symmetry group.

- [1] R. A. Soref, Silicon-based optoelectronics, *Proc. IEEE* **81**, 1687 (1993).  
 [2] P. Y. Yu and M. Cardona, *Fundamentals of Semiconductors: Physics and Materials Properties*, 3rd ed. (Springer, Berlin, 2005).

- [3] S. L. Chuang, *Physics of Photonic Devices*, Wiley Series in Pure and Applied Optics Vol. 80 (John Wiley & Sons, New York, 2012).  
 [4] L. A. Coldren, S. W. Corzine, and M. L. Mashanovitch, *Diode Lasers and Photonic Integrated Circuits*, Wiley Series in Mi-



- crowave and Optical Engineering Vol. 218 (John Wiley & Sons, New York, 2012).
- [5] W. A. Harrison, *Elementary Electronic Structure*, rev. ed. (World Scientific, Singapore, 2005).
- [6] D. J. Chadi and M. L. Cohen, Tight-binding calculations of the valence bands of diamond and zincblende crystals, *Phys. Status Solidi B* **68**, 405 (1975).
- [7] S. G. Louie, New localized-orbital method for calculating the electronic structure of molecules and solids: Covalent semiconductors, *Phys. Rev. B* **22**, 1933 (1980).
- [8] P. Vogl, H. P. Hjalmarson, and J. D. Dow, A Semi-empirical tight-binding theory of the electronic structure of semiconductors, *J. Phys. Chem. Solids* **44**, 365 (1983).
- [9] J.-M. Jancu, R. Scholz, F. Beltram, and F. Bassani, Empirical sp<sup>3</sup>\* tight-binding calculation for cubic semiconductors: General method and material parameters, *Phys. Rev. B* **57**, 6493 (1998).
- [10] Y. Tan, M. Povolotskiy, T. Kubis, T. B. Boykin, and G. Klimeck, Transferable tight-binding model for strained group IV and III-V materials and heterostructures, *Phys. Rev. B* **94**, 045311 (2016).
- [11] D. Liang and J. E. Bowers, Recent progress in lasers on silicon, *Nat. Photonics* **4**, 511 (2010).
- [12] L. Tsybeskov, D. J. Lockwood, and M. Ichikawa, Silicon Photonics: CMOS Going Optical [Scanning the Issue], *Proc. IEEE* **97**, 1161 (2009).
- [13] S. Furukawa and T. Miyasato, Quantum size effects on the optical band gap of microcrystalline Si:H, *Phys. Rev. B* **38**, 5726 (1988).
- [14] L. T. Canham, Silicon quantum wire array fabrication by electrochemical and chemical dissolution of wafers, *Appl. Phys. Lett.* **57**, 1046 (1990).
- [15] J. D. Gallagher, C. Xu, L. Jiang, J. Kouvetakis, and J. Menéndez, Fundamental band gap and direct-indirect crossover in Ge<sub>1-x-y</sub>Si<sub>x</sub>Sn<sub>y</sub> alloys, *Appl. Phys. Lett.* **103**, 202104 (2013).
- [16] S. Wirths, R. Geiger, N. von den Driesch, G. Mussler, T. Stoica, S. Mantl, Z. Ikonik, M. Luysberg, S. Chiussi, J. M. Hartmann *et al.*, Lasing in direct-bandgap GeSn alloy grown on Si, *Nat. Photonics* **9**, 88 (2015).
- [17] P. Zhang, V. H. Crespi, E. Chang, S. G. Louie, and M. L. Cohen, Computational design of direct-bandgap semiconductors that lattice-match silicon, *Nature* **409**, 69 (2001).
- [18] H. J. Xiang, B. Huang, E. Kan, S.-H. Wei, and X. G. Gong, Towards Direct-Gap Silicon Phases by the Inverse Band Structure Design Approach, *Phys. Rev. Lett.* **110**, 118702 (2013).
- [19] D. Y. Kim, S. Stefanoski, O. O. Kurakevych, and T. A. Strobel, Synthesis of an open-framework allotrope of silicon, *Nat. Mater.* **14**, 169 (2014).
- [20] T. Watkins, A. V. Chizmeshya, L. Jiang, D. J. Smith, R. T. Beeler, G. Grzybowski, C. D. Poweleit, J. Menéndez, and J. Kouvetakis, Nanosynthesis routes to new tetrahedral crystalline solids: Silicon-like Si<sub>3</sub>AlP, *J. Am. Chem. Soc.* **133**, 16212 (2011).
- [21] J. Kouvetakis, A. V. Chizmeshya, L. Jiang, T. Watkins, G. Grzybowski, R. T. Beeler, C. Poweleit, and J. Menéndez, Monocrystalline Al(As<sub>1-x</sub>N<sub>x</sub>)Si<sub>3</sub> and Al(P<sub>1-x</sub>N<sub>x</sub>)<sub>y</sub>Si<sub>5-2y</sub> alloys with diamond-like structures: New chemical approaches to semiconductors lattice matched to Si, *Chem. Mater.* **24**, 3219 (2012).
- [22] P. Hohenberg and W. Kohn, Inhomogeneous electron gas, *Phys. Rev.* **136**, B864 (1964).
- [23] W. Kohn, A. D. Becke, and R. G. Parr, Density functional theory of electronic structure, *J. Phys. Chem.* **100**, 12974 (1996).
- [24] W. Kohn and L. J. Sham, Self-consistent equations including exchange and correlation effects, *Phys. Rev.* **140**, A1133 (1965).
- [25] J. P. Perdew, J. A. Chevary, S. H. Vosko, K. A. Jackson, M. R. Pederson, D. J. Singh, and C. Fiolhais, Atoms, molecules, solids, and surfaces: Applications of the generalized gradient approximation for exchange and correlation, *Phys. Rev. B* **46**, 6671 (1992).
- [26] G. Kresse and J. Furthmüller, Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set, *Comput. Mater. Sci.* **6**, 15 (1996).
- [27] P. E. Blöchl, Projector augmented-wave method, *Phys. Rev. B* **50**, 17953 (1994).
- [28] G. Kresse and D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method, *Phys. Rev. B* **59**, 1758 (1999).
- [29] H. J. Monkhorst and J. D. Pack, Special points for Brillouin-zone integrations, *Phys. Rev. B* **13**, 5188 (1976).
- [30] I. Vurgaftman, J. Meyer, and L. Ram-Mohan, Band parameters for III-V compound semiconductors and their alloys, *J. Appl. Phys.* **89**, 5815 (2001).
- [31] I. Vurgaftman and J. R. Meyer, Band parameters for nitrogen-containing semiconductors, *J. Appl. Phys.* **94**, 3675 (2003).
- [32] A. Fleszar and W. Hanke, Electronic structure of II<sup>B</sup>-VI semiconductors in the *GW* approximation, *Phys. Rev. B* **71**, 045207 (2005).
- [33] Y.-S. Kim, M. Marsman, G. Kresse, F. Tran, and P. Blaha, Towards efficient band structure and effective mass calculations for III-V direct band-gap semiconductors, *Phys. Rev. B* **82**, 205212 (2010).
- [34] A. I. Liechtenstein, V. I. Anisimov, and J. Zaanen, Density-functional theory and strong interactions: Orbital ordering in Mott-Hubbard insulators, *Phys. Rev. B* **52**, R5467(R) (1995).
- [35] S. M. Sze and K. K. Ng, *Physics of Semiconductor Devices* (John Wiley & Sons, New York, 2006).
- [36] O. Madelung, Numerical data and functional relationships in science and technology, New Series, 571 (1982).
- [37] C.-Y. Yeh, S.-H. Wei, and A. Zunger, Relationships between the band gaps of the zinc-blende and wurtzite modifications of semiconductors, *Phys. Rev. B* **50**, 2715 (1994).
- [38] S. L. Altmann and P. Herzog, *Point-Group Theory Tables* (Clarendon Press, Oxford, 1994).
- [39] C. Persson and A. Zunger, *s* - *d* coupling in zinc-blende semiconductors, *Phys. Rev. B* **68**, 073205 (2003).
- [40] S. L. Richardson, M. L. Cohen, S. G. Louie, and J. R. Chelikowsky, Electron charge densities at conduction-band edges of semiconductors, *Phys. Rev. B* **33**, 1177 (1986).
- [41] T. Miyake, P. Zhang, M. L. Cohen, and S. G. Louie, Quasiparticle energy of semicore *d* electrons in ZnS: Combined LDA + *U* and *GW* approach, *Phys. Rev. B* **74**, 245213 (2006).
- [42] S. Lany and A. Zunger, Assessment of correction methods for the band-gap problem and for finite-size effects in supercell

- defect calculations: Case studies for ZnO and GaAs, *Phys. Rev. B* **78**, 235104 (2008).
- [43] O. Zakharov, A. Rubio, X. Blase, M. L. Cohen, and S. G. Louie, Quasiparticle band structures of six II-VI compounds: ZnS, ZnSe, ZnTe, CdS, CdSe, and CdTe, *Phys. Rev. B* **50**, 10780 (1994).
- [44] L. Ley, R. Pollak, F. McFeely, S. P. Kowalczyk, and D. Shirley, Total valence-band densities of states of III-V and II-VI compounds from x-ray photoemission spectroscopy, *Phys. Rev. B* **9**, 600 (1974).
- [45] S.-H. Wei and A. Zunger, Predicted band-gap pressure coefficients of all diamond and zinc-blende semiconductors: Chemical trends, *Phys. Rev. B* **60**, 5404 (1999).
- [46] T. N. Morgan, Symmetry of Electron States in GaP, *Phys. Rev. Lett.* **21**, 819 (1968).