Bandgap narrowing in zincblende III–V semiconductors: Finite-temperature full random-phase approximation and general analytical model

A. Schenk

AIP Advances 13, 070702 (2023)
https://doi.org/10.1063/5.0149190
Bandgap narrowing in zincblende III–V semiconductors: Finite-temperature full random-phase approximation and general analytical model

Cite as: AIP Advances 13, 070702 (2023); doi: 10.1063/5.0149190
Submitted: 10 March 2023 • Accepted: 23 June 2023 • Published Online: 12 July 2023

A. Schenk

AFFILIATIONS
Integrated Systems Laboratory, ETH Zurich, Gloriastr. 35, 8092 Zurich, Switzerland

Author to whom correspondence should be addressed: schenk@iis.ee.ethz.ch

ABSTRACT
The bandgap narrowing (BGN) in zincblende III–V semiconductors is calculated in a finite-temperature full Random-Phase Approximation (RPA) formalism based on an isotropic dispersion model. The cases of n-type and p-type quasi-neutral regions and the case of a neutral electron–hole plasma are elaborated for the technologically important materials GaAs, AlAs, InAs, GaP, InP, GaSb, InSb, zb-GaN, zb-InN, Al₀.₃Ga₀.₇As, GaAs₀.₅Sb₀.₅, InP₀.₅₃Sb₀.₄₇, InAs₀.₄₅P₀.₅₅, InAs₀.₄₂Sb₀.₅₈, InAs₀.₃₂Al₀.₆₈As, In₀.₄₉Ga₀.₅₁P, In₀.₅₇Ga₀.₄₃As, In₀.₅₃Ga₀.₄₇Sb, and zb-Ga₀.₅In₀.₅N (60 cases). In quasi-neutral regions, the correlation energy of the interaction between carriers and ionized dopants adds two terms to the total BGN. At low temperatures, inefficient screening makes the hole term dominant in n-type materials with a large ratio of the valence band to the conduction band (CB) density-of-states. The inclusion of the CB nonparabolicity is decisive here, as it prevents a diverging BGN at high concentrations. For all 60 cases, the BGN is evaluated in the temperature range from 0 to 500 K. A strong temperature dependence over the whole density range is observed in all direct n-type materials. Otherwise, the temperature dependence quickly ceases with increasing density. An analytical model of BGN without material-dependent free fit parameters is derived and compared with the full-RPA results.

I. INTRODUCTION
Bandgap narrowing (BGN) in semiconductor devices is a many-body effect that occurs in heavily doped regions and under strong optical or electrical excitation. In the first case, as in bipolar transistors (BTs) and field-effect transistors (FETs), the activated doping is compensated by a one-component plasma either in the emitter of a BT or in the source/drain of a FET. In the second case, a two-component plasma is generated, as in optically and electrically pumped laser diodes. Device characteristics can be altered by BGN, examples being the reduced current gain in a BT due to the increased emitter minority concentration or the density dependence of the laser wavelength. Modeling of BGN is facilitated by the condition of charge neutrality, which holds in quasi-neutral regions (negligible field strength) and in a neutral plasma (background doping is small compared to the plasma density). The physical origin of BGN is the many-body Coulomb interaction between free charge carriers as well as between charge carriers and ionized dopants. The resulting total self-energy shifts the band edges such that the bandgap shrinks. Four contributions have to be considered in a neutral electron–hole (eh) plasma—the exchange and correlation energies of electrons and holes, whereas five components contribute in quasi-neutral regions—the exchange energy of the majority carriers, the free-carrier correlation energies of the majority and minority carriers, and the ionic correlation energies of the majority and minority carriers. In this paper, all self-energies are calculated in the finite-temperature full Random-Phase Approximation (RPA). RPA is increasingly accurate at higher densities where BGN becomes significant for device operation, and it is the only method to cover the whole temperature range. In n-type III–V zincblende semiconductors with small electron effective masses, the temperature dependence becomes extreme even at low densities. This is because
the ionic correlation energy of holes is the largest contributor to the total BGN here. The shift of the minority carrier band is a consequence of the lowering of the energy of the generated/annihilated minority carrier due to Coulomb attraction by the mobile majority carriers and Coulomb repulsion by the sub-system of immobile ionized dopants.

The basic RPA theory as well as results for the eh-plasma in various materials can be found in Ref. 1. The extrinsic semiconductor with a random distribution of dopants was treated in Ref. 4 for $T = 0$ K. Earlier calculations of the self-energy of the carrier-dopant interaction in the $T = 0$ limit were performed with the Hartree–Fock variational method for donors distributed on a regular sub-lattice applying the single plasmon pole (SPP) approximation and with second-order perturbation theory for randomly distributed dopants based on RPA screening. The multi-valley case had been discussed for both dopant arrangements in Refs. 12–14.

Measurements of BGN are accompanied by an intricate analysis. Three types of experiments have been developed and mostly applied to silicon due to its outstanding role in technology: Absorption experiments, photoluminescence (PL) experiments, and electrical measurements of BTs. In absorption experiments, phonon-assisted transitions are present besides the band-band transitions; therefore, the phonon energies must be known as well as the exact position of the Fermi energy. The evaluation of PL experiments is difficult because of the actual initial states and the weak intensity of the spectrum. The electrical method requires the measurement of collector current in the BT as a function of the emitter-base voltage.

Appendix C presents the derivation of the analytical formulas of the BGN for n-type materials with small electron effective masses, the hole masses are calculated from the effective mass. It is shown how the hole masses are calculated from the effective mass. It is shown how the latter effect weakens the BGN in n-type materials with small electron effective masses, preventing the bandgap from shrinking to zero in the relevant density range. Conclusions are given in Sec. V. Appendix A contains the complete presentation of the full-RPA BGN in all 20 materials as a function of density in the temperature range from 0 to 500 K. Each figure comprises quasi-neutral n-type, quasi-neutral p-type, and neutral eh-plasma and also displays the corresponding analytical curves. The density limits can be inferred from the endpoint of the density axis. Appendix B provides the derivation of the analytical expressions of the free-carrier correlation energies in quasi-neutral regions and in the neutral eh-plasma, whereas Appendix C presents the derivation of the analytical formulas of the ion-carrier correlation energies in quasi-neutral regions including the effect of nonparabolicity in n-type material, and the temperature dependence of the donor-hole correlation energy including the effect of nonparabolicity on this temperature dependence.

II. RPA THEORY AND BAND STRUCTURE MODEL

The full-RPA BGN is calculated by the quasi-particle shift (QPS) $\Delta_0(k)$ ($a = e$ for electrons, $a = h$ for holes), which is given by the difference between interacting and free dispersion,

$$\Delta_a(k) = E_a(k) - E_a^0(k),$$

and equals the real part of the self-energy $\Sigma_a$,

$$\Delta_a(k) = \text{Re} \Sigma_a(k, E_0(k) + i0^+).$$

It consists of the three terms

$$\Delta_a(k) = \Delta_a^0(k) + \Delta_a^e(k) + \Delta_a^h(k),$$

the unscreened exchange energy $\Delta_a^0(k)$, the correlation energy of free carriers $\Delta_a^e(k)$, and the correlation energy of the interaction between the carriers and ionized dopants $\Delta_a^h(k)$. A random arrangement on regular lattice sites is assumed for the dopants. Since the
displacement of the QPS in the energy interval between band edge and Fermi energy is rather flat for the relevant densities, the dispersive QPS $\Delta_\gamma(k)$ can be replaced by a rigid shift $\Delta_\gamma$. Then a self-consistent solution of the problem is not necessary because the rigid shift drops out in the energy difference $E_\gamma(k + q) - E_\gamma(k)$, which becomes $E_\gamma^0(k + q) - E_\gamma^0(k)$, i.e., fully determined by the free dispersion. In the distribution functions, $\Delta_\gamma$ is fixed by the given density in the quasi-neutral region or by the plasma density in the case of the neutral e-h-plasma. The condition that the QPS density should not change in first order with respect to $\Delta_\gamma(k) - \Delta_\gamma$ results in
\[ \Delta_\gamma = -\sum_k \frac{\partial f_\gamma(k)}{\partial \epsilon_\gamma} \Delta_\gamma(k) - \frac{1}{\sum_k \partial f_\gamma(k)/\partial \epsilon_\gamma}, \quad (4) \]
with rigidly shifted bands $E_\gamma(k) = E_\gamma^0(k) + \Delta_\gamma$ and Fermi–Dirac functions $f_\gamma(k)$ depending on shifted chemical potentials $\mu_\gamma$,
\[ f_\gamma(k) = \frac{\sum \epsilon a e^{-\epsilon k_n/\beta \Lambda_n}}{\sum \epsilon a e^{-\epsilon k_n/\beta \Lambda_n}}, \quad (5) \]
The first derivative of the Fermi–Dirac function in (4) plays the role of a weight that filters out energies near the Fermi energy in the low-T/high-density limit and energies close to the band edge in the high-T/low-density limit. The dispersive QPS (3) is given explicitly in Refs. 1, 4, and 41. Insertion into Eq. (4) results in the following consolidated forms of the three contributors to the rigid shift:
\[ \Delta_\gamma^x = \frac{\gamma e^2}{4\epsilon_0 e \Lambda_n} F_{1/2}(\beta \Lambda_n), \quad (6) \]
\[ \Delta_\gamma^\delta = \frac{1}{2} \sum_{q} \left[ \epsilon(q, \Omega_n) - 1 \right] \frac{\partial \epsilon(q, \Omega_n)}{\partial n_\gamma}, \quad (7) \]
\[ \Delta_\gamma^\epsilon = \frac{m_i}{2} \left( \frac{\partial \epsilon(q, \Omega_n)}{\partial n_\gamma} \right)^{-1} \int \frac{d^3 q}{(2\pi)^3} \frac{\partial \epsilon(q, 0)}{\partial n_\gamma} \quad (8) \]
In the correlation energies (7) and (8), $\epsilon(q, \Omega_n) = e^2/(\epsilon_0 e^2 q^2)$ is the Fourier transform of the bare Coulomb potential with the static permittivity $\epsilon_0$, $\Omega_n = 2\pi n_\gamma$ denotes the Matsubara frequency ($n_\gamma$ integer) and $n_\gamma$ the concentration of electrically active dopants. The correlation energies are governed by the RPA dielectric function,
\[ \epsilon(q, \Omega_n) = 1 - v(q) \sum_{k} g_{\gamma k} \frac{f_\gamma(k + q) - f_\gamma(k)}{E_\gamma(k + q) - E_\gamma(k) - \hbar \Omega_n}, \quad (9) \]
where the energy difference $E_\gamma(k + q) - E_\gamma(k)$ in the denominator is replaced by $E_\gamma^0(k + q) - E_\gamma^0(k)$ (rigid shift).

In the parabolic band model, the QP density $n_\gamma$ in the rigid shift approximation is defined by
\[ n_\gamma = \sum_k g_{\gamma k} f_\gamma(k) = g_{\gamma} \Lambda_n^3 F_{1/2}(\beta \Lambda_n), \quad (10) \]
with the Fermi integral
\[ F_{1/2} = \frac{(2\pi)^{3/2}}{(2\pi)^3} \left( \frac{2\pi \hbar^2 k^2}{m_e} \right)^{1/2} \]
the thermal wavelength $\Lambda_n = (2\pi \hbar^2 / m_e)^{1/2}$, the inverse thermal energy $\beta = 1/k_B T$, and the Fermi energy $\Lambda_n^\delta = \mu_\gamma - \Delta_\gamma$. The simple form of the exchange energy (6)—just a single integral (the Fermi integral)—is bound to parabolic bands $E_\gamma^0(k) = \hbar^2 k^2/2m_e$. In contrast, the rigid-shift expressions of the correlation energies (7) and (8) are obtained by employing (4) regardless of the specific form of the band dispersion. The exchange-correlation energy of the electron and hole plasmas is rather insensitive to band structure details as a consequence of a compensation effect.

The nonparabolicity of the CB is crucial for the ionic correlation energy $\Delta_\gamma'^{\text{ionic}}$ of certain n-type III–V materials in the low-T/high-density regime, since, due to the small n-density-of-states (DOS), the Fermi level moves deep into the CB. The electron density $n_e$ and the dielectric function $\epsilon(q, \Omega_n)$ in (8) are then calculated with the nonparabolic free dispersion,
\[ E_{\text{ion}}^{0}(q) = \frac{1}{2v} \left( \sqrt{2\chi \hbar^2 q^2/m_e + 1} - 1 \right), \quad (11) \]
where $\chi$ denotes the nonparabolicity parameter (the common symbol $\alpha$ is used for the mass ratios $\alpha_{\text{e/h}}$ in this paper). In addition, for the shift of the majority band edge in n-type material, $\Delta_{\gamma'^{\text{ionic}}}$, and for the shift of the minority band edge in p-type material, $\Delta_{\gamma'^{\text{ionic}}}$, the nonparabolic dispersion (11) is applied for electrons.

Figure 1 demonstrates the nonparabolic effect on $\Delta_{\gamma'^{\text{ionic}}}$, which dominates the total BGN in all n-type materials with $m_e \ll m_h$ ($a_0 \ll 1$). It is particularly strong for the small-gap materials with strong CB-valence band (VB) coupling (Nos. 1–5 in Table 1) but is also important for all other direct materials. The only exception is the indirect materials AIAs and GaP, where $m_e \approx m_h$ ($a_{\text{e/h}} \approx 0.5$).

As the evaluation of (7) comes down to a sevenfold integration and that of (8) to a sixfold integration, an isotropic dispersion model is inevitable. A full numerical treatment based on a more elaborate band structure model is out of reach even in the rigid-shift approximation of BGN. Therefore, the real DOS has to be approximated by an isotropic DOS, and its range of validity must be carefully determined, which sets upper bounds on the allowed densities. The isotropic band structure model is obtained as follows: for all direct zincblende materials, the $\Gamma$-valley is assumed isotropic with
an effective mass $m_e$ taken from Refs. 45 and 44. For the two indirect materials, AlAs and GaP, the anisotropic valleys are mapped to isotropic ones with an effective DOS mass $m_e = (m^2 \gamma^2) / 12$. The valley degeneracy is included in the total degeneracy factor $g_e$. Heavy and light holes are treated as independent isotropic VB valleys, and the effective DOS mass is obtained as $m_h = [(m^2 \gamma^2 + m^2 \gamma^2) / 12] / 12$, which results in valley degeneracies of $2$ for all materials included in the valley degeneracy factors of $2$ for all materials treated in this paper. The split-off band is neglected. In some cases, the values of $m_h$ and $m_{hh}$ were taken from Ref. 44 if based on experimental data. For most materials, however, the Luttinger parameters $\gamma$ of Refs. 45 and 44, and the electronic structure model. In all other cases and, in general, for the neutral material, the limits of activated and intensive research.

The amphoteric nature of these defects leads to a Fermi level pinning in the gap or in the CB, independent of the type or doping level, which explains the saturation of the electrically active doping. Experimental data for the limits of activated doping are only available for some of the 20 materials treated in this paper: InSb, InAs, GaSb, and zb-GaN. These limits were adopted for the density limits since they are, in any case, smaller than the limits imposed by the band structure model. In all other cases and, in general, for the neutral and doped material, the limits of activated and intensive research.

An important aspect in the context of BGN is the possible maximum density. Applying the theory to unreasonable high densities can readily shrink the gap of certain n-type materials to zero. In quasi-neutral regions, the carrier density is equal to the electrically active doping. The intrinsic limitations of the latter have been subject to intensive research.

The amphoteric nature of these defects is a property of the material and not a chemical or electronic feature of the dopant (e.g., formation of inactive complexes or charge compensation due to the amphoteric nature of shallow dopants). Process-related localized defects stabilize the Fermi energy at the Fermi-level stabilization energy $\sim 4.9$ eV below the vacuum level in many III–V materials. The amphoteric nature of these defects leads to a Fermi level pinning in the gap or in the CB, independent of the type or doping level, which explains the saturation of the electrically active doping. Experimental data for the limits of activated doping are only available for some of the 20 materials treated in this paper: InSb, InAs, GaSb, and zb-GaN. These limits were adopted for the density limits since they are, in any case, smaller than the limits imposed by the band structure model. In all other cases and, in general, for the neutral and doped material, the limits of activated and intensive research.

Finally, the value $\gamma$ of Refs. 44, 45, and 69. The nonparabolic free-dislap
Note that $\alpha$ to transform excitonic quantities back to electron/hole quantities.

Fig. 4 for 20 K. Two densities have been chosen for all 60 cases: $10^{18}$ and $10^{19}$ cm$^{-3}$. In some cases, the higher density already exceeds the allowed limit, as discussed earlier; hence, the corresponding values were omitted.

Looking at the room-temperature results first, n-type, p-type, and eh-plasma differ only a little at the lower density (filled circles). The average trend is an increase in BGN from $\sim 15$ to $\sim 50$ meV. At the higher density, p-type and eh-plasma still behave similarly with an average increase from $\sim 30$ to $\sim 100$ meV, but the BGN in direct n-type material is significantly larger due to the small electron mass leading to reduced screening. One observes an average increase from $\sim 40$ to $\sim 120$ meV. The indirect materials AlAs and GaP make an exception since their CB-DOS and VB-DOS are comparable in strength. Therefore, the BGN in n-type material is almost identical to the one in p-type material here. The trends closely follow the trends of the DOS effective masses (see Fig. 2). The maxima produced by the nitrides are reflected in the BGN curves, albeit scaled-down a lot.

Turning to the low-temperature case of 20 K in Fig. 4, the behavior of p-type material and eh-plasma is not only almost identical at the lower density of $10^{19}$ cm$^{-3}$ but also very similar to the one at room temperature. To a lesser extent, this also holds at a

### III. SUMMARY OF FULL-RPA RESULTS

The total full-RPA BGN as a function of material, i.e., in dependence on the increasing bandgap, is shown in Fig. 3 for 300 K and in Fig. 4 for 20 K. Two densities have been chosen for all 60 cases: $10^{18}$ and $10^{19}$ cm$^{-3}$. It is related to their multi-valley nature. The nitrides are strong outliers from $\sim 9$ to $\sim 30$ to $\sim 50$ meV, but the BGN in direct n-type material is significantly larger due to the small electron mass leading to reduced screening. One observes an average increase from $\sim 40$ to $\sim 120$ meV. The indirect materials AlAs and GaP make an exception since their CB-DOS and VB-DOS are comparable in strength. Therefore, the BGN in n-type material is almost identical to the one in p-type material here. The trends closely follow the trends of the DOS effective masses (see Fig. 2). The maxima produced by the nitrides are reflected in the BGN curves, albeit scaled-down a lot.

Turning to the low-temperature case of 20 K in Fig. 4, the behavior of p-type material and eh-plasma is not only almost identical at the lower density of $10^{19}$ cm$^{-3}$ but also very similar to the one at room temperature. To a lesser extent, this also holds at a
higher density of $10^{19}$ cm$^{-3}$. The behavior of n-type material, however, differs drastically from that at 300 K. At lower densities, the BGN exhibits a slightly decreasing tendency up to the indirect materials, followed by an increase for the wide-gap nitrides. Nevertheless, the BGN is relatively independent of material, taking values between $\sim$50 and $\sim$80 meV. In direct n-type materials, the BGN becomes dramatically increased as a consequence of the weak screening of the hole-ion interaction by majority carriers (electrons), an effect that gains importance with decreasing temperature (compare Fig. 5 and the detailed BGN representations in Appendix A). The analytical explanation for this temperature dependence will be given in Sec. IV. One can see an increasing tendency of the BGN from $\sim$130 to $\sim$200 meV, only interrupted by the indirect materials AlAs and GaP, where again the values are the same as in p-type materials due to the similar DOS size.

In Figs. 5 and 6, the BGN of the small-gap direct material InSb is compared with that of the wide-gap indirect material AlAs. The density limit for n-type InSb in the left part of Fig. 5 stems from the measured maximum of activated doping ($7.5 \times 10^{18}$ cm$^{-3}$) and results in a BGN that already slightly exceeds the value of the zero-temperature gap. Interestingly, for a limit of $4.5 \times 10^{18}$ cm$^{-3}$, as found in a standard low-pressure growth process, the gap would exactly shrink to zero at $T = 0$ K. In all other cases in Figs. 5 and 6, the upper bounds of the density are imposed by the band structure model. The striking difference between InSb and AlAs is the behavior of n-type materials. Whereas in AlAs, the BGN is similar to that in p-type and eh-plasma (note the similar BGN scale and the almost identical temperature dependence, which ceases above $10^{12}$ cm$^{-3}$), the BGN in n-type InSb exhibits a strong temperature dependence that becomes even more pronounced at higher densities. At $T = 0$ K and $10^{18}$ cm$^{-3}$, the BGN is 15 times larger than in p-type material. As outlined in Ref. 7, this is due to the small electron effective mass and the resulting weak screening of the hole-ion correlation energy. It will be shown in Sec. IV that the strong temperature dependence arises from the occupation probability of the minority (hole) band.

Appendix A comprises the presentation of BGN for all 20 materials (n-type and p-type) and for the neutral eh-plasma (in the intrinsic form of these materials) as a continuous function of density from $10^{15.5}$ cm$^{-3}$ to the maximum at the temperatures $T = 0$, 20, 77, 300, and 500 K. The full-RPA curves for $T = 0$ K (black solid) start at the Mott density $n_{M} = (p_{M}/a_{ex}a_{b})^{3}$, if $n_{M} > 10^{15.5}$ cm$^{-3}$, using $p_{M} = 0.296$ for the Mott criterion. In all p-type nitrides, $n_{h,M} > n_{\text{max}}$; hence, the $T = 0$ curves are completely missing. Reasons could be (i) the Mott parameter, (ii) the size of the Luttinger parameters from which the VB DOS masses are derived, (iii) the isotropization Eq. (13), or (iv) the size of the spin–orbit energy $E_{so}$ that determines $n_{\text{max}}$. The latter is rather small [(6–19) meV in the ordering of the nitrides given in Table I] and leaves only a narrow energy interval compliant with the 2-valley VB model. The ratio $n_{h,M}/n_{\text{max}}$ increases from 5 to 12. Actually, (i) and (ii) are not independent of each other since $n_{h,M} \sim (p_{M}m_{h})^{3}$. As the Mott parameter is related to the lattice constant, it decreases in zb-InN toward zb-AlN, which could partly explain the discrepancy. The analytical $T = 0$ curves are displayed over the whole density range, neglecting carrier freeze-out.

Figure 7 compares the available experimental data with the full-RPA results. The left panel shows PL data and hetero-structure BT data for p-type GaAs. They fairly match the theoretical curves. In the case of p-type GaSb (right panel), the measured BGN by PL is systematically larger than the RPA one. The opposite is found for n-type InP (absorption data) in the high-density range. Absorption
measurements of the BGN in n-type InAs at room temperature resulted in a sharp peak at intermediate densities. The authors interpreted this peculiar behavior by a large screening length due to the strong correlation of the tellurium dopants (“complete ordering”) when their concentration falls into the interval \((2–4) \times 10^{18} \text{ cm}^{-3}\). Interestingly, full-RPA results (based on the assumption of a random impurity distribution) are in reasonable agreement both at low and very high densities.

**IV. ANALYTICAL MODEL WITHOUT MATERIAL-DEPENDENT FREE FIT PARAMETERS**

In this section, a fully analytical model of BGN as a function of density and temperature in terms of the material and band structure parameters of Table I is presented. The goal is not the highest accuracy but to avoid free fit parameters for each case (as performed in Refs. [77] and [78] for \(T = 0 \text{ K}\) and local-density approximation (LDA)). This facilitates physical insight and the possibility to quickly calculate BGN for other materials or ternary compositions, as long as they obey (or can be approximated to) the isotropic band structure model consisting of one CB valley and two VB valleys. Wurtzite materials are excluded because of their complicated VB structure caused by the crystal-field splitting.

The general approach is to combine the \(T = 0\)-limit \(\Delta E_{g,T=0}\) with the high-temperature (Debye) limit \(\Delta E_{g,T\to\infty}\) by a Padé approximation using a universal switching function, \(\lambda_\beta\). Therefore, \(\Delta E_{g,T=0}\) is replaced by \(\Delta E_{g,0}(T)\) with the analytical model for \(\Delta E_{g,0}(T)\) (see below), but the \(T = 0\)-limits in the case of all other energy contributions. The total BGN is, therefore, calculated by

\[
\Delta E_g(n, T) = \frac{1 + U(n, T)}{1 + \Delta E_{g,T=0}(n, T)} \Delta E_{g,0}(n, T). \tag{15}
\]

Here, \(n\) equals the concentration \(n_i\) of activated dopants in quasi-neutral regions or the plasma density \(n_e = n_0\) of the neutral plasma, respectively. Note that BGN is defined as a positive quantity, i.e., the negative sum of all QPS contributions (6)–(8). The Debye limits read

\[
\Delta E_{g,T\to\infty}(n, T) = \begin{cases} 
8\pi n_0^2 \alpha_s/\delta_0 + 6\sqrt{2\pi n_0^3} & n \to n_i, \quad n - \text{type}, \\
8\pi n_0^2 \alpha_s/\delta_0 + 6\sqrt{2\pi n_0^3} & n \to n_i, \quad p - \text{type}, \\
8\pi n_0^2 (\alpha_s/g_0 + \alpha_0/g_0) + 8\sqrt{2\pi n_0^3} & n \to n_i, \quad \text{eh - pl},
\end{cases} \tag{16}
\]

The expression of the exchange energy \(\Delta E_{g,0}^{1}(e,h)\) is well known, and numerous attempts have been made to fit power laws in terms of density to the RPA free-carrier correlation energies of the electron and hole plasmas \((\Delta E_{g,0}^{1}(e,h))\) (reviewed in Ref. 1). In Ref. 7, the ionic correlation energies \(\Delta E_{g,0}^{1}(e,h)\) at \(T = 0 \text{ K}\) were derived for n- and p-doped quasi-neutral regions of InGaAs, with a first estimate of how strongly nonparabolicity impacts \(\Delta E_{g,0}^{1}(e,h)\). In this paper, all terms are calculated analytically starting from Eqs. (7) and (8) using the RPA dielectric function. A central role in the multiple integrations of complicated functions is played by the length

\[
s_b = \frac{\delta_0}{\pi n_0 q_{e,b}}, \tag{18}
\]

and “\(b\)” is the index of the majority carrier band. Hence, \(s_b\) is proportional to the Fermi wavelength \(\lambda_{eb}\) in units of the excitonic Bohr radius, \(s_b = g_b/(2\pi^2 n_0)\lambda_{eb}\). The assumption \(s_b \ll 1\) is sometimes used in this paper to simplify the integrands. Since \(1/q_{e,b}\) is normalized by the excitonic Bohr radius \(a_{eb}\), this assumption works better as the Fermi level moves deeper into the majority band (increasing density, weak DOS) and the excitonic Bohr radius is large, i.e., in the case of direct n-type III–Vs with small electron effective mass \(m_e\). On the contrary, if \(s_b\) is large and also \(m_e\), as in the case of n-type AlAs and GaP, the assumption \(s_b \ll 1\) is rather poor, as \(s_b = 1\) already corresponds to a density of \(\sim 10^{21} \text{ cm}^{-3}\). This finally shows up in a worse agreement between analytical and numerical \(\Delta E_g(n, T)\)-curves. Therefore, also \(s_b \gg 1\) is tested and utilized if more appropriate. The analytical form of BGN is first given in terms of the length \(s_b\), which is more compact, and then as a function of the normalized density for more convenient use.

**A. Quasi-neutral regions**

In quasi-neutral regions, \(\Delta E_{g,0}(n, T)\) reads in terms of the normalized length \(s_b\)

\[
\Delta E_{g,0}(s_b, T) = \frac{2g_b}{\pi^2 a_{eb}} + \frac{2\sqrt{g_b}}{3\pi^3 a_{0}} \ln \left(1 + \sqrt{2/s_b}\right) + \frac{\delta_0}{\sqrt{3\pi^2 a_{eb}^3 s_b}} + \frac{\delta_0 g_{eb} + \psi_{pe}(s_b)}{3\pi^3 a_{eb}^3 \sqrt{2s_b}} \tag{20}
\]

where the first term is the exchange energy and the second term is the sum of all correlation energies. Nonparabolicity is negligible in the low-density range.

\[
\Delta E_{g,0}(n, T) = \begin{cases} 
-\Delta_e^e - \Delta_e^h - \Delta_e^i - \Delta_e^0(T) & n \to n_i, \\
-\Delta_h^e - \Delta_h^h - \Delta_h^i - \Delta_h^0 & n \to n_i, \\
-\Delta_i^e - \Delta_i^h - \Delta_i^0 & n \to n_i, \quad \text{eh - pl},
\end{cases} \tag{17}
\]
with

\[ \varphi_{np}(s_e) = \sqrt{1 + \frac{4ye_e^2}{\pi^2 a_e^2 s_e^2}}, \tag{21} \]

\[ \varphi_{\text{f}}(s_e) = 1 - \frac{2}{\pi} \arctan \left( \frac{2\pi a_e^2 s_e}{\beta a_e \varphi_{np}(s_e)} \right)^{\frac{3}{2}}. \tag{22} \]

In Eq. (20), “b” is the index of the majority carrier band (b = e for n-type, b = h for p-type material), and “a” is the index of the minority carrier band (a = h for n-type, a = e for p-type material). The ordering of the five terms is the same as in Eq. (17). Note that the third term, the minority-carrier free-carrier correlation energy, only contains the index of the majority carrier band (b), since minority carriers do not contribute to screening. Nevertheless, the effective mass of the minority carriers finally comes into play after multiplication with the normalization energy (excitonic Rydberg).

Equation (21) is the nonparabolicity correction that only applies to n-type material. Note that it increases the electron-ion correlation energy (fourth term) but decreases the hole-ion correlation energy (fifth term) with different powers. Equation (22) describes the reduction of the hole-ion correlation energy with increasing temperature at intermediate and high densities. It also applies to n-type material only and is strongly affected by nonparabolicity (note the occurrence of \( \varphi_{np} \) in the argument of \( \text{arctan} \ldots \)). The zero-temperature limit (\( \beta \to \infty \)) is \( \varphi_{\text{f}} = 1 \).

Figure 8 shows the ability of this model to reproduce the \( T \)-dependence of the hole-ion correlation energy for all densities >10^16.5 cm^-3 over the whole temperature range. By way of example, the cases InSb (small gap), InGaAs (intermediate gap), and zb-GaN (wide gap) have been chosen. One observes a decreasing \( T \)-dependence and a changing shape of the \( n \)-dependence with a rising gap.

From Eq. (8), one can see that the temperature dependence has two origins: (i) the distribution function of minority carriers in the factor \((\partial n_{i}/\partial \varepsilon)_{n}^{-1}\) and (ii) the occupation probability of the majority carriers contained in the dielectric function (9). From a second-order Taylor expansion of the latter by 1/\( \beta \) around the Fermi energy, one obtains a quadratic \( T \)-dependence which, however, is weak and vanishes at high densities. It is negligible compared to (i), which results in the term (22). The derivation of this term is given in Appendix C.

Replacing \( n_{b} \) by the normalized density \( n \) using Eqs. (18) and (19), i.e., \( n_{b} = g_{h}^{1/3}/\{a_{h}(6\pi^{2}n^{3/2})\} \), one obtains

\[ \Delta E_{np}(n, T) = \frac{48n}{\pi g_{e}} \left( \frac{3\pi^{2}}{2} \right)^{1/3} \left( \frac{3n}{4\pi} \right)^{1/6} \left( \delta_{nh} + \varphi_{np}(n) \delta_{he} \right) \]

\[ + \frac{2}{\pi} \left( \frac{3n}{4\pi} \right)^{1/6} \left( \delta_{nh} + \varphi_{\text{f}}(n) \delta_{he} \right), \tag{23} \]

with

\[ \varphi_{np}(n) = \sqrt{1 + \frac{4ye_e^2}{\pi^2 a_e^2 s_e^2}}, \tag{24} \]

\[ \varphi_{\text{f}}(n) = 1 - \frac{2}{\pi} \arctan \left( \frac{2\pi a_e^2 s_e}{\beta a_e \varphi_{np}(n)} \right)^{\frac{3}{2}}. \tag{25} \]

In Eqs. (23)–(25), \( n \) represents the density \( n_{i} \) of ionized dopants (normalized by \( a_{e}^{-3} \)). Note that BGN is normalized by the excitonic Rydberg energy \( E_{R} \) as the thermal energy \( 1/\beta \) and the inverse of the nonparabolicity parameter \( 1/\gamma_{e} \).

### B. Neutral electron-hole plasma

In the case of the neutral eh-plasma, both exchange energies are present, but the carrier-ion correlation energies are absent [see Eq. (17)]. The analytical model for \( T = 0 \) K and parabolic bands reads as a function of the normalized lengths \( s_{eh} \),

\[ \Delta E_{eh}(s_e, s_h, 0 \text{K}) = \frac{2}{\pi} \left( \frac{g_{e}}{a_{e} s_{e}} + \frac{g_{h}}{a_{h} s_{h}} \right) + \frac{2e_{e}}{5\pi} \ln \left( 1 + \frac{8\alpha_{e}}{3} \right)^{\frac{1}{2}} \frac{1}{s_{e}^{1/4}} \]

\[ + \frac{2e_{h}}{3} \left( \Theta(0.85 - \alpha_{e}) \ln \left( 1 + \frac{8\alpha_{h}}{3} \right)^{\frac{1}{2}} \frac{1}{s_{h}^{1/4}} + \Theta(\alpha_{e} - 0.85) \ln \left( 1 + \frac{2\alpha_{h}}{\alpha_{e} - \alpha_{h}} \right) \right), \tag{26} \]

with

\[ \alpha_{he} = \frac{\alpha_{h}}{1 + \left( \frac{\alpha_{h}}{\alpha_{e}} \right)^{3/2}}. \tag{27} \]

The first term in (26) is the sum of the exchange energies; the second and third terms are the correlation energies of electrons and holes, respectively. These terms differ from the corresponding majority-
carrier correlation energy of the one-component plasma [second term in Eq. (20)] because of the stronger screening by the two-component plasma. The derivation of the correlation terms with detailed explanations is presented in Appendix B.

Inserting \( s_{(e,h)} = \frac{g_e^{1/3}}{\alpha} \left( \frac{e^2}{\pi} \right) \left( \frac{6n^2}{n} \right)^{1/3} \), one obtains the model as function of plasma density \( n \),

\[
\Delta E_{g,0}(n, 0K) = \left( \frac{1}{\alpha_e} + \frac{1}{\alpha_h} \right) \left( \frac{48n}{5\pi} \right)^{1/3} + 2\frac{g}{5\pi} \ln \left[ 1 + \frac{2\alpha_e (\pi^2 n)^{1/3}}{2\alpha h} \right] + 2\frac{\alpha h}{\pi} \left[ 0.85 - \alpha_e \right] \ln \left[ 1 + \frac{2\alpha_e (\pi^2 n)^{1/3}}{2\alpha h} \right] + \Theta(\alpha_e - 0.85) \ln \left( 1 + \frac{\alpha_e (6n^2)^{1/6}}{\alpha h} \right). \tag{28}
\]

In Fig. 9, the numerical RPA results of the band edge shifts induced by the free-carrier correlation in InGaAs at \( T = 0 \) K are compared with the outcomes of the analytical models. The freeze-out of carriers was disregarded to display the whole density range. The agreement is excellent for the shift of the CB edge and reasonable for the shift of the VB edge. This difference in accuracy is due to \( s_e \ll s_h \) \( (\alpha_e \gg \alpha_h) \). Figure 9 also shows that the correlation energies in a neutral eh-plasma are very similar to the free-carrier correlation energies of the majority carriers in quasi-neutral regions. Size and slope are only slightly larger, which can be attributed to the stronger screening in a two-component plasma.

The comparison of the analytical BGN model with the numerical full-RPA is presented in Appendix A for all 60 cases (Figs. 10–29). In p-type material and in the eh-plasma, the temperature dependence is only significant at intermediate densities. For better visibility, the analytical curves were plotted just for 0 K and room temperature here. The Padé approximation leads to a relatively large deviation in the transition region, \( n \sim n_c \), which is unavoidable. The density and temperature dependences are correctly reproduced in all
60 cases. Very good agreement is found for the direct n-type materials, except for InAsSb at high densities because of its extreme nonparabolicity. Larger deviations occur for the indirect materials AlAs and GaP, as $s_e << 1$ is a poor assumption here. For many p-type materials, the accuracy is satisfactory as well; exceptions are InAsSb, InSb, and the nitrides. For the neutral eh-plasma, the analytical model works best at high densities but deviates markedly at intermediate densities due to the error caused by the Padé interpolation.
FIG. 19. Total BGN in InAs<sub>0.4</sub>P<sub>0.6</sub>. Densities limited by \( E_L - E_\Gamma \)-separation (left and right) and \( E_\Gamma - E_{so} \)-separation (middle).

FIG. 20. Total BGN in zb-GaIn<sub>0.5</sub>N<sub>0.5</sub>. Densities limited by \( E_X - E_\Gamma \)-separation (left) and \( E_\Gamma - E_{so} \)-separation (middle and right).

FIG. 21. Total BGN in InP. Densities limited by \( E_L - E_\Gamma \)-separation (left and right) and \( E_\Gamma - E_{so} \)-separation (middle). Experimental data on the limits of activatable n-doping\textsuperscript{57,58} cover the range (5.5 \times 10^{19} \text{–} 1.76 \times 10^{20}) \text{ cm}^{-3}.

FIG. 22. Total BGN in In<sub>0.52</sub>Al<sub>0.48</sub>As. Densities limited by \( E_X - E_\Gamma \)-separation (left and right) and \( E_\Gamma - E_{so} \)-separation (middle).

FIG. 23. Total BGN in GaAs. Densities limited by activatable n-doping\textsuperscript{59} (left), activatable p-doping\textsuperscript{59} (middle), and \( E_L - E_\Gamma \)-separation (right).

FIG. 24. Total BGN in Al<sub>0.3</sub>Ga<sub>0.7</sub>As. Densities limited by \( E_L - E_\Gamma \)-separation (left and right) and \( E_\Gamma - E_{so} \)-separation (middle).
V. CONCLUSION

The BGN in quasi-neutral n-type and p-type regions and in a neutral electron–hole plasma of 20 technologically relevant zincblende III–V semiconductors was calculated in full-RPA as a function of density in the temperature range from 0 to 500 K. The sevenfold (plasma) and sixfold (quasi-neutral regions) integration is enabled by a careful isotropization of the band dispersion, including the important CB nonparabolicity in n-type materials. The neglect of higher CB valleys and of the split-off VB imposed by the simplified band structure model results in upper bounds for the allowed density in each case. The available experimental limits of the activated doping are always smaller than these theoretical limits. The gap can never shrink to zero, even at zero temperature. An exception is n-type InSb, where measured doping limits would lead to a vanishing gap at 0 K.

At room temperature, the trend of the BGN as a function of material (gap) follows the trend of the DOS effective masses. At a density of $10^{18}$ cm$^{-3}$, the BGN increases from $\sim 15$ meV (InAsSb) to $\sim 50$ meV (AlN) with only a little difference between n-type, p-type, and eh-plasma. At the higher density of $10^{19}$ cm$^{-3}$, the BGN
of p-type and eh-plasma is still similar with an average increase from ~30 meV (InAsSb) to ~100 meV (AlN), but the BGN in direct n-type material becomes larger due to the small electron mass (reduced screening) and increases from ~40 meV (InAsSb) to ~120 meV (AlN). An exception is given by the indirect materials AlAs and GaP due to their similar CB- and VB-DOS. The nitrides are strong outliers with respect to the VB DOS mass, which becomes visible as maxima in the BGN curve.

At very low temperatures, p-type material and eh-plasma show almost ideal behavior at a density of 10^{18} cm^{-3}, which is not much different from that at 300 K. At the higher density of 10^{19} cm^{-3}, p-type material and eh-plasma are still comparable, but in direct n-type materials, the BGN becomes giant as a consequence of the weak screening of the hole-ion interaction by the majority of electrons. At a density of 10^{18} cm^{-3}, the BGN is relatively independent of material, taking values between ~50 meV and ~80 meV at 20 K. At 10^{19} cm^{-3}, it increases from ~130 meV (InAsSb) to ~200 meV (AlN), again with the exception of the indirect materials AlAs and GaP.

The strong temperature dependence of the BGN in n-type materials with small CB DOS is traced back to the occupation probability of minority holes. It persists at all electron concentrations. In p-type material and eh-plasma, the temperature dependence of the BGN is much weaker; it is only significant at low and intermediate densities and always ceases at high densities.

The second major result of the paper is the derivation of an analytical BGN model without material-dependent free fit parameters, applicable to all densities and temperatures. Compared to numerical full-RPA, it reduces the central processing unit (CPU) time from days (the worst case is a very low but non-zero temperature and a high density) to less than seconds. The analytical model reveals the dependencies on band-structure and material parameters and allows for a quick calculation of the BGN in materials not explicitly treated in this paper. The only restriction here is compliance with an isotropic dispersion model (which excludes wurtzite materials). Besides other model simplifications (e.g., neglect of disorder-induced DOS tails and polaronic mass enhancement), isotropization is the key not only for the numerical full-RPA calculations but also for the analytical model. After the exact angular integrations, threefold integrals over complicated functions are approximately solved for the free-carrier correlation energies at 0 K and for the non-perturbative temperature-dependent hole-ion correlation energy, including the nonparabolicity of the CB. This is enabled by taking advantage of the ratio between Fermi wavelength and Bohr radius and by applying other approximation methods. As a result, the functional dependence of the BGN(n,p,T) on material and band structure parameters can be preserved. Comparison between analytical and numerical curves shows very good agreement in many cases, in particular for the density and temperature dependence of the BGN in n-type materials. Larger deviations occur in the case of n-type InAsSb at high densities (extreme nonparabolicity), for the indirect materials AlAs and GaP (due to their comparable CB and VB DOSs), and for some p-type materials such as InAsSb, InSb, and the nitrides.

A serious attempt has been made to express the BGN of ternary compounds by a composition average of the BGNs of their binary constituents (without or with bowing), as is common for the bandgap. However, although possible to some extent for a particular density and temperature, it fails for a wider density interval. This is probably due to the involved dependence on band filling and screening.

AUTHOR DECLARATIONS
Conflict of Interest
The author has no conflicts to disclose.

Author Contributions
A. Schenk: Conceptualization (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Software (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal).

DATA AVAILABILITY
The data that support the findings of this study are available from the corresponding author upon reasonable request.

APPENDIX A: BANDGAP NARROWING AS FUNCTION OF DENSITY IN THE TEMPERATURE RANGE FROM 0 TO 500 K FOR ALL 60 CASES
This appendix presents the BGN for all 20 materials (n-type and p-type) and for the neutral eh-plasma in these materials as a continuous function of density from 10^{15.5} cm^{-3} to the maximum at the temperatures T = 0, 20, 77, 300, and 500 K. The full-RPA curves for T = 0 K (black solid) start at the Mott density. The end point of each density axis corresponds to the allowed density maximum according to the measured limits of activated doping or to the limitations of the band-structure model. Solid lines are used for numerical full-RPA, i.e., Eqs. (6)–(8) with the RPA dielectric function Eq. (9) and densities calculated with Eq. (10) (parabolic case) or Eq. (11) (non-parabolic case), symbols are used for the analytical model outlined in Sec. IV.

APPENDIX B: ANALYTICAL APPROXIMATION OF THE FREE-CARRIER CORRELATION ENERGIES
This appendix provides the derivation of the analytical expressions of the free-carrier correlation energies in quasi-neutral regions [second and third terms in Eq. (20)] and in the neutral eh-plasma [second and third terms in Eq. (26)].

1. Quasi-neutral regions
The zero-temperature correlation energy of free carriers is calculated by Eq. (7) in the parabolic band approximation (justification given in the main text). The integration of angles can be performed exactly. Furthermore, the limit β → ∞ is utilized. First, in the factor ∂ψ/∂n, the derivative of the distribution function of majority carriers can be generated by partial integration, which turns into a delta-function and removes one integral. Second, the Matsubara sum is replaced by \[ \sum_q \rightarrow 2eta/\pi \int_0^\infty dt \] with \[ t = \pi n/\beta \]. In the momentum integration, the variable q is normalized by the Fermi momentum of the majority carriers: \[ q \rightarrow q_{F,M} = \sqrt{\beta}/\alpha_M \]. This results in
The correlation energy (B1) takes the form

$$\Delta_c^s = -\frac{4\delta_b}{\pi \alpha_b^2} \int_0^\infty dz \int_0^\infty dt \frac{1}{2s_b + z^2 + 3\left(\frac{\tau}{\tau'} + \frac{\tau'}{\tau}\right)} \times \left\{ \frac{1}{2} \tan^{-1} \left[ \frac{z^2}{z^2 + \frac{\tau}{\tau'} + \frac{\tau'}{\tau}} \right] + \frac{\alpha_b}{\alpha_b^2} \right\} \Theta(\delta_b),$$

(B9)

where the integration variable $t$ was changed to $\tau = t/\zeta_b$. The goal is to approximate the integrand such that the correct dependence on density (contained in $s_b$) is preserved as well as possible. After testing various ways, the following turned out to be the best.

### a. Majority-carrier band edge

For the majority-carrier band edge (first term in curly braces), the frequency dependence in the denominator of the SPP screening function [first line in Eq. (B9)] is skipped ($\tau = 0$). This allows us to perform the remaining $\tau$-integration exactly, leading to

$$\Delta_{c,maj}^s \approx -\frac{2\delta_b \delta_b}{\pi \alpha_b^2} \int_0^\infty dz \frac{z \Theta(2-z) + 2 \Theta(z-2)}{2s_b + z^2 + 3z^2/4},$$

(B10)

The numerator of the integrand is replaced by $z \Theta(2-z) + 2 \Theta(z-2)$, which results in

$$\Delta_{c,maj}^s \approx -\frac{4\delta_b \delta_b}{3\pi \alpha_b^2} \int_0^\infty dz \frac{1}{2s_b + z^2 + 3z^2/4},$$

and after exact integration in the analytical expression

$$\Delta_{c,maj}^s \approx -\frac{2\delta_b \delta_b}{\pi \alpha_b^2 \sqrt{\delta_b}} \sqrt{1 + \sqrt{1 - 6s_b} + \sqrt{1 - \sqrt{1 - 6s_b}}}.$$  

(B12)

To obtain the density dependence in the form of a simple power law, the limits $s_b \ll 1/6$ and $s_b \gg 1/6$ are considered,

$$\Delta_{c,maj}^s \approx \frac{2\delta_b \delta_b}{3\pi \alpha_b^2} \frac{1}{\sqrt{s_b}} \left( \frac{2}{3} \right)^{\frac{1}{6}} \quad s_b \ll 1/6,$$

(B13)

and

$$\Delta_{c,maj}^s \approx \frac{2\delta_b \delta_b}{\pi \alpha_b^2} \ln \left( 1 + \frac{2}{\sqrt{s_b}} \right) \quad s_b \gg 1/6.$$  

(B14)

The first limit (high density, short Fermi wavelength) yields an $n^{1/6}$-dependence, and the second limit (low density, long Fermi wavelength) yields an $n^{1/3}$-dependence. In Fig. 30, $s_b$ and $s_b$ at a density of $10^{19}$ cm$^{-3}$ are plotted for all materials. Neither of the two limits is actually valid, as one can see by comparison with the green dashed line (1/6). A possible simple fit to the full-RPA curves (compare Fig. 9) is the following: (i) the limit $s_b \ll 1/6$ is chosen, (ii) a logarithmic weakening$^{51,58}$ is performed, and (iii) the numerical pre-factor is adjusted. The result is [the second term of Eq. (20)],

$$\Delta_{c,maj}^s \approx \frac{2\delta_b \delta_b}{3\pi \alpha_b^2} \ln \left( 1 + \frac{2}{\sqrt{s_b}} \right)$$

i.e., a logarithmic weakened $n^{1/6}$-dependence.
A feasible power-law fit to the full-RPA curves (compare Fig. 9) is obtained as follows: (i) the limit \( s_0 \gg 1/6 \) is chosen, and (ii) the numerical pre-factor is adjusted. The result is [the third term of Eq. (20)]:

\[
\Delta_{\text{min}}^c \approx -\frac{g_b (1 - \delta_b)}{\sqrt{3} \alpha_b s_0^{1/2}} \left[ \begin{array}{c} \frac{\sqrt{2}}{s_0} \\ \frac{2}{s_0^{1/2}} \end{array} \right] \frac{1}{z^{1/4}} \quad \text{for} \quad s_0 \gg 1/6.
\]

A. A. 1/2 dependence.

It should be noted that the power-law fits for majority and minority band edges are somewhat arbitrary, as the applicability of the limits strongly depends on material and density. This might explain the many variants suggested in the past, e.g., \( n^{1/3} \) in Ref. 37, \( n^{1/4} \) in Ref. 38, and \( n^{1/2} \) in Refs. 38 and 41.

### 2. Neutral electron-hole plasma

Screening by electrons and holes with density \( n = n_e = n_h \) determines the free-carrier correlation energies in a neutral e-h plasma. Analytical modeling proceeds similarly to the case of a one-component plasma. Charge neutrality imposes a simple relation between the quasi-Fermi levels (parabolic dispersion, \( T = 0 \text{K} \)),

\[
\zeta^e = \left( \frac{g_b}{g_0} \right)^{2/3} \frac{\alpha_0}{
\alpha_e + \zeta_e}.
\]

In the screening function \( (1 - 1/\epsilon) \), the SPP approximation is now applied to a two-component plasma,

\[
\epsilon_{\text{app},\alpha} (q, \nu) = 1 + \frac{\omega_p^2}{\omega^2 + \frac{h \nu}{\epsilon_0}} - \Omega^2_\nu.
\]

Using relation (B19), plasmon frequency \( \omega_p \) and Thomas-Fermi momentum \( \kappa \) at \( T = 0 \text{K} \) are given by

\[
h^2 \omega_p^2 = \frac{8}{3 \pi} \left( \frac{g_b \sqrt{\alpha_e}}{\sqrt{\alpha_e}} + \frac{g_h \sqrt{\alpha_h}}{\sqrt{\alpha_h}} \right) \frac{8g_b}{3 \pi} \frac{\zeta_e}{\alpha_e} = \frac{8g_h}{3 \pi} \frac{\zeta_h}{\alpha_h},
\]

\[
\kappa^2 = \frac{2}{\pi} \left( \frac{g_b \sqrt{\alpha_e}}{\sqrt{\alpha_e}} + \frac{g_h \sqrt{\alpha_h}}{\alpha_h} \right) \frac{2g_h \sqrt{\alpha_h}}{\pi} \frac{\zeta_h}{\alpha_h} = \frac{2g_b \sqrt{\alpha_e}}{\pi} \frac{\zeta_e}{\alpha_e} = \frac{2g_b \sqrt{\alpha_e}}{\pi} \left( 1 + \frac{g_b}{g_h} \right) \frac{1}{\alpha_e}.
\]

The screening function reads

\[
1 - \epsilon^{-1}_{\text{app},\alpha} \left( \sqrt{\alpha_0/s_0}, t \right) = \frac{2s_0 \alpha_0}{2s_0 \alpha_0 + z^2 \alpha_0 + 3 \left( \frac{\alpha_0}{\alpha_e} \right)^2}.
\]

with

\[
\alpha_0 = \alpha_e + \frac{1}{(\frac{\alpha_0}{\alpha_e})^2}.
\]

The form of the correlation energy at zero temperature is similar to the majority-carrier term in Eq. (B1),

\[
\Delta_0^c = -\frac{2}{\pi \sqrt{\alpha_b s_0}} \frac{\omega_0}{z} \int_0^\infty dt \left[ 1 - \epsilon_{\text{app}, e} (z \sqrt{\alpha_b / \alpha_b, t}) \right] \times \tanh^{-1} \left[ \frac{z}{1 + \left( \frac{\alpha_0}{\alpha_e} \right)^2} \right].
\]

With the same approximations as for Eqs. (B10) and (B11), one obtains

\[
\Delta_0^c \approx -\frac{d_g \alpha_0}{3 \pi^2} \int_0^\infty \frac{dz}{z_s \alpha_0 + \frac{1}{\alpha_0} \alpha_b}.
\]

where \( \alpha_b \) is defined in Eq. (B24). Exact integration yields the analytical expression
\[ \Delta_\alpha \approx -\frac{2g_0}{\pi\sqrt{\alpha_b}} \sqrt{\frac{\alpha_b/\alpha_{ba}}{1 + \sqrt{1 - s_b/\alpha_b} + \sqrt{1 - s_b/\alpha_b}}}. \]  

(B27)

with

\[ \sigma_b = \frac{2\alpha_{ba}}{3\alpha_b}. \]  

(B28)

Again, the limits of very small and very large \( s_b \) are considered in order to simplify the model,

\[ \Delta_\alpha \approx -\frac{g_0}{\pi} \left( \sqrt{\frac{\alpha_b}{\alpha_{ba}}} \frac{2/\sqrt{3}}{s_b} \left\{ \begin{array}{ll} \frac{\alpha_b}{\alpha_{ba}} & \text{if } s_b \ll \sigma_b, \\
\frac{1}{\sqrt{3}} & \text{if } s_b >> \sigma_b. \end{array} \right. \right) \]  

(B29)

As can be seen from Fig. 30, the assumption \( s_c >> \sigma_c \) is fulfilled for all materials, whereas \( s_b >> \sigma_b \) is only fulfilled for the indirect materials AlAs and GaP. For the antimonides and InAs, it rather holds that \( s_b << \sigma_b \). Inspection of the full-RPA correlation energies at a density of \( 10^{19} \text{ cm}^{-3} \) shows that \( \Delta_c \approx \Delta_\alpha \) is only for the indirect materials AlAs and GaP; otherwise, the ratio \( \Delta_c/\Delta_\alpha \) increases from 2 to 6 with decreasing gap. Therefore, the same simple power law for \( \Delta_\alpha \) cannot work in all 20 cases.

To achieve a good agreement with the full-RPA curves in all 20 cases, a distinction is made based on the parameter \( \alpha_e = m_b/(m_0 + m_e) \): if \( \alpha_e > 0.85 \) (materials No. 1–11), the limit \( s_b << \sigma_b \) is used; if \( \alpha_e < 0.85 \) (materials No. 12–20), the limit \( s_b >> \sigma_b \) is applied. A logarithmic weakening \( \Delta_\alpha \) is performed, and the numerical pre-factor is adjusted independently for electrons and holes. The result is [the second and third terms of Eq. (20)],

\[ \Delta_c + \Delta_\alpha \approx -\frac{2g_0}{5\pi} \ln \left[ 1 + \left( \frac{8\alpha_b}{3} \right)^{1/2} \frac{1}{\sqrt[3]{3}} \right] \]

\[ -\frac{2g_0}{\pi} \left( \Theta(0.85 - \alpha_e) \ln \left[ 1 + \left( \frac{8\alpha_b}{3} \right)^{1/2} \frac{1}{\sqrt[3]{3}} \right] \right) \]

\[ + \Theta(\alpha_e - 0.85) \ln \left[ 1 + \sqrt{\frac{2\alpha_b}{\alpha_{he}\alpha_b}} \right]. \]  

(B30)

The correlation energy of the ion-carrier interaction is calculated from Eq. (8). After normalization and integration over angles, it is given by

\[ \Delta_\beta = -m_{s} \frac{2}{\pi} \left( \frac{\partial n_{s}}{\partial \alpha} \right)^{-1} \int_{0}^{\infty} dq \frac{1}{e^{2} (q, 0)} \frac{\partial e(q, 0)}{\partial \alpha}. \]  

(C1)

The static RPA dielectric function reads [compare Ref. 7, Eqs. (A7)–(A9)],

\[ \epsilon(q, 0) = 1 + \frac{1}{\pi a_{q}^{2}} \sum_{k} \frac{g_{k}}{a_{k}^{2}} \int_{0}^{\infty} d\kappa \kappa \left\{ 2\left[ h_{k}(\kappa, q) - h_{k}(\kappa, -q) \right] \right\} + h_{k}(\kappa, 0) \ln \left[ \frac{h_{k}(\kappa, 0) - h_{k}(\kappa, q)}{h_{k}(\kappa, 0) - h_{k}(\kappa, -q)} \right], \]  

(C2)

where

\[ h_{k}(\kappa, q) = \sqrt{1 + 4\gamma_{k}(\kappa + \sqrt{\kappa^{2} q^{2}})}, \]  

(C3)

\[ f_{k}(\kappa) = \left( 1 + e^{\left[ (h_{k}(\omega_{k}) - 1)/2\gamma_{k} \right] \kappa^{\frac{1}{2}}} \right)^{-1}, \]  

(C4)

have to be used. For the derivative of \( e(q, 0) \) with respect to the Fermi energy, one needs

\[ \frac{\partial f_{k}}{\partial \kappa} = -\frac{\partial f_{k}}{\partial h_{k}(\kappa, 0)} \frac{2}{\partial \kappa} = \beta f_{k}(\kappa) \left[ 1 - f_{k}(\kappa) \right], \]  

(C5)

under the \( \kappa \)-integral. To obtain the \( T = 0 \) K-limit, the \( \kappa \)-integral in Eq. (C2) is integrated by parts, where the derivative of the distribution function with respect to \( \kappa \) becomes a delta function,

\[ \frac{\partial f_{k}}{\partial \kappa} = \delta(\kappa - \sqrt{\omega_{k}}(1 + y_{s})), \]  

(C6)

Then

\[ \epsilon_{\gamma}(q, 0) = 1 + \frac{g_{0}}{\pi a_{q}^{2}} \int_{0}^{\infty} d\kappa \kappa \left\{ 2\left[ h_{k}(\kappa, q) - h_{k}(\kappa, -q) \right] \right\} + h_{k}(\kappa, 0) \ln \left[ \frac{h_{k}(\kappa, 0) - h_{k}(\kappa, q)}{h_{k}(\kappa, 0) - h_{k}(\kappa, -q)} \right] \Theta(\kappa), \]  

(C7)

with the abbreviation \( \tilde{\omega}_{k} = \omega_{k}(1 + y_{s} \tilde{\omega}_{s}) \) and “\( \tilde{\omega} \)” denoting the index of the majority carrier band. Only majority carriers contribute to screening in quasi-neutral regions \( (\xi_{\alpha} > 0) \), whereas minority carriers are frozen out. The derivative of the densities with respect to the Fermi energy becomes

\[ \lim_{\beta \rightarrow 0} \frac{\partial n_{s}}{\partial \beta} = -\frac{g_{3}}{4\pi \alpha_{b}^{2} \sqrt[3]{3}} \left\{ \begin{array}{ll} \frac{\sqrt{\pi}}{\pi} \exp(\beta_{s} \tilde{\omega}_{s}) & \text{majority band}, \\
\sqrt{\pi} \exp(-\beta_{s} \tilde{\omega}_{s}) & \text{minority band}. \end{array} \right. \]  

(C8)
For the derivative of the static dielectric function with respect to the Fermi energy \( \partial_e(q, 0) / \partial \xi \) in the limit \( \beta \to \infty \), one has

\[
\lim_{\beta \to \infty} \frac{\partial f_a}{\partial \xi} = -\lim_{\beta \to \infty} \frac{h(k, 0)}{2k} \frac{\partial f_a}{\partial k}
\]

\[
= \left[ \frac{h(k, 0)}{2k \beta} \right] \left( \kappa - \sqrt{\xi} \right) \Theta(\xi) \text{ majority band,} \\
\beta \to \infty \frac{1}{\beta} \Theta(-\xi) \text{ minority band,}
\]

(C9)

which follows from Eq. (C5) neglecting nonparabolicity for the minority carriers. Using this in Eq. (C2) leads to

\[
\lim_{\beta \to \infty} \frac{\partial e(q, 0)}{\partial \xi} = \frac{1}{\pi q^2 a_s^2} \left\{ \frac{h_s(\sqrt{\xi}, q) - h_s(\sqrt{\xi}, -q)}{h_s(\sqrt{\xi}, 0)} + \frac{1}{2} \ln \left( \frac{h_s(\sqrt{\xi}, 0) - h_s(\sqrt{\xi}, q)}{h_s(\sqrt{\xi}, 0) - h_s(\sqrt{\xi}, -q)} \right) \right\} \Theta(\xi) + \lim_{\beta \to \infty} \frac{1}{\pi q^2 a_s^2} \int_0^\infty dx \kappa e^{-\beta x^2} \left( \kappa^2 - (\kappa + \sqrt{\alpha q})^2 \right) \Theta(-\xi),
\]

(C10)

where in the term \( \sim \Theta(-\xi) \) nonparabolicity was again neglected. The \( \kappa \)-integral can be calculated analytically,

\[
\lim_{\beta \to \infty} \frac{1}{\pi q^2 a_s^2} \int_0^\infty dx e^{-\beta x^2} \ln \left( \frac{\kappa^2 - (\kappa + \sqrt{\alpha q})^2}{\kappa^2 - (\kappa - \sqrt{\alpha q})^2} \right) = \frac{1}{\pi q^2 a_s^2} \int_0^\infty dt t e^{-t} \tanh^{-1} \left( \frac{4T}{\alpha \beta q} \right)
\]

(C11)

For the second line, the relation \( \ln((x + 1)/(x - 1)) = 2 \tanh^{-1}(1/x) \) was used. Inserting Eq. (C10) with (C11) and Eq. (C8) into Eq. (C1), the correlation energy of ion-carrier interaction in the \( T = 0 \) K-limit becomes

\[
\lim_{\beta \to \infty} \Delta_s^i = -n_s \frac{8}{\sqrt{\alpha q}} \int_0^\infty \frac{dq}{q} \Theta(q, 0)
\]

\[
\times \left\{ h_s(\sqrt{\xi}, q) - h_s(\sqrt{\xi}, -q) \right\} + \frac{1}{2} \left( 1 + 2\sqrt{\alpha q} \right) \ln \left( \frac{1 + 2\sqrt{\alpha q} - h_s(\sqrt{\xi}, q)}{1 + 2\sqrt{\alpha q} - h_s(\sqrt{\xi}, -q)} \right)
\]

\[-n_s \frac{32}{\alpha q} \int_0^\infty \frac{d\xi}{\xi^{1/2}} \Theta(-\xi)
\]

(C12)

where the static dielectric function \( \epsilon_{\text{static}}(q, 0) \) is defined in Eq. (C7) and the function \( h_s \) in Eq. (C3). Introducing normalized momentum variables of majority carriers by \( z = q / \sqrt{\xi on} \) and \( t = \kappa / \sqrt{\xi_0} \), this can be written in the form

\[
\lim_{\beta \to \infty} \Delta_s^i = -\frac{2g_b}{3\pi^2} \int_0^\infty \frac{d\xi}{\xi^{1/2}} \left[ \frac{h_s(1, z) - h_s(1, -z)}{z^2 + \frac{1}{2} \int_0^1 dt \ h_s(t, z)} \right] \Theta(\xi),
\]

(C13)

with

\[
\begin{align*}
\frac{g_b}{\pi \sqrt{\alpha q_0} \xi_0} = s_b = \frac{g_b}{\pi \sqrt{\alpha q_0} \xi_0} \frac{1}{6n^2 \beta q}.
\end{align*}
\]

(C15)

Therefore, one can assume \( s_b \ll 1 \) for very large \( n_s \) which results in a pronounced maximum of the integrand at some \( z < 1 \). The function \( h_s(t, z) \) can then be linearized in \( z \), which results in

\[
H_s(t, z) \approx \frac{2z}{t} \int_0^1 \frac{d\xi}{\xi} h_s(1, 0) + \frac{4y_0 \xi_s t^2}{h_s(1, 0)}.
\]

(C16)

With that, the \( t \)-integration in the dielectric function can be performed easily, giving

\[
\frac{h_s}{z} \int_0^1 \frac{d\xi}{\xi} h_s(t, z) = s_b \left\{ 2h_s(1, 0) + \left[ h_s(1, 0) - \frac{\sinh^{-1}(4y_0 \xi_s t^2)}{4y_0 \xi_s t^2} \right] \right\}.
\]

(C17)

The second term is negligible compared to the first one as long as \( 4y_0 \xi_s t^2 \ll 1 \) can be assumed, meaning that the Fermi energy in the corresponding parabolic band must be much smaller than the bandgap. For InAsSb and InSb, this becomes critical at higher densities, as can be seen in Figs. 10 and 11. With the linearized form of \( h_s(1, z) \) in the numerator of Eq. (C13) and inserting

\[
\frac{S_b}{z} \int_0^1 \frac{d\xi}{\xi} h_s(t, z) = 2n_b h_s(1, 0),
\]

(C18)
in the denominator, the remaining $z$-integration is trivial and yields
\[
\lim_{\beta \to \infty} \Delta_\text{\scriptscriptstyle A} = - \frac{2\gamma_b}{3\pi} \left( \frac{1}{\sqrt{2\pi n_b h_b(1,0)}} \right) \delta_{ab} + \frac{1}{2n_b h_b(1,0)} \Theta(\zeta_b).
\] (C19)

A last simplification is the replacement
\[
\sqrt{n_b(1,0)} + \frac{8\gamma_b \zeta_b}{h_b^{3/2}(1,0)} \to h_b(1,0),
\] (C20)

that closely matches the numerical RPA result. Using the notation $\varphi_{\text{np}}(s_b) = h_b(1,0)$ for the nonparabolicity correction factor finally results in the ionic correlation energies in Eq. (20). Note that $\gamma_b = 0$.

As discussed in Appendix B 1, the strong temperature dependence of the hole-ion correlation energy $\Delta_{\text{\scriptscriptstyle A},\text{\scriptscriptstyle n-type}}(T)$ mainly arises from the occupation probability of holes, whereas the majority-carrier screening is only weakly $T$-dependent and vanishes at high concentrations. Therefore, Eq. (C1) becomes
\[
\Delta_{\text{\scriptscriptstyle A},\text{\scriptscriptstyle n-type}} = -n_b \left( \frac{\partial h_b}{\partial \zeta_b} \right)^{-1} \int_0^\infty dq \frac{1}{e^{\varphi(q,0)/T} - 1} \frac{\partial \varphi(q,0)}{\partial \zeta_b},
\] (C21)

Since $\gamma_b = 0$ is used throughout, nonparabolicity only needs to be taken into account in the factor $1/e^{\varphi_{\text{\scriptscriptstyle np}}(q,0)}$ (screening by majority electrons). Furthermore, the $T$-dependent factors $(\partial h_b/\partial \zeta_b)^{-1}$ and $\partial \varphi(q,0)/\partial \zeta_b$ can be calculated with Boltzmann statistics. From Eq. (10), one obtains after normalization
\[
\left( \frac{\partial h_b}{\partial \zeta_b} \right)^{-1} = \frac{4\pi n_b}{h_b^{3/2}} \sqrt{\pi} e^{-\zeta_b},
\] (C22)

and from Eq. (C2) after partial integration
\[
\frac{\partial \varphi(q,0)}{\partial \zeta_b} = \frac{2}{2\pi n_b q^2} \sqrt{\zeta_b} \int_0^\infty dk e^{-\zeta_b} \frac{dk}{dk} H_b(k, q),
\] (C23)

with $H_b(k, q)$ defined in Eq. (C14). Its parabolic version reads
\[
H_b(k, q) = \tanh^{-1} \left( \frac{\sqrt{\pi} \alpha q^2}{\kappa^2 + \alpha q^2/4} \right). \tag{C24}
\]

Inserting
\[
\frac{d}{dk} H_b(k, q) = -2q \sqrt{\pi} \frac{k^2 - \alpha q^2/4}{(k^2 - \alpha q^2/4)^2} \left( \frac{q\sqrt{\pi} - \kappa}{2} - \kappa \right),
\] (C25)

into Eq. (C23) and introducing the new integration variables $t = k\sqrt{\beta}$ and $z = q\sqrt{\alpha}/\sqrt{\zeta_b}$, consequently,
\[
\left( \frac{\partial h_b}{\partial \zeta_b} \right)^{-1} \frac{d}{dk} H_b(k, q) = \frac{8\pi n_b}{z^2 \sqrt{\pi}} \tau(z, \beta), \tag{C26}
\]

where
\[
\tau(z, \beta) = \int_0^{\alpha(\beta)} d\beta \beta e^{\beta^2 z^2 - \beta^2}, \tag{C27}
\]

for an analytical solution of the integral in (C30), the error function is approximated as
\[
er\left( \frac{2x}{\sqrt{\pi}} \right) \approx \Phi(1 - x) + \Theta(x - 1). \tag{C31}
\]

This results in
\[
\Delta_{\text{\scriptscriptstyle A},\text{\scriptscriptstyle n-type}}(T) = -\frac{\sqrt{2\pi}}{3\pi n_b} \frac{\varphi_{\text{np}}}{\gamma^3/2} \frac{h_b^{3/2}}{\psi_{\text{np}}(s_b)} \times \left\{ 1 - 2 - \arctan \left( \frac{\alpha(\beta) \sqrt{2\pi} \varphi_{\text{np}}(s_b)}{2\pi} \right) \right\} + \frac{2(2 - \sqrt{\pi})}{\gamma^3/2} \frac{\alpha(\beta) \sqrt{2\pi} \varphi_{\text{np}}(s_b)}{1 + \alpha(\beta) \sqrt{2\pi} \varphi_{\text{np}}(s_b)},
\] (C32)

where $h_b(1,0)$ was renamed by $\varphi_{\text{np}}(s_b)$, which is the symbol for the nonparabolicity correction factor Eq. (21) in the main text. The last term in curly braces in Eq. (C32) is small compared to the second one and will be omitted. Inserting $\alpha(\beta)$ from Eq. (28) gives
\[
\Delta_{\text{\scriptscriptstyle A},\text{\scriptscriptstyle n-type}}(T) = -\frac{\sqrt{2\pi}}{3\pi n_b} \frac{\varphi_{\text{np}}}{\gamma^3/2} \frac{h_b^{3/2}}{\psi_{\text{np}}(s_b)} \times \left\{ 1 - 2 - \arctan \left( \frac{2\pi n_b^{2/3} \varphi_{\text{np}}(s_b)}{\beta \varphi_{\text{np}}(s_b)} \right) \right\} \frac{2\pi n_b^{2/3} \varphi_{\text{np}}(s_b)}{\beta \varphi_{\text{np}}(s_b)}.
\] (C33)

In this approximation, the temperature dependence has the form $\sim \left( 1 - 2 - \arctan \left( \pi \sqrt{T/\pi^3} \right) / \pi \right)$. Due to the assumption of very large Fermi momenta in the dielectric function, $z \ll 1$, it only matches the numerical results at low temperatures and extremely high densities, which are outside the validity range.

The opposite assumption, $z \gg 1$, would lead to $H_b(t, z) \approx 8t/z$. The dielectric function then reads $e^{\varphi_{\text{np}}(z, 0)} = 1 + 8s_t/(3z^2)$, resulting in a temperature and density dependence of the correlation energy given by $\sim \sqrt{T/\pi}$ as in the Debye limit [compare Eq. (16)]. Combining both limits with the static form of the SPP approximation (B7) leads to cumbersome expressions that are not helpful. Therefore, to cover the cases $z \ll 1$ and $z \approx 1$, a viable solution is to
replace arctan[(\cdots)^{1/2}] by arctan[(\cdots)^{3/4}] in (C33). This not only preserves the zero-temperature limit but also results in excellent agreement with the numerical curves for all n-type materials over the whole density and temperature range, as can be seen from Fig. 8 and the left panel of Figs. 10–29.

REFERENCES