Design and exploration of semiconductors from first principles: A review of recent advances

To cite this article: Fumiyasu Oba and Yu Kumagai 2018 Appl. Phys. Express 11 060101

View the article online for updates and enhancements.

Related content
- Point defects in ZnO: an approach from first principles
  Fumiyasu Oba, Minseok Choi, Atsushi Togo et al.
- Native defects in oxide semiconductors: a density functional approach
  Fumiyasu Oba, Minseok Choi, Atsushi Togo et al.
- Computational methods for 2D materials: discovery, property characterization, and application design
  J T Paul, A K Singh, Z Dong et al.

Recent citations
- First-principles many-body study of the electronic and optical properties of CsK$_2$Sb, a semiconducting material for ultra-bright electron sources
  Caterina Cocchi et al
- Improved adaptive sampling method utilizing Gaussian process regression for prediction of spectral peak structures
  Yuki K. Wakabayashi et al
Design and exploration of semiconductors from first principles: A review of recent advances

Fumiyasu Oba1,2,3 and Yu Kumagai2,4

1Laboratory for Materials and Structures, Institute of Innovative Research, Tokyo Institute of Technology, Yokohama 226-8503, Japan
2Materials Research Center for Element Strategy, Tokyo Institute of Technology, Yokohama 226-8503, Japan
3Center for Materials Research by Information Integration, Research and Services Division of Materials Data and Integrated System, National Institute for Materials Science, Tsukuba, Ibaraki 305-0047, Japan
4PRESTO, Japan Science and Technology Agency, Kawaguchi, Saitama 332-0012, Japan

Received January 14, 2018; accepted March 14, 2018; published online May 29, 2018

Recent first-principles approaches to semiconductors are reviewed, with an emphasis on theoretical insight into emerging materials and in silico exploration of as-yet-unreported materials. As relevant theory and methodologies have developed, along with computer performance, it is now feasible to predict a variety of material properties ab initio at the practical level of accuracy required for detailed understanding and elaborate design of semiconductors; these material properties include (i) fundamental bulk properties such as band gaps, effective masses, dielectric constants, and optical absorption coefficients; (ii) the properties of point defects, including native defects, residual impurities, and dopants, such as donor, acceptor, and deep-trap levels, and formation energies, which determine the carrier type and density; and (iii) absolute and relative band positions, including ionization potentials and electron affinities at semiconductor surfaces, band offsets at heterointerfaces between dissimilar semiconductors, and Schottky barrier heights at metal–semiconductor interfaces, which are often discussed systematically using band alignment or lineup diagrams. These predictions from first principles have made it possible to elucidate the characteristics of semiconductors used in industry, including group III–V compounds such as GaN, GaP, and GaAs and their alloys with related Al and In compounds; amorphous oxides, represented by In–Ga–Zn–O; transparent conductive oxides (TCOs), represented by In2O3, SnO2, and ZnO; and photovoltaic absorber and buffer layers, materials such as CdTe and CdS among group II–VI compounds and chalcopyrite CuInSe2, CuGaSe2, and CuIn0.6Ga0.4Se2 (CIGS) alloys, in addition to the prototypical elemental semiconductors Si and Ge. Semiconductors attracting renewed or emerging interest have also been investigated, for instance, divalent tin compounds, including SnO and SnS; wurtzite-derived ternary compounds such as ZnSnN2 and CuGaO2; perovskite oxides such as SrTiO3 and BaSnO3; and organic–inorganic hybrid perovskites, represented by CH3NH3PbI3. Moreover, the deployment of first-principles calculations allows us to predict the crystal structure, stability, and properties of as-yet-unreported materials. Promising materials have been explored via high-throughput screening within either publicly available computational databases or unexplored composition and structure space. Reported examples include the identification of nitride semiconductors, TCOs, solar cell photoabsorber materials, and photocatalysts, some of which have been experimentally verified. Machine learning in combination with first-principles calculations has emerged recently as a technique to accelerate and enhance in silico screening. A blend of computation and experimentation with data science toward the development of materials is often referred to as materials informatics and is currently attracting growing interest. © 2018 The Japan Society of Applied Physics

1. Introduction

Semiconductors play increasingly important roles in modern society. In addition to prototypical group IV elemental semiconductors, III–V and II–VI compounds are currently used in a variety of electronic and optoelectronic devices, photovoltaic cells, and so forth. The functionalities of oxide semiconductors have been explored extensively in recent decades, and industrial applications such as thin-film transistors and transparent electrodes have been realized. Semiconductors doped with magnetic impurities, which are referred to as dilute magnetic semiconductors, have been developed toward spintronic applications. These technologies typically rely on heterostructures fabricated under precisely controlled conditions, where understanding and tailoring of lattice defects, including point defects (native defects, residual impurities, and dopants), dislocations, surfaces, and interfaces, are essential, as well as knowledge of fundamental bulk properties, because such defects affect or even dominate device performance. A fair amount of experimental research on lattice defects has therefore been conducted, in addition to characterization of bulk properties. However, lattice defects are often difficult to fully access by experiment, especially at the atomic and electronic levels, and computational approaches complement experimental investigation. Historically, lattice statics and molecular dynamics simulations using empirical interatomic potentials have been widely applied to defect modeling. Because it typically requires large simulation cells, computationally demanding approaches were infeasible in the past. The recent development of electron theory and relevant computational methods, as well as computer performance, now allows for the use of first-principles, or ab initio, calculations based on quantum mechanics for defect-related problems. Such first-principles studies have provided useful insights toward detailed understanding and elaborate tailoring of lattice defects as well as bulk properties.

Materials with diverse properties are needed to further expand the applications of semiconductors, but the number of industrially used semiconducting materials is thus far very limited. With the aid of advanced computational methodologies, in particular those for high-throughput screening using first-principles calculations, previously unreported but promising materials can now be efficiently explored. These in silico screening approaches have been developing rapidly, so not only the stability and bulk properties of candidate materials but also more complex properties such as those of lattice defects are increasingly considered toward more realistic searches for materials. Furthermore, data-science-based schemes can be combined with first-principles calculations to accelerate or enhance screening. Some of the materials identified by in silico
screening have been verified experimentally, demonstrating its predictive power.

In this article, we review recent progress in first-principles studies of semiconductors. Our discussion covers not only traditional computational approaches to understanding the behavior of currently used and emerging semiconductors but also the design and exploration of as-yet-unreported materials. The remainder of this article is organized as follows: In Sect. 2, we briefly describe the theoretical and computational background of first-principles calculations. We then discuss the prediction of fundamental bulk properties such as band gaps, effective masses, dielectric constants, and optical absorption coefficients in Sect. 3. Also mentioned are point defect properties, including formation energies, local atomicistic structures, and electronic states. Among the properties of semiconductor surfaces and interfaces, we focus on the absolute band positions at surfaces and band offsets at heterointerfaces, which constitute a basis for the design of heterostructure devices. In Sect. 4, we discuss the computational design and exploration of as-yet-unreported semiconductors, as well as experimental verification of theoretical predictions. A combination of first-principles calculations and data science techniques, which is currently attracting growing interest, is also mentioned. Finally, concluding remarks are given in Sect. 5.

2. Theoretical background of first-principles calculations

In this section, we provide a brief overview of basic theory and approximations in first-principles calculations, with an emphasis on those frequently applied to semiconductors. Whereas the solution of Schrödinger equations is originally based on wavefunction theory, density functional theory (DFT) is widely employed in recent calculations. In DFT, the fundamental variable is the electron density $\rho(r)$ of all electrons involved in the system, instead of many-body wavefunctions. Although the existence of the universal total-energy functional $E[\rho]$ has been proved, the exact functional form is unknown.

To determine the ground-state electron density and the corresponding total energy, the Kohn–Sham scheme is typically adopted, where the exchange–correlation term expresses the electron–electron interactions other than the Hartree term. The exchange–correlation term is often treated using a density functional that is constructed within the local (spin) density approximation (LSDA) or generalized gradient approximation (GGA), which are based on solutions for an electron gas. The GGA functionals in particular include various forms, such as Becke 88 (B88), Perdew–Burke–Ernzerhof (PBE), and Armiento–Mattsson 05 (AM05). When van der Waals interactions play critical roles in the prediction of the geometry, for instance, in layered structures, additional correlation terms are applied to model the van der Waals interactions.

Typically, the LDA and GGA or these approximations with van der Waals corrections reproduce the lattice parameters of crystals within errors of a few percent. However, the band gaps of semiconductors and insulators are severely underestimated, and spatially localized states are not well described. To remedy the latter problem, the LDA and GGA plus the Hubbard $U$ corrections, that is, LDA+$U$ and GGA+$U$, have been widely employed. Although this approach was originally developed to describe strongly correlated systems, it also improves the band gaps of band insulators when localized states are involved in their valence or conduction bands, as discussed later in Sect. 3.1.1.

Meta-GGAs such as Tao–Perdew–Staroverov–Scuseria (TPSS) and the strongly constrained and appropriately normed (SCAN) functionals have been reported to outperform the LDA and GGA in terms of band-gap prediction, but underestimation is still obvious in many cases. The modified Becke-Johnson (mBJ) exchange potential, in combination with the L(S)DA correlation, is often used in current band structure calculations. By appropriately choosing the parameters, the band gaps of various materials can be well reproduced, although the valence and conduction band widths tend to be underestimated.

The Gritsenko–Leeuwen–Lenthe–Baerends potential including the correlation for solids (GLLB-sc), in which an exchange–correlation derivative discontinuity term is added to the Kohn–Sham gaps, shows significant improvement in band-gap prediction over the LDA and GGA. Efficient and improved prediction of band gaps has also been reported using the $\Delta$-sol method, which is a generalization of the Delta self-consistent-field method to infinite solids.

Hybrid functionals, in which unscreened or screened nonlocal Fock exchange is partially incorporated, have been shown to describe the band structures of diverse systems as well as screened exchange methods. Both approaches include wavefunctions in the exchange term but can be treated similarly to local and semilocal functionals in the framework of the generalized Kohn–Sham scheme. Various forms of hybrid functionals have been proposed to date. Among them, PBE0 and Heyd–Scuseria–Ernzerhof (HSE06) are often adopted in calculations of solids, including semiconductors. The Becke three-parameter Lee–Yang–Parr functional (B3LYP) was originally used for molecules but has also been applied to semiconductors. An issue with hybrid functionals is that the optimal amount of nonlocal exchange is system-dependent, reflecting the strength of the electronic screening, and therefore it must be tuned to reproduce band gaps, especially those of wide-gap materials. This is empirically possible so that experimental band structures are well reproduced, although recently proposed dielectric-dependent hybrid functionals can provide well-balanced descriptions of a wide variety of systems without empirical tuning. Another issue is that the computational costs are significantly greater than those of LDA and GGA calculations when plane-wave basis functions are used in hybrid functional calculations.

The GW approximation based on many-body perturbation theory allows for even more accurate prediction of the band gaps of many materials, although the calculations are rather demanding. Several levels of approximation have been proposed in terms of the treatment of the screened Coulomb interaction, self-consistency, and vertex corrections in the screened Coulomb interaction and self-energy. In particular, self-energy vertex corrections have been reported to be important for accurate reproduction of the absolute band positions and localized states, as discussed later. Although the GW approximation provides quasiparticle energies for dressed electrons and holes on the basis of many-
body perturbation theory, the total energies can also be obtained in a related manner using the adiabatic-connection fluctuation–dissipation theorem.\textsuperscript{73} A large number of program codes for first-principles calculations have been developed, and many of them cover all or some of these approximations. In particular, fundamental approximations to DFT or hybrid DFT, such as the LDA, GGA, and hybrid functionals, are available in many codes. It has been shown recently that 15 widely used codes yield reproducible results when the same approximation to the exchange–correlation functional in DFT is used, irrespective of differences in the approximations other than the exchange–correlation functional, basis functions, and implementation details.\textsuperscript{20}

3. Prediction of bulk, alloy, and defect properties of semiconductors

3.1 Fundamental bulk properties

3.1.1 Band structure and band gaps. Needless to say, the most fundamental property of semiconductors is the band gap. Except in strongly correlated systems, band gaps are derived from one-electron (single-particle) or quasiparticle band structures. As mentioned in Sect. 2, however, accurate prediction of the band gaps of diverse materials is challenging, as it requires a rather high level of approximation. In addition, spin–orbit coupling reduces band gaps, particularly when heavy elements are the main contributors to electronic states near the gaps. For instance, the reported gap reduction amounts to <0.05 eV for C, Si, GaN, ZnO, GaP, ZnS, and CdS; ~0.1 eV for Ge, GaAs, ZnSe, and CdSe; and ~0.3 eV for GaSb, ZnTe, and CdTe; these reductions are attributed mainly to upward shifts of the valence bands.\textsuperscript{76} The shifts are larger when the main components of the valence bands, which are typically anions in the case of compounds, are heavier. Spin–orbit coupling is safely neglected for some materials composed of light elements but not for others. Significant spin–orbit coupling effects of ~0.3 and ~0.7 eV have also been found for PbS and PbTe, respectively, where the electronic states around both the valence band maximum (VBM) and conduction band minimum (CBM) are strongly affected owing to the contributions of Pb(II), as well as Te, in PbTe.\textsuperscript{78} This is again in contrast to the case of SnS, with lighter Sn(II) and S constituents, which shows a reduction of only 0.02 eV.\textsuperscript{79} For the organic–inorganic hybrid perovskite CH\textsubscript{3}NH\textsubscript{3}PbI\textsubscript{3}, which has recently attracted tremendous interest as a solar cell photoabsorber material, a gap reduction of as large as ~1 eV has been reported.\textsuperscript{80–84}

The importance of electron–phonon coupling has also been pointed out, even in predictions of band gaps at the ground state. The reported band gap reduction due to such zero-point renormalization amounts to 0.1 and 0.15 eV in Si\textsuperscript{85} and ZnO,\textsuperscript{86} respectively, and it is especially large in diamond, 0.6 eV.\textsuperscript{87} We need to keep this fact in mind, but theoretical band gaps excluding electron–phonon coupling effects are typically considered in practical first-principles studies of semiconducting materials.

The determination of band gaps requires knowledge of the wave-vector dependence of the band energies. This information is obtained by drawing band structure diagrams; note, however, that the band extrema, that is, band edges, are not necessarily located at high-symmetry points sampled in band structure diagrams, and an analysis of the entire first Brillouin zone may be required in some cases.\textsuperscript{79} Systematic construction of band structure diagrams for materials with diverse crystal structures is actually a hard task, as the band path is crystal-structure-dependent. Fortunately, standard primitive cells and band paths have been reported for all types of crystal structures by Setyawan and Curtarolo\textsuperscript{86} and by Hinuma et al.,\textsuperscript{89} and are available online at the AFLOW Distributed Materials Property Repository (www.aflowlib.org) and the SeeK-path website (www.materialscloud.org/tools/seekpath), respectively; in particular, the latter considers band paths that are compatible with crystallographic convention.

Figure 1 shows the band gaps of prototypical semiconductors obtained using various approximations with spin–orbit coupling included.\textsuperscript{76,80} Overall, the tendency mentioned in Sect. 2 is recognized: severe underestimation of band gaps by the GGA and improvement by using the hybrid functional and the GW approximation. The GW\textsuperscript{T} approximation, which includes vertex corrections in the self-energy, yields slightly overestimated band gaps. This is a theoretically reasonable tendency, given that the band gaps are reduced from these values if the electron–phonon coupling effects are also considered.

The electronic band structures of In compound semiconductors are shown in Fig. 2.\textsuperscript{91} Spin–orbit coupling has been taken into account to appropriately describe the valence band structures composed of heavy-hole, light-hole, and split-off bands. On the other hand, many other studies, in which an outline of the band structure is of interest, do not include spin–orbit coupling. For example, band gaps can be reproduced well without considering spin–orbit coupling when the target materials do not consist of heavy elements, as mentioned above.

Among the systems included in Fig. 1, the band structure prediction of Zn compounds is challenging owing to the presence of localized Zn 3d states slightly below anion p states. Such localized states are underbound, in other words, too high in energy, under the LDA and GGA, and therefore are over-hybridized with anion states in the valence band. This is an especially serious problem in ZnO, as discussed in previous studies.\textsuperscript{63,76,92–97} Figure 3 shows the electronic density of states (DOS) of ZnO obtained using the PBE-GGA and HSE\textsubscript{(α = 0.375)} hybrid functionals, the latter of which uses a tuned nonlocal exchange amount of 0.375 to reproduce an experimental band gap.\textsuperscript{83} Electronic states forming sharp peaks are recognized around ~5 and ~6.5 eV with respect to the VBM in the PBE-GGA and HSE\textsubscript{(α = 0.375)} results, respectively. These Zn 3d band positions are too shallow compared to an experimentally reported position of ~7.5 eV,\textsuperscript{98} which results in overhybridization with O 2p states located above them and thereby affects the band gap. The results are enumerated in Table I alongside values from other approximations. PBE-GGA+U can control the 3d position by tuning the effective U value (U\textsubscript{eff}), and the experimental 3d position is almost reproduced with U\textsubscript{eff} = 7.5 e V. This approach improves the band gap over that of the standard PBE-GGA without +U corrections, but the underestimation is still severe. The GW\textsubscript{0}/PBE approximation predicts the band gap accurately, although the 3d position is too high owing to the residual self-interaction error.\textsuperscript{76} This can be
improved by vertex corrections in the self-energy, as in the \( GW_\Gamma @HSE06 \) result, but the calculation becomes much more demanding. The biorthogonal transcorrelated method also yields a reasonably good description of the overall electronic structure of ZnO, as recently reported by Ochi et al.\(^99\)

Another characteristic band structure is found in SnO, where the 5s and 5p orbitals of Sn(II) make significant contributions to electronic states near the VBM. Prominent hybridization between the Sn 5sp and O 2p states is recognized in the electronic DOS shown in Fig. 4. Such cationic orbital contributions are favorable for p-type doping as they tend to raise the VBM toward the vacuum level and expand the width of the valence bands,\(^103\) as discussed later in Sect. 4.1.

Organometal halide perovskites with Pb(II) have analogous band structures. Figure 5 compares the band structures of CH\(_3\)NH\(_3\)PbI\(_3\) and NH\(_4\)PbI\(_3\) obtained from quasiparticle self-consistent GW and LDA calculations, where the former approach has been confirmed to reproduce well an experimental band gap of CH\(_3\)NH\(_3\)PbI\(_3\).\(^81\) The differences between the GW and LDA results are evident in not only the band gap values but also the band dispersions and relative band positions, particularly those of the highest valence bands. The sizable differences, that is, the large band-dependent quasiparticle shifts, are attributed in part to the complicated band structures with the aforementioned Pb(II) orbital contributions; for simpler band structures, the band topologies are typically less affected by quasiparticle shifts, aside from band gap changes. It should again be noted that spin–orbit coupling reduces the band gap of CH\(_3\)NH\(_3\)PbI\(_3\) by as much as \( \sim 1 \) eV.\(^80\–84\) This value is comparable to the quasiparticle shift of 1.1 eV.\(^81\) As a consequence of error cancellation, LDA/GGA calculations neglecting spin–orbit coupling accidentally yield band gaps close to the experimental value.

Because many-body approaches such as the \( GW \) approximation are computationally rather demanding, a more cost-effective but sufficiently accurate approach is required in studies that consider a large number of systems. The use of...
hybrid functionals would be a good compromise, as would the less expensive approaches such as mBJ, GLLB-sc, and Δ-sol mentioned in Sect. 2. For a typical unit cell size, hybrid functional calculations are one or two orders of magnitude more time-consuming than those using the LDA and GGA when a plane-wave basis set is used. A recently proposed non-self-consistent approach is useful for accelerating calculations while retaining sufficient accuracy, at least for semiconductors and insulators with relatively simple band structures.68,106) Another issue with hybrid functionals is that the optimal amount of nonlocal exchange is system-dependent, as discussed in Sect. 2, but this can be resolved nonempirically by the aforementioned dielectric-dependent hybrid functionals; the results for a material set similar to that shown in Fig. 1 have been reported in Ref.68, and applications to other systems can be found in Refs.61, 64–67, and 69, all of which demonstrate good performance of dielectric-dependent hybrid functionals in band-gap prediction for wide-gap systems.

### Table I. Band gap ($\varepsilon_g$) and average position of the Zn 3d states with respect to the VBM ($\varepsilon_{3d}$) of ZnO obtained using various approximations76,90,96,99) compared to experimental values.98,100)

| Approximation         | $\varepsilon_g$ (eV) | $\varepsilon_{3d}$ (eV) | Reference          |
|-----------------------|-----------------------|-------------------------|--------------------|
| PBE-GGA               | 0.80                  | −4.9                    | 76, 90, 96         |
| TPSS-meta-GGA        | 1.45                  | −5.1                    | 101                |
| PBE-GGA+U ($U_{eff} = 5.0$ eV) | 1.42                  | −6.3                    | 96                 |
| PBE-GGA+U ($U_{eff} = 7.5$ eV) | 1.81                  | −7.2                    | 96                 |
| HSE06-hybrid         | 2.55                  | −6.2                    | 76                 |
| HSE($\alpha = 0.375$)-hybrid | 3.42                  | −6.4                    | 63                 |
| Screened exchange    | 3.41                  | −7.0                    | 102                |
| GW$_{\Delta}$PBE     | 3.40                  | −6.7                    | 76, 90             |
| GW$_{TC}$-TC@HSE06   | 3.68                  | −6.9                    | 76, 90             |
| GW$_{TC}$-TC@HSE06   | 3.88                  | −7.1                    | 76, 90             |
| Biorthogonal transcorrelated | 3.1                   | −9.1                    | 99                 |
| Experiment           | 3.44                  | −7.5                    | 98, 100            |

### 3.1.2 Effective masses.

The electron and hole effective masses are relevant to most applications of semiconductors, as they are decisive factors affecting the carrier mobilities. Effective masses can be calculated with acceptable accuracy using the LDA and GGA, especially for materials with relatively simple band structures.107) Although improved accuracy has reportedly been obtained by using meta-GGA or hybrid functionals, for example, for InP and GaAs.108) This is because each of the valence and conduction band structures is reasonably well reproduced even when the band gaps are severely underestimated by the LDA and GGA. Serious errors could occur for systems with localized states around their valence or conduction bands, for example, in the Zn and Cu(I) compounds mentioned above and below. In these

![Fig. 3. Electronic DOS for ZnO obtained using the (a) PBE-GGA and (b) HSE($\alpha = 0.375$) hybrid functionals. Zero energy is set to the VBM. Adapted from Ref. 96. © 2010 IOP Publishing.](image)

![Fig. 4. Electronic band structure and projected electronic DOS for SnO obtained using the PW91-GGA functional. Zero energy is set to the VBM. The indirect minimum band gap is 0.3 eV smaller than an experimental value of 0.7 eV owing to the use of the GGA. Adapted from Ref. 105. © 2006 American Physical Society.](image)

![Fig. 5. Electronic band structures of the organometal halide perovskites (a) CH$_3$NH$_3$PbI$_3$ and (b) NH$_4$PbI$_3$ obtained using the quasiparticle self-consistent GW approximation including spin–orbit coupling, alongside schematic crystal structures. Zero energy is set to the VBM. Bands are colored according to their orbital nature: Pb 6s (blue), Pb 6p (red), and I 5p (green). The band structures calculated using the LDA are also shown as light gray dashed lines. Reprinted from Ref. 81. © 2014 American Physical Society.](image)
materials, the position of the localized 3d states affects the valence band structure and thus the effective masses. This problem is often remedied by the use of LDA+U/GGA+U, as well as hybrid functionals and the GW approximation.

A typical procedure for the evaluation of effective masses is the analysis of band curvatures along wave vectors of interest. This can be done on the basis of theoretical band structure diagrams through fine k-point sampling in the vicinity of the band extrema. We again note that, to model the detailed valence band structures, particularly the light-hole and split-off bands, spin–orbit coupling needs to be taken into account. Another approach includes the computation of so-called DOS effective masses using a theoretical electronic DOS.109) The derivation of average effective mass tensors, which takes into account the effects of nonparabolicity, multiple bands, multiple minima, and anisotropy, has also been reported.110,111) The calculated effective masses for a variety of semiconductors have recently been compiled by Ricci et al.,107) along with other electronic transport properties.

3.1.3 Dielectric constants. Dielectric constants including both the ionic and electronic contributions are typically calculated using the LDA or GGA on the basis of density functional perturbation theory.112) For evaluation of the electronic contributions, a finite electric field approach113) has also been applied; it can be used in combination with hybrid functionals, as well as the LDA and GGA. The electronic contributions to the dielectric constants are especially sensitive to the quality of the predicted band structures. Table II summarizes the static electronic dielectric constants obtained using the PBE-GGA (with and without +U corrections for the Zn 3d states in the Zn compounds) and PBE0 hybrid functionals. The PBE0 hybrid functional has been reported to well reproduce or slightly overestimate the band gaps of the listed materials. The resultant dielectric constants are close to the experimental and PBE0 values.

Table II. Static electronic dielectric constants obtained using the PBE-GGA functional with the random phase approximation ($\varepsilon_{\text{PBE-RPA}}$) and local-field effects ($\varepsilon_{\text{PBE}}$), and the PBE0 hybrid functional with local-field effects ($\varepsilon_{\text{PBE0}}$)116) compared to experimental values.114) The PBE-GGA+U results with $U_{\text{eff}}=5.0$ eV for the Zn 3d states are also shown in parentheses. Mean absolute errors (MAEs) with respect to the experimental values are listed on the bottom line; the values in parentheses are obtained using the PBE-GGA results with and without +U corrections for the Zn compounds and the others, respectively.

| Material     | $\varepsilon_{\text{PBE-RPA}}$ | $\varepsilon_{\text{PBE}}$ | $\varepsilon_{\text{PBE0}}$ | Experiment |
|--------------|-------------------------------|-----------------------------|-----------------------------|------------|
| C            | 5.5                           | 5.9                         | 5.4                         | 5.7        |
| Si           | 12.1                          | 12.9                        | 10.9                        | 11.6       |
| BN           | 4.4                           | 4.6                         | 4.4                         | 4.46       |
| AlP          | 7.5                           | 8.1                         | 7.1                         | 7.4        |
| AlAs         | 8.6                           | 9.3                         | 8.0                         | 8.16       |
| AlSb         | 10.7                          | 11.5                        | 9.8                         | 9.88       |
| GaN          | 5.5                           | 5.8                         | 5.0                         | 5.4        |
| GaP          | 9.7                           | 10.4                        | 8.7                         | 8.8        |
| GaAs         | 12.9                          | 13.7                        | 10.1                        | 10.86      |
| InP          | 10.3                          | 11.0                        | 8.7                         | 9.9        |
| MgO          | 3.0                           | 3.1                         | 2.9                         | 3.1        |
| ZnO          | 4.8 (4.1)                     | 5.0 (4.3)                   | 3.6                         | 3.7        |
| ZnS          | 5.6 (5.3)                     | 5.9 (5.6)                   | 5.1                         | 5.1        |
| ZnSe         | 6.7 (6.3)                     | 7.2 (6.8)                   | 5.9                         | 5.9        |
| ZnTe         | 8.3 (7.8)                     | 8.9 (8.4)                   | 7.2                         | 6.9        |
| CdS          | 5.7                           | 6.1                         | 5.1                         | 5.4        |
| CdSe         | 7.2                           | 7.6                         | 5.9                         | 6.2        |
| CdTe         | 8.1                           | 8.6                         | 6.9                         | 7.1        |
| MAE          | 0.7 (0.5)                     | 1.1 (1.0)                   | 0.3                         |            |

Technically speaking, theoretical dielectric constants are fortuitously improved by the use of a more approximate approach, that is, the random phase approximation. The +U corrections to the Zn 3d states improve the description of the band structures of Zn compounds, as noted above for ZnO. The resultant dielectric constants are closer to the experimental and PBE0 values.

3.1.4 Optical properties. Optical absorption and emission are especially important for photovoltaic cell and optoelectronic device applications of semiconductors. Optical absorption spectra can be obtained via calculations of the electronic part of the complex dielectric functions, ideally considering excitonic or electron–hole coupling effects115) and phonon-assisted electronic transitions, especially for indirect-type semiconductors.116,117) The optical absorption and emission properties have been discussed in many first-principles studies, a notable example of which is the interpretation of the absorption spectrum of In$_2$O$_3$ reported by Walsh et al.118)

Undoped In$_2$O$_3$ shows an optical absorption onset at about 3.7 eV, which is accompanied by weak absorption with a threshold of about 2.6 eV. This weak absorption degrades the optical transparency slightly and has been attributed to either phonon-assisted absorption via an indirect gap or defect-induced absorption, as mentioned in Ref. 118. According to the first-principles calculations by Walsh et al.118) and by Fuchs and Bechstedt,119) the highest valence band has a very small wave-vector dependence in In$_2$O$_3$. Therefore, the 1.1 eV difference between the lower weak and higher strong absorption onsets cannot be attributed to the indirect and direct electronic transitions. Instead, Walsh et al. explained this behavior in terms of symmetry- or parity-forbidden electronic transitions from upper valence states to lower conduction states, leading to the strong absorption onset approximately 0.8 eV above the fundamental band gap value118) (Fig. 6). This is consistent with their X-ray photoelectron spectroscopy results, indicating that the fundamental gap of In$_2$O$_3$ is less than 2.9 eV.

Sn-doped In$_2$O$_3$ allows for simultaneous realization of good optical transparency and high electron conductivity. These properties are exploited in transparent electrodes. In such heavily doped systems, carrier electrons fill the lower part of the conduction band, and the Fermi level is located within the conduction band. This results in a blue shift of the absorption onset, called the Burstein–Moss effect, because only the unoccupied states above the Fermi level contribute to the electronic transition from occupied valence states. Theoretical investigation of the Burstein–Moss effects on the optical absorption spectra and optical band gaps has been reported for various semiconductors, including n-type transparent conductive oxides (TCOs) such as In$_2$O$_3$, ZnO, SnO$_2$, and BaSnO$_3$,120) p-type oxides such as SnO$^{120}$ and CuAlO$_2$,120) and heavily doped nitrides such as GaN,121) InN,121) Zn$_3$N$_2$,122) ScN,123,124) and ZnSnN$_2$.125)
Another example of the use of theoretical optical absorption spectra includes performance evaluation of solar cell photoabsorber materials. Yu and Zunger calculated the photovoltaic energy conversion efficiencies as a function of absorber thickness by extending the Shockley–Queisser model126 and using absorption coefficients obtained from first-principles calculations.127 The effects of nonradiative recombination are approximated in their model. Yu, Zunger, and their coworkers applied this approach to various Cu ternary and Ag ternary chalcogenides,127,128 examples of which are shown in Fig. 7, and ABX half-Heusler compounds;129 their results are also mentioned in Sect. 4.3 on in silico high-throughput screening.

The same or analogous approaches have been taken by other researchers to predict the conversion efficiencies of emerging or hypothetical photoabsorber materials, for example, ZnSnP2 and its alloys, compared to industrially used GaAs, CdTe, CuInS2, and CuGaSe2 by Yokoyama et al.,130 Cu ternary and Ag ternary chalcogenides by Bercx et al.,131 and Ag–Cu sulfides and Pb-free halide double perovskites by Savory et al.132,133 Blank et al. have evaluated the theoretical efficiencies of relevant sulfides and selenides including CuInS2, CuGaSe2, Cu2ZnSnS3 (CZTS), Cu2ZnSnSe4 (CZTSe), CuSbS2, Sb2S3, CuSbSe2, and Sb2Se3 using an extended scheme, where the contribution of the refractive index was also taken into account.134

3.2 Alloy properties

Alloying semiconductors is an essential technique for band gap engineering. To model alloys of semiconductors as well as metals, the cluster expansion method135 and Monte Carlo simulations are often employed in conjunction with first-principles calculations, allowing us to construct phase diagrams and predict alloy properties such as composition-dependent band gaps. As an example, the band gaps of CuIn1−xGaSe2 (CIGS) and CuIn1−xAlxSe2 (CIAS) pseudobinary alloys obtained using such a combined approach are shown in Fig. 8.138 Ideally random alloys are assumed here, which have been shown to yield band gaps close to those more rigorously predicted via Monte Carlo simulations in the temperature range of 500–1000 K.138

The band gaps of CIGS and CIAS show not a linear but a nearly quadratic composition dependence. This behavior is typical of semiconductor alloys, and their band gaps are often discussed using the band bowing parameter in quadratic fitting of the composition dependence of band gaps as

$$\varepsilon_g(x) = (1-x)\varepsilon_g^A + x\varepsilon_g^B - bx(1-x)$$

where $\varepsilon_g^A$ and $\varepsilon_g^B$ are the band gaps of components A and B in
an A–B binary or pseudobinary alloy, respectively, and \( b \) denotes the bowing parameter.

Theoretical bowing parameters of 0.16 and 0.57 eV for CIGS and CIAS, respectively, are obtained by fitting the calculated band gaps in the entire composition range.\(^{138}\) It is clear, however, that the composition dependence of CIAS is not quadratic, as it shows stronger bowing at higher CAS contents. Similar behavior has been reported for InGaN alloys.\(^{139}\)

Because explicit modeling of alloys is computationally demanding, special quasi-random structures (SQSs)\(^{140}\) are often adopted to mimic random alloys concisely in terms of atomic correlation functions. For CIGS, a bowing parameter of 0.13 eV is obtained using SQSs,\(^ {130}\) which is in reasonable agreement with the value from cluster expansion. SQSs can also be used to discuss the properties of compounds with disordered atomic arrangements. Note, however, that random atomic arrangements based on SQSs could be inappropriate for systems that have strong tendencies to preserve specific local environments; examples include ZnSnP\(_2\) and ZnSnN\(_2\), where disordering of heterovalent Zn and Sn ions is likely to occur under the octet rule or local charge neutrality conservation at typical growth temperatures, as reported by Ma et al.\(^ {141}\) and by Lany et al.\(^ {142}\) respectively.

Alloys that consist of isostructural semiconductors having analogous band structures are often considered for band gap engineering, for example, industrially used III–V and II–VI compound semiconductor alloys. Recently, Holder et al. proposed that heterostructural semiconductor alloys can have a wider range of metastable compositions than isostructural alloys, the metastability of which is severely limited by binodal and spinodal decomposition, as shown in Fig. 9.\(^ {143}\) Such heterostructural alloys are also of interest in view of their properties; we will return to this issue in Sect. 4.1.

3.3 Band alignment: Ionization potentials, electron affinities, and heterojunction band offsets

Band alignment carries fundamental information for the design of semiconductor heterojunctions.\(^ {144-146}\) The ionization potential (IP) and electron affinity (EA), which are the VBM and CBM measured from the vacuum level, respectively, are key parameters often used to discuss band alignment; note that the sign is taken so that the IP and EA are positive when the VBM and CBM are lower than the vacuum level, respectively. The IP and EA involve surface dipole contributions.\(^ {147-149}\) Therefore, we should keep in mind that IP- and EA-based band alignment depends on the surface orientation, composition, local atomicistic and electronic structure, and adsorption or contamination, which affect the surface dipole. Nevertheless, IPs and EAs from some representative surfaces can be used to predict heterojunction band offsets with reasonable accuracy, especially when the interface constituents have similar crystal and electronic structures.\(^ {90,150,151}\) This corresponds to the so-called EA model for the prediction of interfacial conduction band offsets, where charge transfer effects are not taken into account as in the case of the Schottky limit for metal–semiconductor interfaces.\(^ {146}\) The EA model has been used directly for the estimation of band offsets at semiconductor heterointerfaces\(^ {90,152}\) or modified by charge neutrality levels\(^ {153}\) and Schottky pinning parameters accounting for charge transfer effects.\(^ {146}\) The Schottky barrier heights at metal–semiconductor interfaces have been predicted in an analogous manner,\(^ {146,154}\) where the work functions of metals can also be evaluated through first-principles calculations of surfaces. Moreover, surface band alignment is useful for systematic discussion of the doping limits of semiconductors.\(^ {103}\) Other types of band alignment thus far considered include those using branch point energies or charge neutrality levels,\(^ {90,146,153}\) impurity levels such as that of hydrogen,\(^ {155}\) and averaged interfacial band offsets.\(^ {90,156,157}\) In this section, we first discuss surface band alignment and subsequently mention band offsets directly calculated using heterointerface models.

The IP (I) and EA (A) are typically obtained by combining the surface and bulk calculations as

\[
I = E_{\text{VBM}} - E_{\text{CBM}},
\]

\[
A = E_{\text{bulk}} - E_{\text{bulk}},
\]

where

\[
E_{\text{vac}} \quad \text{and} \quad E_{\text{ref}}
\]

are the surface and bulk values, respectively.
where $\epsilon_{\text{vac}}$ and $\epsilon_{\text{ref,far}}$ are the vacuum level and the electrostatic reference level in the bulk-like region far from the surface, respectively. These values can be obtained by a surface calculation using a slab model (supercell), where two-dimensional slab and vacuum regions are stacked in the cell and repeated under three-dimensional periodic boundary conditions. $\epsilon_{\text{VBM,bulk}}, \epsilon_{\text{CBM,bulk}},$ and $\epsilon_{\text{bulk}}$ are the VBM, CBM, and electrostatic reference level from a bulk calculation, respectively. This procedure is illustrated for IPs in Fig. 10. The macroscopically averaged electrostatic potential is used widely as the reference level,\(^71,76,90,158,160,161\) whereas the local potential at atomic sites is also employed.\(^76,90,159\)

In principle, the IPs and EAs can be estimated using only surface models, but the bulk-surface combined approach has the following advantages. First, the effects of in-gap surface states are excluded from the estimation of the VBM. The EAs are obtained in a similar manner, where the CBM from the bulk model is taken instead of the VBM.

The band alignment of various semiconductor surfaces obtained using various approximations.\(^76,90\) The VBM positions with respect to the vacuum level, that is, the negatives of IPs, from the PBE-GGA calculations are much higher than the experimental values. HSE06 improves the VBM positions, but they are still too high. $GW_0@PBE$ and $GW^\Gamma@HSE06$ yield band gaps rather close to the experiment, but the band positions are found to be systematically too low, not only for GaAs and ZnSe, as shown in Fig. 11, but also for other semiconductors.\(^76,90\) This is remedied by self-energy vertex corrections, as in the $GW^\Gamma@HSE06$ results.

The band alignment of various semiconductor surfaces obtained using $GW^\Gamma@HSE06$ is shown in Fig. 12.\(^76,90\) Overall, good agreement is found between theory and experiment, demonstrating the method’s predictive power. This approach is, however, computationally rather expensive, and the use of dielectric-dependent hybrid functionals for band alignment is a good compromise in terms of balancing accuracy and efficiency.\(^68\)}

**Fig. 10.** Schematic showing a computational procedure for the evaluation of IPs using a supercell approach. The vacuum level from a surface slab model and the VBM from a bulk model are aligned using the common electrostatic reference levels in the bulk-like region of the surface model and the bulk model. This approach can exclude the effects of in-gap surface states from the estimation of the VBM. The EAs are obtained in a similar manner, where the CBM from the bulk model is taken instead of the VBM.

**Fig. 11.** Band alignment at (a) GaAs and (b) ZnSe (110) surfaces. The VBM and CBM positions with respect to the vacuum level (the negatives of the IP and EA, respectively) obtained using various approximations\(^76,90\) are shown, along with experimental values.\(^63\)


Generally speaking, modeling of surfaces is a complicated issue because a vast number of possible configurations exist in terms of the surface orientation and atomistic structures. Regarding the former point, automatic generation of nonpolar surface models has been reported recently in Refs. 164 and 165; for polar surfaces, modeling must be based on the mechanism that cancels out the internal electric field associated with the polarity, including accumulation of carrier electrons or holes, formation of surface point defects, adsorption, and so on. Moreover, general searches for surface atomistic structures without prior experimental knowledge require global structure optimization techniques.166)

When actual heterointerfaces are formed, specific structural reconstruction at the atomistic and electronic levels takes place, leaving interfacial dipoles. The relative band position of two materials that constitute an interface, that is, the band offset, is affected by such interfacial dipole contributions.144–146,153 Therefore, an explicit model of the heterointerface of interest would be desired. However, it is not straightforward except when two materials that have analogous crystal structures and lattice parameters form coherent interfaces; modeling of semicoherent interfaces, where misfit dislocations exist to relieve strain associated with lattice misfits,167,168) is limited by the affordable number of atoms in the simulation cells, and consideration of incoherent interfaces additionally requires statistical treatment, as these interfaces do not have periodicity. Furthermore, actual interfaces may show other types of atomic reconstructions169 and/or interdiffusion.170 Here we discuss the relatively clear and well-established cases of zinc-blende coherent heterointerfaces with small lattice misfits.

The valence band offsets at selected heterointerfaces calculated using various approximations are listed in Table III. The values are those of the so-called natural band offsets, where two interface constituent materials are assumed to be unstrained; the computational procedures adopted for natural band offset evaluation are described in Refs. 90 and 150. In addition, misfit dislocations are not considered in the models; their effects on macroscopic band offsets have been reported to be typically as small as ~0.1 eV or less for CdTe/CdS, CdS/ZnS, and InP/GaP (110) interfaces with relatively large lattice misfits and therefore high dislocation densities.171) Compared to that of the IPs and EAs, the dependence on the approximation is small. The experimental values are taken from Refs. 173–180 and include not only the values for the (110) interfaces considered in the calculations but also those for the (100) interfaces.

Table III. Natural valence band offsets at zinc-blende (110) heterointerfaces obtained using various approximations, compared to experimental values (in eV).190 Positive values mean that the VBM of semiconductor A is higher than that of semiconductor B at the A/B heterointerface. Interfaces with small lattice misfits between the constituent crystals (less than 0.5%) are selected, as the ambiguity associated with possible strain effects in experimental offsets is expected to be small. The experimental values are taken from Refs. 173–180 and include not only the values for the (110) interfaces considered in the calculations but also those for the (100) interfaces.

| Interface (A/B) | PBE-GGA | HSE06-hybrid | GW0@PBE | Experiment |
|----------------|---------|--------------|---------|------------|
| GaP/GaAs       | 0.46    | 0.53         | 0.56    | 0.57,173,174 |
| GaAs/AIAs      | 0.45    | 0.47         | 0.56    | 0.40,175,176 |
| GaAs/ZnSe      | 0.65    | 0.93         | 1.20    | 0.96,177 |
| AlSb/ZnSe      | 0.24    | 0.42         | 0.55    | 0.42 ± 0.07,178 |
| ZnTe/CdSe      | 0.62    | 0.79         | 0.77    | 0.64 ± 0.07,180 |

3.4 Point defects: Native defects, impurities, and dopants

3.4.1 Defect formation energy. Point defects in semiconductors involve various species, charge states, and spin states, resulting in diverse defect-induced material functionalities. The most fundamental quantity associated with these point defects is the energy of defect formation.181–187) With

Fig. 12. Band alignment of various semiconductor surfaces obtained using GWT4@HSE06.76,190 The experimental values are taken from Ref. 163 and are reported for nonpolar surfaces with orientations and reconstructed structures identical to those considered in the calculations; an exception is GaN, for which experimental values for nonpolar surfaces are not available, and those for polar {0001} surfaces are shown instead.
the configurational entropy, it determines the equilibrium defect concentrations and carrier densities. Furthermore, the thermodynamic transition levels, which correspond to donor, acceptor, or deep-trap levels, are given by the formation energies for different charge states.

In materials science, the Gibbs free energy of defect formation, \( \Delta G_i \), is often used because its fundamental variables, that is, the temperature, pressure, and composition, are relatively easy to control in experiments on solids, compared to the entropy, volume, and chemical potential. To obtain \( \Delta G_i \) theoretically, however, the evaluation of vibrational effects is required, for example, using phonon calculations and the quasi-harmonic approximation.\(^{187,188}\) This is still challenging, especially for charged defects, in terms of both computational complexity and expense. Assuming that the vibrational effects are not significant, at least at relatively low temperature, the total energy change associated with defect formation at 0 K and 0 GPa is typically considered. This quantity is simply called the formation energy, \( \Delta E_i \), and we use this formulation in the following: the vibrational contributions evaluated separately can be added to \( \Delta E_i \), if necessary.

For a point defect \( D \) in charge state \( q \) (\( \text{D}^q \)), the formation energy is given as

\[
\Delta E_i[\text{D}^q] = E[\text{D}^q] - \sum_i N_i \mu_i + q\varepsilon_F, \tag{4}
\]

where \( E[\text{D}^q] \) denotes the total energy of a simulation model containing defect \( \text{D}^q \). \( N_i \) and \( \mu_i \) are the number and chemical potential of a constituent atom of type \( i \), respectively, and \( \varepsilon_F \) is the Fermi level.

The atomic chemical potentials are related to the crystal growth and/or doping conditions. They can vary under the constraints determined by phase equilibria. Taking the case of ZnO as an example (\( i = \text{Zn}, \text{O} \)), the relevant constraints are written as

\[
\mu_{\text{Zn}} + \mu_{\text{O}} = \mu_{\text{ZnO(bulk)}},
\mu_{\text{Zn}} \leq \mu_{\text{Zn(metal)}},
\mu_{\text{O}} \leq \frac{1}{2} \mu_{\text{O}_2(molecule)}, \tag{5}
\]

where \( \mu_{\text{ZnO(bulk)}}, \mu_{\text{Zn(metal)}}, \) and \( \mu_{\text{O}_2(molecule)} \) denote the chemical potentials of bulk ZnO, Zn metal, and the O\(_2\) molecule, respectively. The calculated total energies of each phase are typically used for these values, although more elaborate models consider the Gibbs free energies at a given temperature and pressure.

The upper limit of \( \mu_{\text{Zn}} \), in other words, the lower limit of \( \mu_{\text{O}} \), corresponds to the Zn-rich (O-poor) limit, which is described as

\[
\mu_{\text{Zn}} = \mu_{\text{Zn(metal)}},
\mu_{\text{O}} = \mu_{\text{ZnO(bulk)}} - \mu_{\text{Zn(metal)}}. \tag{6}
\]

The other limit, that is, the O-rich (Zn-poor) limit, is given as

\[
\mu_{\text{Zn}} = \mu_{\text{ZnO(bulk)}} - \frac{1}{2} \mu_{\text{O}_2(molecule)},
\mu_{\text{O}} = \frac{1}{2} \mu_{\text{O}_2(molecule)}. \tag{7}
\]

Note that these limiting conditions for the chemical potentials may be determined by other compounds when they are present. For instance, the O-rich limit in SnO is given as the equilibrium with bulk SnO\(_2\) rather than the O\(_2\) molecule. These cation-rich (O-poor) and O-rich (cation-poor) limits are often considered as representatives of low and high O\(_2\) partial pressure conditions, respectively; if necessary, the O\(_2\) partial pressure dependence can be explicitly discussed via thermodynamic modeling of the O\(_2\) gas phase, as described later in Sect. 3.4.3. In ternary systems, several competing phases may be present, including elementary substances, binary compounds, and ternary compounds other than the target material. The limiting conditions for the chemical potentials are readily determined using chemical potential diagrams, as reported, for example, in Refs. 159, 189, and 190. Systems with more components can be treated similarly.

Except for degenerate semiconductors where the Fermi level is located within the valence or conduction band, \( \varepsilon_F \) is typically assumed to vary between the VBM and CBM. The Fermi level is often measured with respect to the VBM level \( \varepsilon_{\text{VBM}} \) as

\[
\Delta \varepsilon_F = \varepsilon_F - \varepsilon_{\text{VBM}}. \tag{8}
\]

Then the variable range of \( \Delta \varepsilon_F \) becomes

\[
0 \leq \Delta \varepsilon_F \leq \varepsilon_F. \tag{9}
\]

When some crystal growth conditions or chemical potential values are assumed, the formation energies are thus described as a function of \( \Delta \varepsilon_F \). Although the formation energies of neutral defects do not depend on \( \Delta \varepsilon_F \), those of charged defects show specific dependences. Here, the positive and negative charge states (\( q > 0 \) and \( q < 0 \)) indicate donor and acceptor behavior of defects, respectively, and these defects are associated with carrier generation and compensation.

Diagrams showing the formation energy versus the Fermi level are often used for general discussion. This is illustrated in Fig. 13, taking as an example the O vacancy in ZnO at the Zn-rich (O-poor) limit given by Eq. (6). It is found that the 2+ charge state is energetically most favorable at low Fermi

Fig. 13. (a) Formation energies of the O vacancy in the neutral, +, and 2+ charge states in ZnO at the Zn-rich (O-poor) limit. The values are based on the total energies from the HSE06 = 0.375 hybrid functional calculations with finite cell-size corrections reported in Ref. 63 and shown as a function of the Fermi level according to the definition of the defect formation energy in Eq. (4). The bold line denotes the most energetically preferable charge states at a given Fermi level position. The Fermi level at which the lines for two charge states intersect corresponds to the thermodynamic transition level, as given by Eq. (10). The O vacancy in the + charge state is unstable at any Fermi level position; therefore, the \( \varepsilon(2+/0) \) level appears, as depicted by the open circle. (b) \( \varepsilon(2+/0) \) thermodynamic transition level corresponding to the open circle in (a), presented in an energy level diagram. The 2+ and 0 charge states are energetically most favorable for Fermi level positions below and above the \( \varepsilon(2+/0) \) level, respectively.
level positions, whereas the neutral charge state is energetically most favorable at high Fermi level positions, as in n-type ZnO. The + charge state is never stable. This behavior is often denoted as negative $U$ and is also found for point defects in other materials.

### 3.4.2 Thermodynamic transition level and optical transition energy

The thermodynamic transition levels of point defects are often simply called defect levels and are further categorized into donor, acceptor, and deep-trap levels in discussions of semiconductors. These thermodynamic levels involve atomic relaxation effects and therefore are different from the optical transition energies described below. By definition, the thermodynamic transition level corresponds to the Fermi level at which the formation energies of a defect in two charge states, $q$ and $q'$, are equal.\(^{181-183,191}\)

When measured from the VBM, it is given as

$$e(q/q') = \frac{\Delta E_i[D^q, \Delta E_F = 0] - \Delta E_i[D^{q'}, \Delta E_F = 0]}{q' - q},$$

where $\Delta E_i[D^q, \Delta E_F = 0]$ denotes the formation energy of defect $D^q$ for $\Delta E_F = 0$. In Fig. 13, the thermodynamic transition level of the O vacancy in ZnO is given as an example; other examples are also given later in Sect. 3.4.6.

The optical (vertical) transition energies associated with the absorption and emission of photons can be obtained using the Franck–Condon principle.\(^{182,187}\) Typically, it is evaluated similarly to the thermodynamic transition level but with the atomic coordinates frozen at those of the initial state. Alternatively, single-particle or quasiparticle energies can be used to estimate the optical transition energies using the atomic geometries in the initial state. Note that these schemes neglect the effects of electron–hole interactions on the electronic transition, that is, excitonic effects, which are sizable when localized defect states are involved. For instance, a GW and Bethe–Salpeter study has shown that excitonic effects lead to a ~0.2 eV red shift for the C vacancy in SiC.\(^{192}\) Such an approach including excitonic effects is ideal but rather computationally demanding when applied to point defects in supercells.

### 3.4.3 Equilibrium Fermi level position, defect concentration, and carrier density

Once the formation energies of the relevant defects are obtained, the equilibrium Fermi level at a given temperature is derived simultaneously with the equilibrium charged defect concentrations and carrier densities via the charge neutrality condition.\(^{105,182,187,189,191,193,194}\) If the vibrational entropy contribution to $\Delta G_i$ and the pressure effect are negligibly small, the concentration of dilute defect $D^q$ is given using $\Delta E_i[D^q]$ as

$$C[D^q] = N[D^q] \exp \left( \frac{-\Delta E_i[D^q]}{k_B T} \right),$$

where $N[D^q]$ is the number of sites per unit volume for defect $D^q$ times its spin degeneracy, $k_B$ is the Boltzmann constant, and $T$ is the absolute temperature. The hole density in the valence band ($p$) and the electron density in the conduction band ($n$) are given as

$$p = \int_{E_{VBM}}^{E_{FBM}} D(e) \frac{1}{\exp[(e - e_F)/k_B T] + 1} \, de,$$

$$n = \int_{E_{FBM}}^{E_{VBM}} D(e) \frac{1}{\exp[(e - e_F)/k_B T] + 1} \, de,$$

where $D(e)$ is the electronic DOS for the perfect crystal, assuming that the presence of dilute defects does not essentially affect the DOS. The defect concentrations and carrier densities are constrained by the charge neutrality condition as

$$p - n + \sum q_i C[D^{q_i}] = 0.$$  \(\text{(14)}\)

Under given chemical potential conditions, the concentrations of neutral defects at $T$ are determined only by Eq. (11). Self-consistent solutions to Eqs. (11)–(14) provide the concentrations of charged defects, the densities of carrier electrons and holes, and the Fermi level under thermal equilibrium. It should be noted that the thermodynamic transition levels, that is, the donor, acceptor, and deep-trap levels, are irrelevant to the resultant defect concentrations and carrier densities.

In real materials, point defects introduced during crystal growth or doping at elevated temperature are somewhat or strongly quenched to room temperature. These situations can also be simulated under the assumption that the defect concentrations are frozen at the temperature before quenching, and electrons and holes are redistributed between the defect levels and host bands.\(^{189,191,194}\) The thermodynamic transition levels now play crucial roles, as some of redistributed carriers are trapped there. From another viewpoint, carriers are thermally excited from the donor and acceptor levels in this situation; this scenario is often considered in the discussion of real semiconducting materials.

The gas partial pressure dependence of the defect concentrations and carrier densities is often of interest, in addition to the temperature dependence. In this case, the temperature and partial pressure dependence of the chemical potentials of the gas phases is taken into account using the ideal gas model;\(^{182,195}\) for high-pressure conditions, the fugacity should be used instead of the partial pressure.\(^{190,196}\) A more rigorous approach also considers the vibrational contributions of solid phases via phonon calculations or molecular dynamics simulations.\(^{187,188}\)

### 3.4.4 Supercell approach to point defects

In practical first-principles calculations of point defects, the supercell approach is often adopted. A point defect is placed in a supercell, which is constructed by expanding a primitive or conventional unit cell so that the bulk-like region far from the defect is sufficiently large. The number of atoms in the resulting supercells is typically 50 to a few hundred, or sometimes even larger, depending on the characteristics of the host materials and defects as well as the balance between the required accuracy and affordable computational costs. As in calculations for perfect crystals based on band theory, the supercell, and therefore a point defect therein, is repeated under three-dimensional periodic boundary conditions. To model defects at the dilute limit, corrections are needed to eliminate spurious interactions, particularly electrostatic interactions between a charged defect, its periodic images, and a charge-compensating jellium background, that is, image-charge interactions.\(^{197,198}\)

On the basis of Eqs. (4) and (8), the defect formation energy is obtained using the supercell approach as

$$\Delta E_i[D^q] = E[D^q] + E_i[D^{q_i}] - E_p - \sum \Delta N_i \mu_i + q(e_{VBM} + \Delta e_F),$$

where $\Delta N_i \mu_i$ is the partial pressure dependence of the chemical potentials of the gas phases.
where $E[D_i]$ and $E_p$ are the total energy of the supercell containing defect $D_i$ and that of the perfect-crystal supercell, respectively. $\Delta N$ is the difference in the number of an atom of type $i$ between these supercells. For instance, $\Delta N_{Zn} = 0$ and $\Delta N_{O} = -1$ for the O vacancy, and $\Delta N_{Zn} = 1$ and $\Delta N_{O} = 0$ for the Zn interstitial in ZnO.

The correction term $E_c[D_i]$ is required for charged defects to eliminate the energy contribution of the image-charge interactions. This issue has been extensively discussed, and several correction schemes have been proposed. Among them, we prefer the scheme proposed by Freysoldt et al. because of its theoretical elegance and generality. Additionally, we have extended their scheme so that the corrections are effective for anisotropic systems such as layered structures and the atomic geometries after relaxation, which allows for application of the correction scheme to diverse three-dimensional systems. Correction schemes for one- and two-dimensional systems have also been developed recently by several researchers. Historically, potential alignment corrections have been applied to $\epsilon_{vBM}$. but it has been proved recently that potential alignment is unnecessary if the image-charge correction is properly applied.

Other necessary corrections include those associated with band-filling effects. When defects induce shallow acceptor (donor) states with host valence (conduction) band characteristics, which are called perturbed host or hydrogenic effective-mass states, the valence (conduction) bands are filled by holes (electrons) in finite-sized supercells. If heavily doped systems are of interest, such models with high carrier densities may be used to discuss the Burstein–Moss effects (see Sect. 3.1.4), keeping in mind that defects are regularly repeated in the supercells, in contrast to real materials. Otherwise, corrections to the band-filling effects are required for predicting the defect formation energies at the dilute limit.

It should be noted that defects with perturbed host states do not behave as expected from the assumed charge states, because the defect-induced holes or electrons spread over supercells. For example, shallow donor defects of this type in the neutral charge state are essentially identical to those in positive charge states. Therefore, special treatments are required for the evaluation or correction of the formation energies of defects with perturbed host states. Moreover, determination of the shallow defect levels associated with hydrogenic states is challenging, as these states could extend to several unit cells, that is, far beyond the supercell size conventionally used. Zhang et al. have reported a computational approach to tackle this issue, which combines the GW calculation for the central-cell potential with a potential patching method for large systems containing 64,000 atoms to describe the wavefunctions of hydrogenic defect states. They demonstrated good reproduction of experimental values of acceptor levels in Si and GaAs.

Defect formation energies and thermodynamic transition levels are thus discussed on the basis of the total energies of supercell calculations including appropriate finite-cell-size corrections. Hydrogenic shallow states are, however, difficult to treat using conventional supercells and are often discussed only qualitatively. The one-electron (single-particle) or quasiparticle energies of defect supercells, as well as their total energies, can be used to estimate the optical transition energies, where the atomic coordinates are frozen upon the electronic transition; ideally, excitonic effects are considered here, as mentioned in Sect. 3.4.2. Such single-particle or quasiparticle energies are also useful for qualitative understanding of defect states, as illustrated in the following section; note that image-charge corrections are needed for single-particle and quasiparticle levels associated with charged defects, as well as the total energies of their supercells. Supercell approaches to point defects and related issues are also described elsewhere, for example, in Refs. 97, 181–183, and 185–187.

### 3.4.5 Approximation to exchange–correlation interaction

The accuracy of the calculated total energies is central to reliable prediction of the defect formation energies and thermodynamic transition levels, where an appropriate choice of the approximation to the exchange–correlation interaction is most important. Further, as the formation energies of charged defects depend on the Fermi level, accurate prediction of the VBM and CBM, which serve as the references for the Fermi level, is essential as well. These band edge positions are also highly dependent on the approximation, as discussed in Sects. 3.1.1 and 3.3.

As an approach satisfying both requirements sufficiently, hybrid functionals are widely employed in recent theoretical investigations into point defects in semiconductors, although the calculations are much more expensive than those using the LDA(+U) and GGA(+U). To accelerate the calculations, a combination of the LDA(+U)/GGA(+U) total energies with band edge determination using hybrid functionals or the GW approximation has been considered. This approach is justified by the reasonable correspondence between the defect levels from LDA(+U)/GGA(+U) and hybrid functional calculations when the band edges are corrected, as demonstrated for defects in SiC, AlN, GaN, Cu3N, Zn3N2, and ZnO. Note, however, that the LDA(+U)/GGA(+U) cannot well describe the polaronic electron and hole localization associated with point defect formation, except when the +U corrections can be directly applied to these polaronic states. Such localization is typically coupled with particular atomic relaxations, potentially leading to large energy errors when the localized states are not appropriately described. To remedy this, a scheme relevant to the +U approach has been suggested by Lany and Zunger in which corrections are applied so that defect states meet a generalized Koopmans condition, that is, independence of defect-induced energy levels from the electron occupation number.

Another combined approach has been applied in studies of the self-interstitial in Si by Rinke et al. and the O vacancy in HfO2 by Choi and Chang, where the vertical transition energy contributions to the defect formation energies are estimated using the GW approximation, and the other contributions are evaluated using the LDA (or other approximations such as the GGA). Jain et al. later used a similar approach to investigate the O vacancy in HfO2, additionally including electrostatic corrections. Many-body methods for total energy evaluation have also been applied to point defects in semiconductors, such as random phase approximation of the correlation energy and quantum Monte Carlo methods. These results are important as benchmarks, although the calculations are demanding.
3.4.6 Applications of point defect calculations to defect-related issues in semiconductors. Many first-principles studies of point defects have been reported for semiconducting materials thus far, revealing the behavior of defects at the atomic and electronic levels as well as the defect energetics. Some examples of such studies are reviewed here.

Energetics and electronic levels of native defects and impurities in ZnO. It is well recognized that ZnO is a representative oxide semiconductor with a wide variety of applications. The fundamental characteristics of its native defects and residual impurities have been investigated both experimentally and theoretically to reveal the origins of observed properties such as the off-stoichiometry with Zn excess (O deficiency), n-type conductivity, optical absorption and emission, and so forth, as reviewed, for example, in Refs. 97 and 226–232. Either the O vacancy or Zn interstitial is often assumed to be a dominant native defect species and is associated with the off-stoichiometry and n-type conductivity. However, such assumptions are controversial, and this issue has been addressed using first-principles calculations. In early studies,\textsuperscript{193,233–241} the calculations were carried out using the LDA, GGA, LDA+U, or GGA+U, which resulted in severe band gap underestimation, as mentioned in Sect. 3.1.1. Discussion of the defect formation energies and electronic levels was therefore challenging, even though several types of post-calculation corrections were suggested.\textsuperscript{193,234,238,239} Later hybrid functional studies, which reproduced the band gap well, yielded different results even qualitatively for some characteristics.\textsuperscript{63,242–249}

As an example, Fig. 14 shows the formation energies and thermodynamic transition levels of native defects in ZnO from HSE(\(\alpha = 0.375\)) hybrid functional calculations.\textsuperscript{97} The dependence of the formation energies on the chemical potential conditions, which vary between the Zn-rich (O-poor) and O-rich (Zn-poor) limits in Eqs. (6) and (7), respectively, is understandable from the definition given by Eqs. (4) and (15). For instance, the formation energies of defects associated with O deficiency or Zn excess, that is, the O vacancy, Zn interstitial, and Zn antisite, are lowest, whereas those associated with Zn deficiency or O excess, that is, the Zn vacancy, O interstitial, and O antisite, are highest at the Zn-rich (O-poor) limit [Fig. 14(a)], and vice versa at the O-rich (Zn-poor) limit [Fig. 14(b)]; note that the O interstitial and antisite are not included in Fig. 14 because of their relatively high formation energies. In other words, the O vacancy, Zn interstitial, and Zn antisite are relatively easily formed due to weak O\(_2\) partial pressure, whereas the Zn vacancy, O interstitial, and O antisite are easily formed under high partial pressure. Another important observation from Figs. 14(a) and 14(b) is the donor or acceptor behavior of each type of defect; positive and negative gradients, that is, charge states, indicate donors and acceptors, respectively. Finally, the defect levels are visible in Figs. 14(a) and 14(b), and are also shown in Fig. 14(c) in the form of an energy level diagram. The primary conclusions on native donor-type defects from the theoretical defect energetics are as follows: the O vacancy shows a low formation energy and a rather deep \(s^2(2+/0)\) donor level owing to the negative \(U\) behavior, as mentioned in Sect. 3.4.1, and the Zn interstitial and Zn antisite have high formation energies, although they are shallow donors. The observed nonstoichiometry is understandable from the easy formation of the O vacancy, whereas the n-type conductivity is explained by neither the O vacancy nor the Zn interstitial.

The characteristics of the O vacancy and Zn interstitial can also be determined from their one-electron structures and local atomistic structures, which are shown in Fig. 15. The neutral O vacancy gives rise to a localized one-electron state deep in the band gap and causes inward relaxation of adjacent Zn ions by as much as 10%. The 2+ charge state behaves differently, inducing no localized in-gap state and weak perturbation of the lowest conduction band. The neighboring Zn ions show outward relaxation by 23%. Similarly, the neutral Zn interstitial perturbs the lowest conduction band only slightly. Except for the electron occupancy, the band structures are essentially the same for the + and 2+ charge states.\textsuperscript{83} These features in the one-electron structure are typical of shallow donors with hydrogenic, perturbed host states. Note again that as mentioned in Sect. 3.4.2, the one-electron defect levels do not correspond to the thermodynamic transition levels determined by the formation energies of the two relevant charge

![Fig. 14. Formation energies and thermodynamic transition levels of the O vacancy (V\(_O\)), Zn interstitial (Zn\(_i\)), Zn antisite (Zn\(_0\)), and Zn vacancy (V\(_Zn\)) in ZnO at the (a) Zn-rich (O-poor) and (b) O-rich (Zn-poor) limits. The values are based on the total energies from the HSE\((\alpha = 0.375)\) hybrid functional calculations with finite-cell-size corrections reported in Ref. 63. The lower and upper limits of the Fermi level are set to the VBM and CBM, respectively. In contrast to Fig. 13, only the charge states that are energetically the most favorable at a given Fermi level are shown for each defect species. The thermodynamic transition levels as defined by Eq. (10) are indicated by open circles. The \(e^{2+/+}\) and \(e^{+/0}\) levels of Zn and Zn\(_0\) are very close to each other; they are denoted as the \(e^{2+/+}\) levels and indicated by single circles. (c) Thermodynamic transition levels corresponding to the open circles in (a) and (b), presented in an energy level diagram. The levels for the H impurity at an interstitial site, specifically, a Zn–O bond center OH\(-\)like configuration (H\(_i\)), and that at the O site (H\(_0\)) are also included. A polaronic hole trap around the Zn vacancy reported in Refs. 246 and 248, which results in a lower formation energy of the neutral charge state and the appearance of the + and 2+ charge states at low Fermi level positions, is not considered here (see text and Fig. 16 for details). Adapted from Ref. 97. © 2011 National Institute for Materials Science.](https://doi.org/10.7567/0003-6907-11-6-060101-14)
states in their own equilibrium geometries. Nevertheless, the depth of the defect levels can be qualitatively understood in terms of the one-electron structures, as illustrated here by the cases of the O vacancy and Zn interstitial in ZnO.

Van de Walle and Janotti proposed that H impurities unintentionally incorporated into an interstitial (bond-center) site and the substitutional O site are both shallow donors in ZnO on the basis of their LDA and LDA+U calculations. The shallow donor behavior of the H impurities is reproduced by HSE(α = 0.375) hybrid functional calculations, as shown in Fig. 14(c). Other proposed sources of shallow donors in ZnO include a complex of a Zn interstitial and N impurity and the Zn interstitial when it is stabilized in the presence of a high concentration of O vacancies owing to strong attractive interactions. Metastable shallow donor behavior of the O vacancy and the (2H–VZn) complex has also been reported to elucidate the persistent photoconductivity observed in ZnO.

The Zn vacancy has been considered the dominant acceptor-type defect in ZnO. As shown in Fig. 14, it generates deep acceptor levels. The formation energy is generally high and positive even for the Fermi level position at the CBM, especially under the Zn-rich condition. Therefore, the Zn vacancy does not significantly compensate for carrier electrons, except when the Fermi level lies inside the conduction band, as in heavily doped ZnO in transparent electrode applications. Recently, a polaronic hole trap around the Zn vacancy has been proposed by Petretto and Bruneval and Frodason et al. but it is not considered in the results presented in Fig. 14. Up to four holes are localized at the O ions adjacent to the Zn vacancy, which is accompanied by distinct outward, off-symmetric relaxation of the O ions, as shown in Fig. 16. Because of the polaronic trap, the formation energy of the neutral Zn vacancy becomes lower than that shown in Fig. 14, and the + and 2+ charge states appear at low Fermi level positions.

$p$-type doping of ZnO is known to be rather challenging. Both experimental and theoretical efforts have been made to overcome the difficulty, as reviewed by Avrutin et al. in Ref. 230. Reported first-principles studies suggest the following fundamental reasons: First, a strong energetic preference for donor-type native defects leads to compensation for doped holes, as discussed above. Rather high or even non-equilibrium O-rich growth conditions are required to suppress the formation of such defects. Further, residual donor impurities such as H and Al should be eliminated. Second, isolated dopants that act as shallow acceptors are unavailable, according to recent systematic calculations by Yim et al. Impurity doping in a defect complex form may be a possible route to resolving this issue.

Native defects and doping in tin sulfides. The tin sulfides SnS, SnS2, and SnS2 are attracting growing interest. The best studied among the three phases is presumably SnS with Sn(II), which has been considered for use in solar cell absorber applications in recent decades. SnS2 with Sn(IV) takes a characteristic layered structure and has been examined for use as, for instance, an electrode in Li ion batteries and a visible-light photocatalyst for water splitting. SnS3 is a well-established mixed-valence Sn binary compound that inherits the structural and electronic features of both Sn(II) and Sn(IV). Typically, undoped SnS exhibits $p$-type conductivity whereas SnS2 and SnS2 are $n$-type conductors. To understand this carrier polarity, their native defects have been investigated using first-principles calculations; a doping strategy toward inversion of the carrier type has also been proposed.

As an example of such studies, Fig. 17 shows the theoretical formation energies of native defects in SnS, SnS2, and SnS2 reported by Kumagai et al. For all the phases, the overall defect energetics are consistent with the aforementioned carrier types. The specific behavior of native defects in each phase is now presented.

![Fig. 15.](image1.png) **Fig. 15.** Electronic band structure for $4 \times 4 \times 3$ supercells of the perfect ZnO crystal, the O vacancy ($V_O$) in the neutral and $2^+$ charge states, and the Zn interstitial ($Zn_i$) in the neutral charge state, obtained using the HSE($\alpha = 0.375$) hybrid functional. Zero energy is set to the VBM. Presented below are the charge densities (squared wavefunctions) of the states indicated by arrows, which are plotted in the middle of the (0001) Zn and O planes adjacent to the defect, alongside the projected Zn (green) and O (yellow) ion positions. Adapted from Ref. 63 (© 2008 American Physical Society) and Ref. 97 (© 2011 National Institute for Materials Science).

![Fig. 16.](image2.png) **Fig. 16.** Hole density around the Zn vacancy in ZnO obtained using the HSE($\alpha = 0.375$) hybrid functional. The gray and red circles denote Zn and O ions, respectively; the O ions at azimuthal (az) and $c$-axial (ax) positions are indicated by arrows. The Zn vacancies in the $-$, 0, $+$, and $2^+$ charge states trap one to four holes at the adjacent O ions, respectively. Reprinted from Ref. 248. © 2017 American Physical Society.
In SnS, the Sn vacancy is a dominant defect at most Fermi level positions under the Sn-poor condition, as shown in Fig. 17(a). It gives rise to a hydrogenic shallow acceptor state located almost at the VBM, although such hydrogenic states are not included in the figure owing to the difficulty of treating them by the conventional supercell approach, as discussed in Sect. 3.4.4. This implies that SnS prepared under such conditions should exhibit p-type behavior, although some carrier compensation by the donor-type S vacancy is expected at low Fermi level positions. Under the Sn-rich condition, the formation energy of the Sn vacancy is not extremely low even when the Fermi level is located near the CBM, indicating a good chance for conversion into the n-type by the use of donor impurities. Indeed, the fabrication of n-type SnS by Cl doping was reported recently by Yanagi et al.262) According to a combined experimental and theoretical study by Ran et al., n-type doping of SnS is also possible via alloying with PbS.257) The role of alloying has been explained in terms of the expansion of the interstitial sites, which facilitates the formation of donor-type Sn and Pb interstitials.

Moving on to Sn$_2$S$_3$ [Fig. 17(b)], the donor-type Sn interstitial exhibits a low formation energy, as expected from its relatively large open space at the interstitial site. The formation energy of the Sn interstitial is negative at Fermi level positions. Under the Sn-rich condition, the formation energy of the Sn vacancy is not extremely low even when the Fermi level is located near the CBM, indicating a
level positions below \( \sim 0.2 \) eV, where spontaneous hole compensation takes place. Nonetheless, p-type conversion could be obtained using an acceptor dopant because the lower limit of the Fermi level at \( \sim 0.2 \) eV is close to the VBM. On the basis of systematic calculations of candidate acceptor dopants, K is proposed to be an effective acceptor.79) The K dopant is preferentially located at the Sn(II) site in SnS\(_2\), rather than at the interstitial site, because its ionic radius is sufficiently large. This is an important consequence because interstitial K dopants act as donors.

Four types of donor-type defects are easily formed in SnS\(_2\), as shown in Fig. 17(c). Among them, the Sn interstitial has the lowest formation energy under most chemical potential and Fermi level conditions. This behavior can be attributed to the large interlayer space in the crystal structure of SnS\(_2\). Although the Sn interstitial is a donor-type defect, it exhibits somewhat deep levels below the CBM. Therefore, a high carrier density is not expected in the absence of intentional or unintentional impurities.

These features of native defects specific to the three tin sulfide phases are systematically understood in terms of the valence of Sn in each phase and the resultant crystal and band structures.79)

Native defects in organic–inorganic hybrid perovskite \( CH_3NH_3PbI_3 \). The organometal halide perovskite \( CH_3NH_3PbI_3 \) is currently a major target of research on highly efficient and low-cost solar cell applications.263) Yin et al. systematically investigated the formation energies and electronic levels of native point defects in \( CH_3NH_3PbI_3 \) using first-principles calculations.264) The reported theoretical defect levels are shown in Fig. 18. They found that the dominant native defects are the Pb vacancy and the \( CH_3NH_3^+ \) interstitial under most chemical potential conditions, and these defects act as a shallow acceptor and a shallow donor, respectively. Not only the Pb vacancy and the \( CH_3NH_3^+ \) interstitial but also other defects with low formation energies do not generate deep in-gap levels, which is advantageous for minimizing carrier recombination in the photosorber. These defect properties can be understood in terms of the electronic structure characteristic of hybridization of Pb 6s and I 5p,264) as discussed in Sects. 3.1.1 and 4.1.

Native defect concentrations and carrier densities in SnO. Using the calculated DOS and defect formation energies, the defect concentrations and carrier densities are evaluated as described in Sect. 3.4.3. Figure 19 shows the calculated equilibrium native defect concentrations and carrier densities in SnO at the O-rich limit. Defect formation energies obtained using the PW91-GGA functional with a band gap correction were used in the evaluation. \( V_{Sn} \), \( V_O \), and \( O_i \) denote the concentrations of the Sn vacancy, O vacancy, and O interstitial, respectively, where the sum of the values for the relevant charge states is taken for each defect species; the contributions of each charge state are also shown for \( V_{Sn} \) and \( V_O \). The concentration of the Sn interstitial should be rather low owing to its high formation energy (\( >3 \) eV) and is not considered here. \( p \) and \( n \) indicate the hole and electron densities, respectively. Adapted from Ref. 105. © 2006 American Physical Society.

Coupling of point defect states with local atomic structures: DX centers and vacancies. Localized defect-induced states are often accompanied by specific local atomic relaxations, as discussed in Sect. 3.4.5. The best-known example of such strong defect state-atomic structure coupling is probably the DX center in Al,Ga,\( _{1-x} \)As alloys.267) According to first-principles calculations by Chadi and Chang, a substitutional Si dopant is substantially displaced toward an
interstitial site. This results in the formation of deep in-gap donor states, as opposed to the shallow donor behavior of conventional substitutional dopants. Lany and Zunger have theoretically proposed the presence of intrinsic DX centers in chalcopyrite CuInSe$_2$ and CuGaSe$_2$, where the neutral In and Ga antisite defects displaced from the Cu site act as DX centers. This is illustrated for CuGaSe$_2$ in Fig. 20. The deep donor nature of the intrinsic DX center in CuGaSe$_2$ could make it difficult to maintain high efficiencies in CIGS solar cells when the Ga content is increased. Kumagai et al. later reported that the neutral Sn-on-Zn antisite in chalcopyrite ZnSnP$_2$ likewise becomes an intrinsic DX center.

Analogous behavior of Ti antisite defects in perovskite SrTiO$_3$ has been reported by Choi et al. on the basis of their first-principles calculations. As shown in Fig. 21, the Ti antisite at the Sr site is displaced in either the [100] or [110] direction, resulting in two types of off-centered configurations. These off-centered Ti antisites have several interesting features, such as low formation energies comparable to that of the O vacancy under Ti-rich conditions, the generation of local dipole moments resulting from the off-centering, switching of the directions of the off-centering and dipole moments with a low activation energy of ~0.2 eV, and deep donor levels associated with the localization of Ti 3d electrons. The experimentally observed ferroelectricity in nonstoichiometric SrTiO$_3$ films can be explained by the formation of such antisite defects.

Another example of strong defect state-atomic structure coupling is polaronic hole trapping by the Zn vacancy in ZnO, where a distinct off-symmetric relaxation of the hole-trapping O ions occurs, as shown in Fig. 16. A polaronic trap of electrons and a concurrent atomic displacement have also been reported for the O vacancy in SrTiO$_3$, whereas related metastable vacancy configurations have been found in other perovskites such as BaTiO$_3$, KTaO$_3$, and NaTaO$_3$. Another type of characteristic electronic and atomic configuration of the O vacancy in SrTiO$_3$ has been proposed by Choi et al., in which the O vacancy is accompanied by local TiO$_6$ octahedral rotation even in the cubic perovskite phase, which is otherwise free from octahedral rotation. In addition, the formation of O vacancy clusters in SrTiO$_3$ has been reported both experimentally and theoretically. The relevance of these isolated O vacancies and clusters to the observed n-type conductivity and optical properties, particularly blue light emission from reduced SrTiO$_3$, has been discussed on the basis of the results of first-principles calculations.

**Defect complexes.** The atomistic and electronic structure of defect complexes in semiconductors, as well as their binding or association energies, has been investigated using first-principles calculations. Complexes may consist only of native defects or involve residual impurities or dopants. O vacancy clusters in SrTiO$_3$ are an example of the former. As an example of a case where dopants are involved, Fig. 22(a) shows the structure of a complex consisting of a Ce dopant and four B vacancies in cubic BN predicted using first-principles calculations. The Ce dopant is unlikely to be located at the cationic B site in BN owing to the huge size mismatch. Indeed, first-principles calculations predicted an extremely high formation energy for this simple substitution. As a result of systematic calculations for diverse complex configurations, the structure shown in Fig. 22(a) was iden-
Semiconductors have attracted great interest recently. Inves-
tigation into H impurities is especially important for oxide
semiconductors, understanding of which is limited compared
to that of, for instance, Si and GaN. Although it has been
difficult to detect a low concentration of H impurities experi-
mentally, recent efforts have enabled the improvement of the
detection limit down to ~1 x 10^{15} cm^{-3}. H impurities in
oxide semiconductors often act as donors to generate carrier
electrons or to hinder p-type doping by hole compensation, as
illustrated by the aforementioned case of ZnO. These roles of
H impurities have also been discussed for other oxide
semiconductors on the basis of first-principles calculations.
Janotti and Van de Walle theoretically proposed the abundance
and shallow donor behavior of H impurities at host O sites, where the H impurities form multicenter bonds with
adjacent cations. This has provided better understanding of
the important roles of H impurities, in addition to typical
interstitial-like H ions that are bonded to host O ions in OH-
like configurations.

**Strategies for carrier doping and carrier-mediated ferromagnetism.** Carrier doping is extremely difficult for
some materials, particularly those having exceedingly wide
band gaps. Other materials are readily dopable into either
p- or n-type semiconductors, but inversion of the carrier type
could be challenging, as in the case of ZnO. First-principles
studies have been used to investigate these issues and propose possible doping strategies for various systems. Such
insights are especially useful for emerging or hypothetical
materials with unknown dopability. A typical approach is the
search for effective dopants by investigating the formation
energies and electronic levels of candidate elements at
various sites. Additionally, carrier compensation by native
defects and unintentional impurities, as well as incorporation
of the dopants themselves into unexpected sites, should be
examined. Such theoretical studies have been conducted,
for instance, to discuss n-type doping of SnS{sup 257,258,261} and
p-type doping of SnS{sub 2},{sup 259} as mentioned above, and to assess
the dopability of emerging materials such as cubic Si_{1-x}N_{x}
and Ge_{1-x}N_{x}.{sup 285,286} Investigation into defect properties is also
useful for in silico screening of candidate materials such as
p-type TCOs,{sup 110} 18-valence-electron compounds, and
ternary nitrides, as detailed later in Sect. 4.3. When the
ystals have large open spaces, as in layered structures,
doping into interstitial or interlayer sites could be effective, as
proposed for hexagonal BN.

Furthermore, doped carriers could be localized as polarons,
which would substantially degrade the carrier mobilities. The assessment of such polaron formation is especially important
for conventional oxide semiconductors, where relatively
localized O 2p states at the VBM could facilitate the forma-
tion of hole polarons, as systematically investigated by
Varley et al.

Magnetic dopants in semiconductors have been exten-
sively investigated for spintronic applications. Such doped
systems, particularly those showing carrier-mediated ferro-
magnetism, are referred to as dilute magnetic semiconduc-
tors and represented by Mn-doped GaAs. On the basis of the results of first-principles calculations, the local
atomistic, electronic, and magnetic structure of magnetic
dopants and the resultant macroscopic properties have been
discussed for a variety of established and potential dilute
magnetic semiconductors, as reviewed in Ref. 291.

**Amorphous In–Ga–Zn–O semiconductors.** The In–Ga–
Zn–O system is now recognized as a representative
amorphous semiconductor with increasing industrial applica-
tions in thin-film transistors. Notably, its conduction band
structure resembles that of the corresponding crystalline phase
owing to the spatially extended nature of the In, Ga, and Zn s
components, which leads to high electron mobility even in the
amorphous phase. This behavior is in stark contrast to

![Diagram of a Ce dopant structure in cubic BN](image-url)
amorphous Si with covalent bonds. Both experimental and theoretical studies indicate that it is important to understand and control defects in amorphous In–Ga–Zn–O, as well as in crystalline semiconductors. Several types of defects and their roles in carrier generation and trapping have been discussed on the basis of the results of first-principles calculations, for example, in Refs. 8 and 292–295.

4. Materials design and exploration: Design principles, high-throughput screening, and materials informatics

The above first-principles approaches are applicable to the design and exploration of novel materials. Continuous development of computer performance now makes it possible to conduct in silico high-throughput screening of a large number of candidate compounds using first-principles calculations and/or prediction models constructed via machine learning of computational data sets. Most in silico screening of semiconductors thus far reported used only first-principles calculations. The search space accessible by this approach is limited by the computational costs, except in cases where only entries in preexisting databases are inspected. The exploration of novel materials typically considers particular chemical compositions and crystal structures that are selected on the basis of some design principles. We begin this section with a review of reported design principles for band structures and defect properties. Publicly open databases compiling the results of first-principles calculations are then described. Finally, in silico screening using first-principles calculations either within or outside the entries of such databases is discussed, along with the construction of prediction models using data science approaches.

4.1 Design of band structure and defect characteristics

4.1.1 Band structure design based on chemical composition. Among compounds, nitrides and oxides are attractive owing to their earth-abundant nitrogen and oxygen constituents and thermal stability, especially for oxides. Although both p- and n-type doping of traditionally used compound semiconductors such as phosphides and arsenides is possible, many nitrides and oxides become only the n-type. To overcome this difficulty with p-type doping, design principles for nitrides, oxides, and mixed-anion compounds based on them have been developed, as described below.

In typical nitrides and oxides, the N 2p and O 2p states are the main contributors to electronic states near the VBM. These states are deep below the vacuum level compared to the anion p states in phosphides and arsenides; this chemical tendency is illustrated in Fig. 12, although, strictly speaking, the band positions depend on the chemical composition, crystal structure, and surface orientation, as mentioned in Sect. 3.3. The deep VBM tends to make hole doping difficult. In contrast, metal nitrides and oxides composed of elements such as Cu(I), Ag(I), Ti(I), Sn(II), Pb(II), Sb(III), and Bi(III) have cation orbital components near the N 2p and O 2p states, and the cationic and anionic states are hybridized with each other. This is shown schematically in Fig. 23 and illustrated by the case of SnO with Sn(II) in Fig. 4. Similarly, sizable hybridization near the VBM is identified in Sn(II) ternary oxides and Ba₂BiTaO₆ with Bi(III). An analogous effect is expected to some extent for oxides and nitrides of Zn and Cd, where the Zn 3d and Cd 4d states are located slightly below the O 2p and N 2p states; examples include ZnO (Fig. 3) and binary and ternary Zn nitrides, which are discussed later, in Sect. 4.3.

As a consequence of the presence of cationic states in the valence bands, these compounds typically have relatively high VBMs and small hole effective masses, both of which are favorable for p-type doping. Indeed, many Cu oxides readily become the p-type even without extrinsic doping. p-type conductivity has also been reported for Sn and Bi oxides such as SnO and Ba₂BiTaO₆ and this behavior is supported by first-principles studies of electronic structure and point defects.

Related electronic band structures are found in oxide-based mixed-anion compounds such as LaCuOS₁₋ₓSeₓ, as reported by Hiramatsu et al. and tetragonal ZrOS, as reported by Arai et al. In these compounds, the S or Se p states are located above the O 2p states in the valence bands and contribute mainly to the electronic states around the VBM. In other words, the anion states other than O 2p play a role analogous to that of the cation states. One may think that the resultant band structures around band gaps are similar to those of pure sulfides or selenides, but this does not necessarily hold because the presence of smaller and more ionic O ions could produce specific chemical environments for the S or Se ions. Such tunability of the band structures in mixed-anion compounds is attractive, as reviewed recently in Ref. 306.

The same concept appears to be useful for the design of other types of compounds. For instance, SnS with Sn(II) is known as a p-type semiconductor, as mentioned in Sect. 3.4.6, and its relatively small hole effective mass could be attributed in part to the Sn(II) contributions to the electronic states near the VBM. Organometal halide perovskites with Pb(II) have similar features in their band structures, as shown in Fig. 5. Ternary Cu(I) phosphides such as CaCuP, SrCuP, and BaCuP possess rather small hole effective masses, as reported recently by Williamson et al.
Another reported example of band structure design leading to semiconductivity is the consideration of 18-valence-electron ABX ternary compounds by Zakutayev et al.\textsuperscript{287} and Gautier et al.\textsuperscript{288} These compounds can have finite band gaps, by analogy with eight-electron AX binaries such as conventional III–V and II–VI compound semiconductors. As also mentioned later, they searched for previously unreported chemical compositions on the basis of these design principles and identified several promising semiconductors both theoretically and experimentally.\textsuperscript{287,288}

Alloying of heterostructural semiconductors, which is mentioned in Sect. 3.2, can yield electronic structures and properties that are dissimilar to those of alloy components. Peng et al.\textsuperscript{308} and Holder et al.\textsuperscript{143} reported that Mn$_1$–Zn$_2$O alloys in the wurtzite structure, which is originally taken by ZnO but not by MnO, exhibit valence band structures that support hole transport and band gap tuning in the visible region, enabling their application to photoelectrochemical water splitting.

4.1.2 Band structure design based on crystal structure and symmetry. The crystal structure and symmetry are important degrees of freedom for manipulating the band structure. Mizoguchi et al. reported that a high-pressure phase of SrGeO$_3$ with cubic perovskite structure has an exceptionally narrow band gap of 2.7 eV among binary and ternary Ge oxides.\textsuperscript{309} The band structure type is indirect, and the optical absorption onset corresponds to its direct gap of 3.5 eV. Furthermore, La doping transforms it into a degenerate n-type conductor, demonstrating the realization of a Ge-based TCO. The experimentally observed narrow gap behavior is understandable in terms of the calculated conduction band structure and the concept of superdegeneracy, which prohibits hybridization of relevant orbitals at the Γ point owing to a symmetry specific to the cubic perovskite structure (Fig. 24). Arai et al. later reported the presence of a nonbonding CBM state in a tetragonal phase of ZrOS and demonstrated its bipolar doping, together with a wide optical gap of ∼2.5 eV, deploying the parity-forbidden nature of the band-edge electronic transitions.\textsuperscript{305} Xiao et al. used an analogous orbital interaction approach to elucidate the exceedingly narrow band gap of β-BaZn$_2$As$_2$.\textsuperscript{310}

A symmetry-relevant band structure design of transparent bipolar-dopable semiconductors has been reported by Nie et al.\textsuperscript{311} and Hosono et al.\textsuperscript{266} where indirect-type band structures or direct-type band structures with parity-forbidden band-edge electronic transitions are advantageous for obtaining wide optical band gaps and bipolar doping simultaneously. Such band structures are recognized for CuInO$_2$\textsuperscript{311} as well as SnO\textsuperscript{104,105} (see Fig. 4), In$_2$O\textsuperscript{110} (see Fig. 6), and tetragonal ZrOS,\textsuperscript{305} as mentioned above; among these, SnO and ZrOS are indeed bipolar dopable.

Omata et al. reported the synthesis of β-CuGaO$_2$ in a wurtzite-derived β-NaFeO$_2$ structure via ion exchange of Na$^+$ ions in the β-NaGa$_2$O$_4$ precursor with Cu$^+$ ions in CuCl.\textsuperscript{312} In contrast to the ordinary α phase in the delafossite structure, the β phase shows direct-type band structure with a gap of 1.47 eV. This band structure matches the solar spectrum well, indicating the potential of β-CuGaO$_2$ for use as a solar cell photoabsorber.

First-principles studies have predicted that nearly free electron states, which spread spatially toward surfaces or interlayer sites, constitute the CBM in the two-dimensional BN monolayer and bilayer.\textsuperscript{289,313} Matsushita et al. reported that analogous states also exist and even appear at the CBM in three-dimensional covalent semiconductors such as SiC and AlN.\textsuperscript{314} The floating nature of these electronic states at interstitial sites explains the band-gap variation and anisotropic effective masses of SiC polymorphs in terms of the length of the interstitial channels\textsuperscript{315} and produces specific interfacial states.\textsuperscript{316} Such electronic structures are related to those of inorganic electrides\textsuperscript{317,318} and therefore could lead to interesting functionalities.

4.1.3 Design of defect properties. In practical applications of semiconductors, lattice defects are often crucial to the material properties or device performance. In particular, deep electronic states inside the band gap can trap carrier electrons and/or holes, degrading the efficiencies of light-emitting devices and photovoltaic cells. Ideally, the formation of point defects, dislocations, surfaces, and interfaces with such deep states is avoided as far as possible. This issue depends on both material’s intrinsic characteristics and the processes used for material synthesis and device fabrication. First-principles calculations can provide useful insights, especially into the former.

Zakutayev et al. have proposed the concept of defect-tolerant semiconductors, where dangling bond states associated with point defects and surfaces tend to be shallow in some particular band structures.\textsuperscript{299} For example, these band structures occur in Cu compounds, where antibonding interactions between the Cu 3d and anion p states constitute electronic states near the VBM, as mentioned above. When vacancies or surfaces are formed, dangling bond states would appear below such antibonding states, that is, within the valence bands. Zakutayev et al. developed their discussion of defect tolerance in Ref. 299 by taking Cu$_2$N as an example. Supporting their concept, relatively shallow cation vacancy states have been observed in Sn(II) binary and ternary oxides,\textsuperscript{105,110,302} as well as another Cu(I) binary, Cu$_2$O.\textsuperscript{298,319} The analogous defect properties of SnS with Sn(II) and CH$_3$NH$_3$PbI$_3$ with Pb(II), which are mentioned in Sect. 3.4.6, can also be understood from this viewpoint.

An essentially identical concept is reported for valence-orbital-derived interfacial states, on the basis of the analysis...
of grain boundaries in CuInSe₂ by Yan et al.320) and in ZnO by Oba et al.321) In ZnO, especially strong bonding interactions between the Zn 4s states constitute the CBM state; therefore, similar effects are expected for interfacial dangling bond states originating from the Zn 4s states.

### 4.2 Computational databases

Databases containing the results of first-principles calculations are useful in many ways. For instance, screening for materials exploration within databases is readily carried out; data are used to construct prediction models via machine learning; and data analysis leads to systematic understanding of materials.322–324 Several computational databases are currently open to the public, examples of which include the following (information as of January 7, 2018):

- **The Materials Project** at [www.materialsproject.org325](http://www.materialsproject.org) contains 69,640 entries for inorganic compounds, 530,243 entries for nanoporous materials, and so forth. This database covers a variety of material, thermodynamic, and electrochemical properties such as electronic band structures, elastic tensors, piezoelectric tensors, phase diagrams, and Pourbaix diagrams. These data are based on the results of first-principles calculations at the level of the GGA and GGA+U using the Vienna Ab initio Simulation Package (VASP)326,327) with additional energy corrections. Data search and analysis are available using the Materials Application Programming Interface (API)328) as well as the website.
- **The AFLLOW Distributed Materials Property Repository** at [www.aflowlib.org](http://www.aflowlib.org), where AFLLOW stands for Automatic-FLOW for Materials Discovery,329) contains 1,748,704 entries for materials, specifically, for a large number of alloys, and covers properties such as the electronic band structures and elastic properties. It supports the VASP and Quantum Espresso30,331) codes. The data in this repository are accessible via the website and the AFLLOWLIB REST API332).
- **The Open Quantum Materials Database (OQMD)** at [oqmd.org333](http://oqmd.org) contains 471,857 entries obtained using the PBE-GGA functional as implemented in the VASP code and covers thermodynamic and structural properties. In particular, phase diagrams can be visualized online for not only two- to four-component systems but also larger component systems. An API called qmpy is provided for data analysis.
- **The Electronic Structure Project (ESP)** at [gurka.physics.uu.se/esp](http://gurka.physics.uu.se/esp) contains ~60,000 entries for theoretical electronic structures obtained using the full-potential linear muffin-tin orbital method and the LDA.334) The ESP data can be downloaded via a web interface.
- **The Open Materials Database** at [openmaterialsdb.se](http://openmaterialsdb.se) contains 205,264 entries for materials and 2787 for computations. It is designed so that researchers can programmatically interact, search, and process the data using the High-Throughput Toolkit.
- **The Novel Materials Discovery (NOMAD) Repository** at [repository.nomad-coe.eu](http://repository.nomad-coe.eu) contains 44,179,006 entries on the results of first-principles calculations.335) A notable feature of this repository is that it compiles computational results obtained using various program codes by different users. It currently supports 14 types of codes, and support for more is planned. Repository users can upload their own data, which may be restricted to the owners or made open to the public.
- **The Theoretical Crystallography Open Database (TCOD)** at [www.crystallography.net/tcod](http://www.crystallography.net/tcod) collects theoretically calculated or refined crystal structures of organic, inorganic, and metal–organic compounds and minerals, excluding biopolymers. There are currently 2682 entries.

To construct databases efficiently, first-principles calculations need to be performed for diverse systems automatically or at least semi-automatically. Several groups have developed program codes for this purpose, including those used for the Materials Project, AFLLOW, OQMD, ESP, and Open Materials Database329,333,334,336) and others, for example, by Pizzi et al. in the Automated Interactive Infrastructure and Database for Computational Science (AiIDA) project337) and by Yim et al.338) Automation of fundamental and defect property calculations has also been reported, for example, band structure diagram construction,88,89) evaluation of electronic transport properties including effective masses,107) surface modeling,164,165) and point defect modeling.249,339–342)

### 4.3 In silico high-throughput screening and materials informatics toward materials discovery

Screening using first-principles calculations is effective for exploring materials having desired functionalities if the target functionalities are closely related to the fundamental bulk and defect properties and if these properties are calculated with sufficient accuracy and speed. This typically holds for semiconductors; the predictable bulk and defect properties mentioned in Sect. 3 are excellent descriptors in many cases. Materials are screened for such properties on the basis of design principles, for instance, those mentioned above. The procedure is schematically illustrated in Fig. 25. A straightforward approach is to search within preexisting computational databases. When previously unreported chemical compositions are explored, candidate systems are often constructed by replacing constituent elements of known crystals.343) However, novel materials could take as-yet-unreported crystal structures; therefore, global crystal structure searches are also performed for a given chemical composition,166) for example, using evolutionary,344–346) random-search,347) minima hopping,348) and conformational space annealing algorithms.349) Further, thermodynamic stability against decomposition into other phases and dynamic stability against lattice vibration need to be assessed for previously unreported materials; these characteristics are evaluated via phase diagram and phonon calculations, respectively, as done in Refs. 190, 287, 288, and 350. As a result of such first-principles screening, several studies successfully identified promising semiconductors, some of which have been verified experimentally, as described below.

**p-type TCOs.** Hautier et al. explored p-type TCOs using first-principles calculations of the band gaps, effective masses, and dopability.110) The considered candidates are 3052 oxides that are existing minerals or already synthesized materials and are reported in the Materials Project database.325) As a result of screening, they identified the promising systems shown in Fig. 26. Some of the identified compounds contain Sn(II), Pb(II), and Ti(I), for example, K₂Sn₃O₇, K₂Pb₂O₃, PbTiO₃, and Ti₃V₂O₇, where electronic states around the VBM consist mainly of cationic orbitals; Others are mixed-anion systems...
such as Ca$_4$P$_2$O, ZrOS, and B$_6$O, where B is reported to behave as both a cation and an anion in B$_6$O. Also identified is Sb$_4$Cl$_2$O$_5$, which exhibits an Sb(III) orbital contribution to the VBM and has two types of anion components; therefore, this compound is categorized as belonging to both classes. These features in the band structure are compatible with the design principles discussed in Sect. 4.1.

Sarmadian et al. used an analogous approach to search for p-type TCOs, including mixed-anion systems. A total of 12,211 candidate materials taken from the AFLLOW Distributed Materials Property Repository were considered in their first-principles screening, from which A$_2$SeO$_2$ systems, where A = La, Pr, Nd, and Gd, were identified as potential p-type transparent conductors.

Li et al. explored previously unreported ternary oxides of Sn(II) and alkaline-earth metals. Through global structure searches using a combination of first-principles calculations and a particle swarm optimization technique, they identified SrSn$_2$O$_3$ and BaSn$_2$O$_3$ as stable compounds with electronic structures suitable for p-type conductivity.

n-type TCOs. First-principles screening has also been carried out to identify not-yet-considered but promising n-type TCOs. Hautier et al. reported the screening of more than 4000 binary and ternary oxides for this purpose. Further, they derived chemical rules producing small electron effective masses via analysis of a large data set. Binary and ternary n-type TCOs with bixbyite structure have also been explored by Sarmadian et al.

Solar cell photoabsorbers. As mentioned in Sect. 3.1.4, Yu, Zunger, and their coworkers calculated the photovoltaic energy conversion efficiencies as a function of absorber thickness and screened photoabsorber candidates on the basis of the predicted efficiencies. They considered Cu ternary and Ag ternary chalcogenides and ABX half-Heusler compounds as target systems and identified promising materials. Examples of the identified compounds are shown in Fig. 27, where the predicted efficiencies are compared to the Shockley–Queisser efficiency limit.

Metal chalcohalides. Davies et al. performed high-throughput screening of ternary metal chalcohalides to...
explore new photoactive semiconductors. A combination of element substitution and global structure searches was used after target systems were efficiently selected using compositional descriptors. As a result, Cd$_5$S$_4$Cl$_2$ and Cd$_4$SF$_6$, which have band gaps in the visible range, small electron and hole effective masses, and good optical absorption properties, were identified.

18-valence-electron systems. Zakutayev et al. and Gautier et al. performed first-principles screening of 18-valence-electron ABX ternary compounds using a combination of element substitution and evolutionary algorithm approaches, as mentioned in Sect. 4.1. Gautier et al. identified 54 previously unreported but thermodynamically stable compounds, 15 of which have been verified experimentally. Some of the predicted and characterized compounds are proposed to be potential transparent conductors, thermoelectric materials, and topological semimetals.

Ternary copper nitrides. Nitride semiconductors, particularly systems with three or more components, are much less explored than oxide semiconductors, even after the successful industrial application of GaN and its alloys. This is partly because of a difficulty in the synthesis of nitrides, so computational searches would be especially useful. Zakutayev et al. explored ternary Cu nitrides for photovoltaic absorber applications using first-principles calculations and an element substitution approach. They included thermodynamically metastable compounds in the target materials, motivated by their successful synthesis of the metastable CuNbN$_2$. As a result, Cu$_3$Ga$_2$N$_3$, Cu$_3$In$_2$N$_3$, and their alloys were identified as promising candidates in view of their band gaps, absorption coefficients, and effective masses.

Ternary zinc nitrides. Another example of the identification of nitride semiconductors is a search for novel ternary zinc nitride semiconductors, followed by experimental verification of the most promising compound, by Hinuma et al. Zn nitrides have valence band structures favorable for p-type doping, as mentioned in Sect. 4.1. Unfortunately, p-type doping of the binary zinc nitride Zn$_3$N$_2$ would be difficult, despite its promising valence band structure, because strong hole compensation is caused by N vacancies, as well as H and O impurities. In contrast, such compensation is likely to be well suppressed in some ternary systems, as described below.

Figure 28(a) shows ternary zinc nitrides identified by first-principles screening of approximately 600 candidates using a combination of element substitution and evolutionary algorithm approaches. These 21 systems are dynamically stable against lattice vibration and thermodynamically stable or slightly metastable (formation energies less than 50 meV/atom), and they have effective masses as small as or smaller than that of GaN for electrons, holes, or both. The identified nitrides fall into three categories: (i) previously reported Zn nitride semiconductors, (ii) Zn nitrides whose synthesis is reported but whose semiconductor properties are unexplored, and (iii) Zn nitrides unreported in the Inorganic Crystal Structure Database (ICSD).

The identification of group (i) validates the in silico screening methods, procedures, and criteria. Systems in groups (ii) and (iii) are of interest in light of their novelty as semiconductors. The authors focused on Cu$_2$Zn$_2$N$_2$ in group (iii) because of several promising features, as shown in Fig. 28(b): its earth-abundant constituent elements; direct-type band structure that is suitable for light emission and harvesting, specifically, the red light emission expected from the theoretical gap of 1.8 eV; the small electron and hole effective masses; and the bipolar dopability predicted by assessment of native point defects and dopants. This novel phase has been realized by high-pressure synthesis at 1200°C and 5.0 GPa for 1 h, which were predicted to be viable synthesis conditions from phase diagram calculations. The theoretically predicted crystal structure and optical properties, including red light emission [Fig. 28(b)], have been experimentally verified. The resultant experimental direct band gap of 1.9 eV is in good agreement with the predicted value of 1.8 eV. Experimental verification of the transport properties through film growth and carrier doping is awaited.

Summary and other examples of first-principles screening. These studies on TCOs, photoabsorbers, 18-electron ABX compounds, and ternary Cu and Zn nitrides demonstrate the power of first-principles screening for the identification of materials whose semiconducting properties or even synthesis have not been reported previously. Related examples of screening studies include searches for nonoxide p-type transparent conductors, where BP has been identified as a promising system, for perovskites as potential semiconductors and dielectrics, and for photocatalysts such as oxide and mixed-anion perovskites, ternary Sm(II) oxides, and ternary vanadates. Materials exploration using first-principles calculations has also been reviewed, for example, in Refs. 21 and 22.

Data-science-based approaches. In silico screening of candidate materials can be accelerated and enhanced by data science approaches. Efforts have been made to predict various properties, among which the band gap is particularly important for the exploration of novel semiconductors. Ward et al. reported band-gap prediction using the chemical compositions and information on the constituent elements of materials. After testing several machine learning algo-
rithms, they obtained mean absolute errors of as small as 0.06 eV in a 10-fold cross-validation test of a single model trained on the DFT results for 228,676 crystalline compounds selected from the OQMD.333) Dey et al. predicted the band gaps of over 200 new chalcopyrite compounds with as-yet-unreported chemistries using an ensemble data mining approach.365) Pilania et al. reported band-gap prediction for quinary double perovskites using kernel ridge regression.366) Such multicomponent systems are difficult to tackle using only first-principles calculations because a vast number of possible configurations of constituent elements exist even within a single crystal structure type.

Lee et al. took a pragmatic approach in which theoretical band gaps at the level of the $GW$ approximation were predicted from the results of computationally much less demanding PBE-GGA and mBJ-meta-GGA calculations, along with information on the constituent elements of the target materials.367) They obtained a band-gap prediction model with a root-mean-square error of 0.24 eV by learning 270 inorganic compounds using nonlinear support vector regression. This approach could eventually reach higher accuracy than machine learning using only the information on the chemical composition and constituent elements, but this advantage would be offset by the computational costs required for calculations at the LDA, GGA, or meta-GGA level.

In addition to the prediction of band gaps and other properties of semiconductors, machine learning has been used to identify materials that are as yet unreported but likely to be synthesized.368,369) Such approaches would be useful for accelerating the discovery of novel materials not only for semiconductor applications but also for diverse purposes.

5. Concluding remarks
A tremendous number of theoretical and computational studies of semiconductors have thus far been reported from both academia and industry, only some of which are covered in this review. Historically, first-principles approaches have been used to understand known semiconducting materials in terms of the fundamental, thermodynamic, and defect properties that are directly accessible by calculations. With the development of relevant theory and computational schemes, as well as computer performance, high-throughput screening for the exploration of novel materials can now be done in silico. More than ten million results of first-principles calculations are already available in public repositories, and searching for materials within these databases is a compelling approach. Another option is to explore materials with as-yet-unreported compositions and/or crystal structures. In both cases, reliable design principles, as well as accurate and efficient computational schemes, are key requirements for successful identification of target materials and functionalities. Theoretical calculations of band structures and defect properties are helpful for deriving such design principles, given that their systematic investigation and understanding are usually difficult to realize by experiment alone. Moreover, data-science-oriented approaches to issues in materials science and engineering are emerging. With the deployment of high-throughput computation and experimentation in addition to available databases, we expect a growing number of successful applications of such materials informatics approaches to practical development of materials.
Acknowledgments

The authors are grateful to Dr. Yoyo Hinuma for valuable discussions. This work was supported by the MEXT Elements Strategy Initiative to Form Core Research Center program, the Creation of Life Innovation Materials for Interdisciplinary and International Researcher Development project from MEXT, Grants-in-Aid for Scientific Research A (Grant No. 17H01318) and Young Scientists A (Grant No. 15H05541) from JSPS, PRESTO (Grant No. JPMJPR16N4), and the Support Program for Starting Up Innovation Hub MI² from JST, Japan.

1) S. M. Sze, Semiconductor Devices: Physics and Technology (Wiley, New York, 2002).
2) F. A. Ponce and D. P. Bour, Nature 386, 351 (1997).
3) M. A. Green and S. P. Bremner, Nat. Mater. 16, 23 (2017).
4) A. Polman, M. Knight, E. C. Garnett, B. Ehler, and W. C. Sinke, Science 352, 448 (2016).
5) M. A. Green, Y. Hishikawa, W. Warta, E. D. Dunlop, D. H. Levy, J. Hohlf-Ebinguer, and A. W. Ho-Baillie, Prog. Photovolt. 26, 688 (2017).
6) Transparent Electronics: From Synthesis to Applications, ed. A. Facchetti and T. J. Marks (Wiley, New York, 2010).
7) K. Nomura, H. Ohba, A. Takagi, T. Kamiya, M. Hirano, and H. Hosono, Nature 432, 485 (2004).
8) T. Kamiya, K. Nomura, and H. Hosono, Sci. Technol. Adv. Mater. 11, 044305 (2010).
9) E. Fortunato, P. Barquinha, and R. Martins, Adv. Mater. 24, 2945 (2012).
10) H. Liu, V. Avrutin, N. Izyumskaya, U. Özgür, and H. Morköç, Superlattices Microstruct. 48, 458 (2010).
11) H. Ohno, H. Menekata, T. Penney, s. von Molnár, and L. L. Chang, Phys. Rev. Lett. 66, 2664 (1992).
12) H. Ohno, A. Shen, F. Matsukura, A. Oiwa, A. Endo, S. Katsumoto, and Y. Iye, Appl. Phys. Lett. 69, 363 (1996).
13) T. Dief and H. Ohno, Rev. Mod. Phys. 86, 187 (2014).
14) B. J. Alder and T. E. Wainwright, J. Chem. Phys. 31, 459 (1959).
15) G. V. Lewis and C. R. A. Catlow, J. Phys. C 18, 11-09 (1985).
16) R. Catlow, R. Bell, F. Cora, S. A. French, B. Slater, and A. Sokol, Annu. Rep. Prog. Chem., Sect. A 101, 513 (2005).
17) J. Kohanoff, Electronic Structure Calculations for Solids and Molecules: Theory and Computational Methods (Cambridge University Press, New York, 2006).
18) R. M. Martin, Electronic Structure: Basic Theory and Practical Methods (Cambridge University Press, New York, 2008).
19) F. Giustino, Materials Modelling Using Density Functional Theory: Properties and Predictions (Oxford University Press, New York, 2014).
20) K. Lejaeghere, G. Bibhaimayer, T. Björkman, P. Blaha, S. Blügel, S. Blüm, K. Lejaeghere, G. Bihlmayer, T. Björkman, P. Blaha, S. Blügel, V. Blum, I. E. Castelli, F. Hüser, M. Pandey, H. Li, K. S. Thygesen, B. Seger, A. Jain, K. A. Persson, G. Ceder, and K. W. Jacobsen, Adv. Energy Mater. 5, 1400915 (2015).
21) F. Tran, S. Elsan, and P. Blaha, Phys. Rev. Mater. 2, 023802 (2018).
22) M. K. Y. Chan and G. Ceder, Phys. Rev. Lett. 105, 196403 (2010).
23) A. V. Krukau, O. A. Vydrov, A. F. Izmaylov, and G. E. Scuseria, J. Chem. Phys. 125, 224106 (2006).
24) J. Paier, M. Marsman, K. Hummer, G. Kresse, I. C. Gerber, and J. G. Angyán, J. Chem. Phys. 124, 154709 (2006).
25) Y. Matsushita, K. Nakamura, and A. Oshiyama, Phys. Rev. B 84, 075205 (2011).
26) D. M. Bylander and L. Kleinman, Phys. Rev. B 41, 7868 (1990).
27) R. Asahi, W. Mannhardt, and A. J. Freeman, Phys. Rev. B 59, 7486 (1999).
28) S. J. Clark and J. Robertson, Phys. Rev. B 82, 085208 (2010).
29) A. Seidl, A. Görling, P. Vogl, J. A. Majewski, and M. Levy, Phys. Rev. B 53, 3764 (1996).
30) J. P. Perdew, M. Ernzerhof, and K. Burke, J. Chem. Phys. 105, 9982 (1996).
31) C. Adamo and V. Barone, J. Chem. Phys. 110, 6158 (1999).
32) M. Ernzerhof and G. E. Scuseria, J. Chem. Phys. 110, 5029 (1999).
33) J. Heyd, G. E. Scuseria, and M. Ernzerhof, J. Chem. Phys. 118, 8207 (2003).
34) J. Heyd, G. E. Scuseria, and M. Ernzerhof, J. Chem. Phys. 124, 219906 (2006).
35) A. D. Becke, J. Chem. Phys. 98, 5648 (1993).
36) M. A. L. Marques, J. Vidal, M. J. T. Oliveira, L. Reining, and S. Botti, Phys. Rev. B 83, 035119 (2011).
37) A. Alkauskas, P. Brouqvit, and A. Pasquarello, Phys. Status Solidi B 248, 775 (2011).
38) F. Oba, A. Togo, I. Tanaka, J. Paier, and G. Kresse, Phys. Rev. B 77, 245202 (2008).
39) J. H. Skone, M. Govoni, and G. Galli, Phys. Rev. B 89, 195112 (2014).
40) T. Shimazaki and T. Nakajima, J. Chem. Phys. 141, 114109 (2014).
41) M. Gerosa, C. E. Bottani, L. Caramella, G. Onida, C. D. Valentin, and G. Pacchioni, Phys. Rev. B 91, 155201 (2015).
42) J. H. Skone, M. Govoni, and G. Galli, Phys. Rev. B 93, 235106 (2016).
43) Y. Hinuma, Y. Kumagai, I. Tanaka, and F. Oba, Phys. Rev. B 95, 075302 (2017).
44) M. Geoa, C. E. Bottani, C. D. Valentin, G. Onida, and G. Pacchioni, J. Phys.: Condens. Matter 30, 044003 (2018).
45) L. Hedin, Phys. Rev. 139, A796 (1965).
46) M. S. Hybertson and S. G. Louie, Phys. Rev. B 34, 5390 (1986).
47) M. Shishkin, M. Marsman, and G. Kresse, Phys. Rev. Lett. 99, 246403 (2007).
48) M. van Schilfgaarde, T. Kotani, and S. Faleev, Phys. Rev. Lett. 96, 226402 (2006).
49) F. Bruneval, N. Vast, and L. Reining, Phys. Rev. B 74, 045102 (2006).
50) R. Del Sole, L. Reining, and R. W. Godby, Phys. Rev. B 49, 8024 (1994).
51) A. Grünies, G. Kresse, Y. Hinuma, and F. Oba, Phys. Rev. Lett. 112, 096401 (2014).
52) D. C. Langreth and J. P. Perdew, Phys. Rev. B 15, 2884 (1977).
