Auger generation as an intrinsic limit to tunneling field-effect transistor performance

Cite as: J. Appl. Phys. **120**, 084507 (2016); https://doi.org/10.1063/1.4960571
Submitted: 07 June 2016 . Accepted: 26 July 2016 . Published Online: 29 August 2016

James T. Teherani, Sapan Agarwal, Winston Chern, Paul M. Solomon, Eli Yablonovitch, and Dimitri A. Antoniadis

Auger generation as an intrinsic limit to tunneling field-effect transistor performance

James T. Teherani,[1,a)] Sapan Agarwal,[2] Winston Chern,[3] Paul M. Solomon,[4] Eli Yablonovitch,[5] and Dimitri A. Antoniadis[3]
[1]*Department of Electrical Engineering, Columbia University, New York, New York 10027, USA*
[2]*Sandia National Laboratories, Albuquerque, New Mexico 87123, USA*
[3]*Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*
[4]*IBM T.J. Watson Research Center, Yorktown Heights, New York 10598, USA*
[5]*Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, California 94720, USA*

(Received 7 June 2016; accepted 26 July 2016; published online 29 August 2016)

Many in the microelectronics field view tunneling field-effect transistors (TFETs) as society's best hope for achieving a $>10\times$ power reduction for electronic devices; however, despite a decade of considerable worldwide research, experimental TFET results have significantly underperformed simulations and conventional MOSFETs. To explain the discrepancy between TFET experiments and simulations, we investigate the parasitic leakage current due to Auger generation, an intrinsic mechanism that cannot be mitigated with improved material quality or better device processing. We expose the intrinsic link between the Auger and band-to-band tunneling rates, highlighting the difficulty of increasing one without the other. From this link, we show that Auger generation imposes a fundamental limit on ultimate TFET performance. *Published by AIP Publishing.*
[http://dx.doi.org/10.1063/1.4960571]

I. INTRODUCTION

Deceleration of Moore's law scaling of complementary metal-oxide-semiconductor (CMOS) field-effect transistors has prompted an urgent search for next-generation logic devices.[1] Of all exploratory devices under consideration, tunneling field-effect transistors (TFETs) offer the most promise due to their potential speed, energy efficiency, and compatibility with existing CMOS infrastructure.[2,3]

The key advantage of TFETs over existing technologies is the possibility for significant energy reduction achieved through energy-efficient switching of band-to-band tunneling (BTBT) current. The switching characteristics of transistors are assessed by the subthreshold swing, $dV_G/d\log_{10}(I_D)$, which gives the change in gate voltage to produce a $10\times$ increase in drain current. The quantum-mechanical nature of BTBT allows significantly sharper switching for TFETs compared to conventional transistors. [The subthreshold swing of conventional transistors is governed by the Fermi-Dirac distribution of carriers above a potential barrier whose height is modulated by the gate voltage. The thermionic current resulting from carriers above this barrier produces the often-cited 60-mV/decade subthreshold swing at 300 K. TFETs, on the other hand, can overcome the 60-mV/decade thermal limit because they switch by modulating tunneling through a barrier, instead of over it.] Sharper switching of TFETs, corresponding to smaller subthreshold swings, enables lower voltage operation yielding substantial energy savings over conventional devices.

However, despite many years of considerable worldwide effort, TFETs with switching characteristics superior to MOSFETs (i.e., TFETs with subthreshold swings <60 mV/decade at 300 K) for currents greater than 10 nA/μm have not been demonstrated. (See Lu and Seabaugh's review[4] for a compilation of the most recent TFET results.) Previous work focused on trap-related phenomena, such as Shockley–Read–Hall generation, degraded gate efficiency, and trap-assisted tunneling, to explain non-ideal device characteristics.[5–8] More recently, Teherani *et al.*[9] suggest that *intrinsic* phenomena, such as Auger, phonon-assisted, and radiative generation, may also limit TFET performance and explain the 60-mV/decade room-temperature swings observed in several experimental papers.[10–12] This work expands on this idea and describes an intrinsic mechanism—Auger generation—that may limit the minimum subthreshold swing, off-current, and on/off ratio of experimental TFETs. Auger generation must be understood to improve TFET switching characteristics so that sizeable energy reduction can be attained.

II. BACKGROUND ON AUGER GENERATION

Auger generation, also called impact ionization, occurs when a high-energy carrier collides with an electron in the valence band, resulting in the high-energy carrier losing much of its momentum and exciting the valence-band electron to the conduction band, creating an electron–hole pair. Auger generation is the reverse of Auger recombination, where the recombination of an electron–hole pair excites a third carrier to a high-kinetic-energy state. For both generation and recombination, the high-energy carrier can be either an electron or a hole (Figure 1).

[a)]Author to whom correspondence should be addressed. Electronic mail: j.teherani@columbia.edu

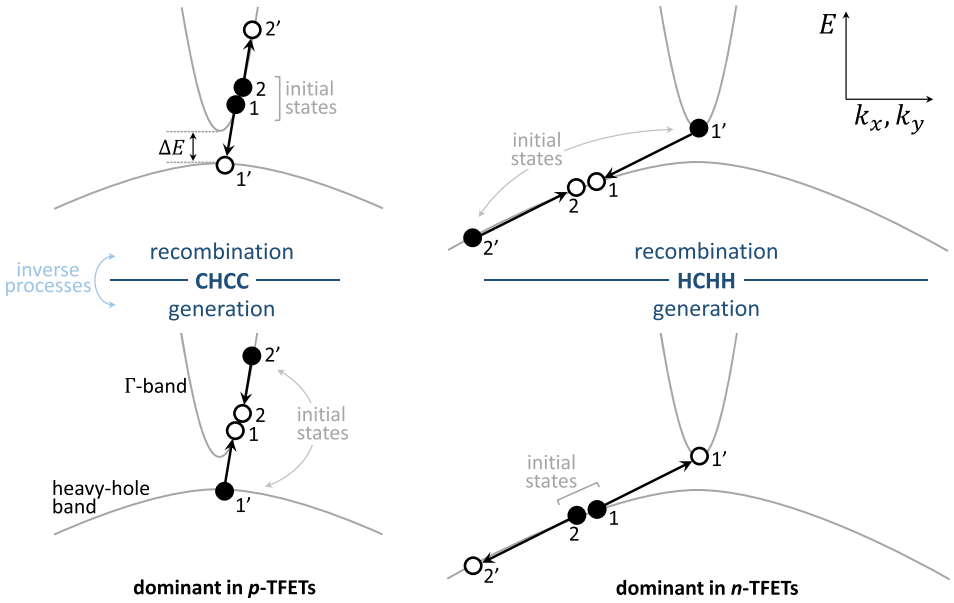

FIG. 1. CHCC and HCHH generation and recombination processes. C and H refer to states in the Γ-*C*onduction and *H*eavy-hole bands. Recombination results in the loss of the electron–hole pair 1-1′, whereas generation, the inverse process, results in the creation of the electron–hole pair. The arrows show the transitions of electrons from occupied states (filled circles) to vacant states (empty circles) that occur for each process. By convention, 2′ denotes the high-energy state, and 1′ denotes the state in a different band. Energy and momentum must be conserved in all processes.

Auger generation is thought to be significant at very high carrier densities or very large electric fields. While this observation is true for conventional semiconductors, Auger generation can also play a fundamental role in very small band-gap semiconductors, even at moderate carrier densities. As a practical example, consider Auger generation in long-wavelength HgCdTe infrared photodiodes ($E_G \sim 0.1$ eV). In these devices, it is well understood that, at room temperature, Auger generation dominates the reverse-bias dark-current of high-quality devices and limits the detector sensitivity.[13,14] The sizable Auger-induced leakage current in these devices supports the idea that Auger generation must be considered in small band-gap materials.

Figure 2 plots the reverse-bias dark-current-density for ~70 photodiodes made from high-quality HgCdTe,[15] InSb,[16] and InGaAs (lattice-matched to InP).[16,17] Reverse-bias currents for experimental devices follow an exponential trend (labeled "Rule 07" after the original work) that increases as the band gap shrinks and the temperature increases. Tennant[15] shows that a simple model of Auger generation [which depends on the probability of a high-energy carrier with enough energy to create an electron-hole pair] explains the exponential behavior and gives excellent agreement over 13 orders of magnitude of dark-current-density. The trend line in Figure 2 suggests a sizable leakage current for a room-temperature device with a small band gap.

The total current for a *pn* junction is the combination of diffusion current due to the decay of excess carriers in the quasi-neutral regions and depletion current due to generation and recombination in the depletion region. Generation and recombination mechanisms, such as Auger, will create depletion currents for non-zero bias. These same generation and recombination mechanisms also determine the minority carrier lifetime that influences the diffusion current. The generation and recombination rates, however, vary as a function of position due to different carrier concentrations and electric fields.

In a typical reverse-biased photodiode, Auger generation is most dominant in the quasi-neutral regions as compared to the depletion region because few carriers exist in the depletion region. Auger generation reduces the minority carrier lifetime and increases the reverse-bias diffusion current leading to the observed dependence plotted in Figure 2. While other generation and recombination processes are also exponentially dependent on the band gap and temperature, it is well established, through theory and experiment,[14] that Auger generation is the dominant process in high-quality small-band-gap HgCdTe photodetectors.]

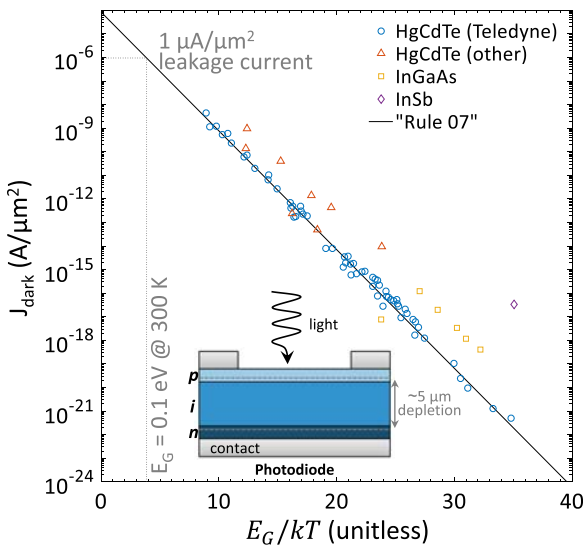

FIG. 2. Reverse-bias dark-current-density measured for a variety of different photodetectors, adapted from Refs. 15 and 16. The large dataset for HgCdTe consists of devices with band gaps ranging from 80 to 500 meV. The devices were measured at a variety of temperatures from 60 to 313 K. "Rule 07" is an empirical fit to the measured data suggested by Ref. 16 that captures the experimental trend over 13 orders of magnitude. The plot shows that the leakage current for a photodiode increases exponentially as the band gap decreases or the temperature rises, in agreement with an Auger-induced mechanism. The trend line suggests a leakage current of ~1 μA/μm^2 for a room-temperature device with a 0.1-eV band gap. The inset shows a typical infrared photodiode structure.

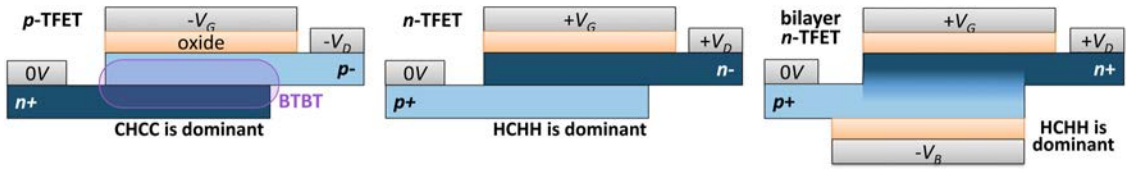

FIG. 3. Device structure of perpendicular *p*-, *n*-, and bilayer TFETs. The terminal voltages indicate a typical biasing scheme, and the purple oval indicates the region over which band-to-band tunneling (BTBT) nominally takes place. The *p* and *n* layers can be made from the same material or different materials to form a homostructure or heterostructure device. The bilayer TFET[29] has no intentional channel doping, and uses top and bottom gates to electrostatically induce electrons and holes. The bilayer *n*-TFET is often biased with a constant negative back gate voltage ($-V_B$) to create a hole-rich layer along the bottom of the channel, while the top gate voltage ($+V_G$) is varied to modify the energy alignment of the electron and hole eigenstates.

Though many TFETs are fabricated from conventional semiconductors with relatively large band gaps, the energy separation between the electron and hole eigenstates of the tunneling region becomes exceedingly small as the device is turned on. For example, the perpendicular *n*-TFET (Figure 3) is switched on by biasing the structure so that the hole eigenstate of the source aligns in energy with the electron eigenstate of the channel. Ideally, the device remains off and no current flows until the electron and hole eigenstates align in energy (Figure 4(c)). But right before the bands align, a small energy separation exists between the electron and hole eigenstates in the tunneling region. *The eigenstate energy separation of the tunneling region resembles a small-band-gap semiconductor whose band gap changes with applied gate voltage*.

Consequently, as will be shown, Auger generation, in addition to other generation and recombination processes, will occur at significantly higher rates than traditionally expected due to the small eigenstate energy separation of the tunneling region as compared to the large band gap of the bulk material.

To further underscore the importance of intrinsic generation and recombination in TFETs, consider a device just before ideal band-to-band tunneling occurs where the electron and hole states of the tunneling junction are separated by the thermal energy, kT (e.g., Figure 4(b)). Tunneling occurs only when the states align in energy if there is *some* spatial overlap of the electron and hole states at the tunneling junction. [A zero spatial overlap signifies that the electrons and holes are completely disconnected from each other. Even with abrupt band-edges, the amplitudes of the electron and hole wavefunctions have an approximately exponential decay in the band gap of the material. Under the WKB framework, this decay is viewed as the exponentially decreasing tunneling probability with barrier thickness.] Given the spatial overlap of electrons and holes and the very small energy separation between the eigenstates, non-defect-mediated generation and recombination processes, often negligible in conventional devices, can couple the electron and hole states with significant transition rates. In this work, we address this problem by first establishing a framework for analyzing Auger generation, and then calculating the Auger rate as a function of the energy separation between the electron and hole eigenstates. The rate is then related to a minimum current, and its impact on subthreshold swing is shown to set intrinsic limits to TFET device performance.

III. DEVICE STRUCTURE AND OPERATION

We analyze the Auger transition rate for the perpendicular TFET structures depicted in Figure 3. The ideal turn-on of the device occurs when a sufficient gate bias is applied such that the electron eigenstate of the *n*-layer aligns in energy with the hole eigenstate of the *p*-layer (Figure 4(c)).

Two doping versions of the structure exist, the *n*-TFET and *p*-TFET (Figure 3). They are named in analogy with *n*- and *p*-MOSFETs since they follow the same biasing conventions.

Generally, the potential bias of the *p*+ source of the *n*-TFET sets the quasi-Fermi level for holes while the *n*+ drain sets the quasi-Fermi level for electrons. For an *n*-TFET,

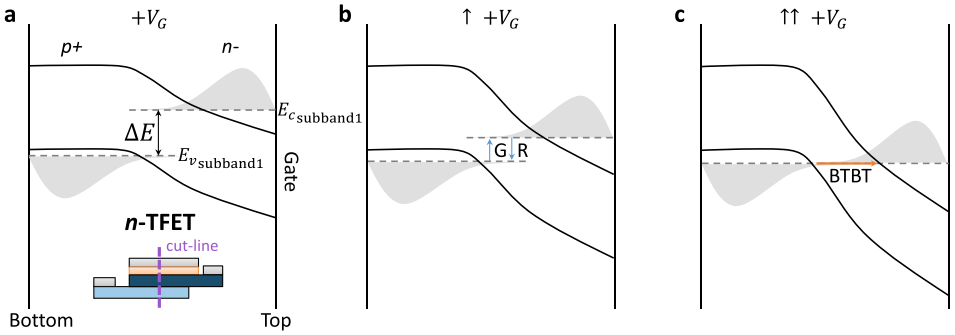

FIG. 4. Energy-band diagram across the tunneling region of an *n*-TFET. The gate voltage is increased from a to c. The dashed lines represent the ground-state subband energy for the electron and hole eigenstates. The gray area represents the electron and hole density as a function of position. (a) ΔE is the energy separation between the first electron and hole eigenstates, which varies with gate voltage. (b) As the gate voltage is increased, band bending occurs mostly in the *n*-region reducing ΔE. *G*eneration and *R*ecombination transition rates increase exponentially as the energy separation becomes smaller than the bulk band gap. (c) At a sufficient gate voltage, the eigenstates align in energy ($\Delta E = 0$), and ideal band-to-band tunneling (BTBT) can occur.

the drain's communication with electrons in the channel is unimpeded, whereas communication with the source is restricted by the tunneling barrier. The drain bias is typically much larger than the thermal voltage ($V_D \gg kT/q$), and electrons in the channel are depleted by the drain as fast as they are supplied (assuming ballistic transport). If the tunneling probability from the source is low, the electron occupancy of the channel will also be low. The strength of Auger generation in an *n*-TFET does not depend on the precise electron occupancy, only that the electron concentration is low enough to prevent efficient recombination of the generated carriers. A related situation occurs for *p*-TFETs, where the hole concentration in the channel is decreased by the negative drain bias.

To achieve reasonable tunneling current and gate control, the thickness of the lightly doped layer must be quite thin (~10 nm), resulting in carrier-energy quantization. This quantization may or may not occur in the heavily doped layer, depending on its thickness. Previous work[18] suggested beneficial device performance for 2D-to-2D tunneling as compared to 2D-to-3D tunneling. In our work, both electron and hole energies are considered to be quantized in their respective layers, a treatment that can be extended to devices with thick layers by considering carriers to be quantized but with a very small energy separation between eigenstates. This approach allows our methodology to be applied to a wider variety of device structures.

IV. THEORY OF AUGER GENERATION AND RECOMBINATION

Auger transitions can occur among many different bands. The type of transition is denoted by a four-letter initialism that identifies the bands for the initial and final states of the two particles involved in the process. C, H, S, and L correspond to the Γ-*C*onduction, *H*eavy-hole, *S*plit-off hole, and *L*ight-hole bands. We focus on Auger generation for two cases: (1) the CHCC process, when a high-energy electron in the conduction band collides with an electron in the heavy-hole band, knocking it into the conduction band; and (2) the HCHH process, when an energetic hole collides with a valence electron, exciting it into the conduction band. These processes are schematically illustrated in Figure 1. In both processes, energy and momentum must be conserved. While we address CHCC and HCHH processes, other transitions, such as HCHS and HCHL, also occur. HCHS and HCHL processes occur at much higher rates than HCHH in most *p*-type bulk materials with HCHS dominating in materials with band gaps greater than or equal to the split-off energy ($E_G \geq \Delta$).[19,20] We investigate HCHH in this work due to its similarity with CHCC and as a means to establish the theoretical framework. Additional work is needed to compare HCHH, HCHL, and HCHS rates as a function of eigenstate energy separation.

The relative rates of the different transitions depend on specific parameters, such as the split-off energy, effective-mass, and quantization energy of the different bands. Our work is restricted to transitions between the first heavy-hole subband and the first Γ subband to demonstrate the key dependencies of the Auger transition rate. Transitions not considered in this work increase the transition rate but do not significantly alter our overall conclusion—that Auger transitions play a key role in determining the subthreshold behavior of TFETs.

The dominance of CHCC or HCHH in a specific device depends on the relative electron and hole concentrations in the channel. CHCC generation requires a reasonable electron concentration because multiple electrons are involved in the transition; HCHH generation requires a reasonable hole concentration. Low electron concentration in the *n*-TFET channel (due to the drain bias, explained in Section III) causes HCHH to dominate, whereas low hole concentration in *p*-TFETs causes CHCC to dominate. We derive rates for the CHCC and HCHH processes, but typically only one is dominant in a specific TFET structure due to the doping profile and the applied bias.

The expression for the net Auger recombination rate per unit area (U) is derived using Fermi's Golden Rule to give[21]

$$U = R - G = \frac{1}{A}\frac{2\pi}{\hbar}\sum_{1,\,1',\,2,\,2'} P(1,1',2,2')\,|M|^2\,\delta(E_1 - E_{1'} + E_2 - E_{2'}), \tag{1}$$

where $R - G$ is the difference between recombination and generation, and the sum is over all initial and final states including spin. Here, A is the in-plane area of the quantum well, $\hbar$ is the reduced Planck constant, P is the occupation probability function, M is the matrix element that couples the initial and final states, and $\delta(E_1 + E_2 - E_{1'} - E_{2'})$ ensures energy conservation. Components of the Auger rate are described in Sections IV A–IV C.

A. Occupation probability function

The driving force for net recombination is the occupation probability function (P) that gives the difference between the occupation of states required for the forward (recombination) and reverse (generation) processes. Following the diagram of Figure 1, P is expressed by

$$\begin{aligned} P_{CHCC}(1,1',2,2') &= f_c(E_1)f_c(E_2)f_v'(E_{1'})f_c'(E_{2'}) \\ &\quad - f_c'(E_1)f_c'(E_2)f_v(E_{1'})f_c(E_{2'}) \\ &\approx f_c(E_1)f_c(E_2)f_v'(E_{1'}) - f_c(E_{2'}), \end{aligned} \tag{2a}$$

$$\begin{aligned} P_{HCHH}(1,1',2,2') &= f_v'(E_1)f_v'(E_2)f_c(E_{1'})f_v(E_{2'}) \\ &\quad - f_v(E_1)f_v(E_2)f_c'(E_{1'})f_v'(E_{2'}) \\ &\approx f_v'(E_1)f_v'(E_2)f_c(E_{1'}) - f_v'(E_{2'}), \end{aligned} \tag{2b}$$

where

$$f_{c,v}(E) = \frac{1}{1 + \exp\dfrac{E - E_{Fn,p}}{kT}} \quad \text{and} \quad f_{c,v}'(E) = 1 - f_{c,v}(E), \tag{3}$$

f_c and f_v are the Fermi-Dirac distributions for a state to be occupied in the conduction and valence bands with quasi-Fermi levels E_{Fn} and E_{Fp}, respectively, and kT is the thermal energy. Similarly, f' is the probability for a state to be vacant.

The approximation to P in Eq. (2) assumes $f_c' \approx 1$ and $f_v \approx 1$, appropriate for non-degenerate carrier concentrations (discussed further below Eq. (6)). Equations (2a) and (2b) can be manipulated into their typical forms

$$P_{CHCC}(1,1',2,2') \approx \frac{n}{N_c}\left[\frac{np}{n_o p_o} - 1\right]\exp\left(-\frac{E_{2'} - E_c}{kT}\right), \quad (4a)$$

$$P_{HCHH}(1,1',2,2') \approx \frac{p}{N_v}\left[\frac{np}{n_o p_o} - 1\right]\exp\left(\frac{E_{2'} - E_v}{kT}\right), \quad (4b)$$

by employing Maxwell–Boltzmann statistics.[22] Here, n and p are the 2D densities of electrons and holes, and n_o and p_o are their values in quasi-equilibrium. [Quasi-equilibrium means $V_{DS} = 0$, but V_{GS} can vary.] N_c and N_v are the 2D densities of states of the conduction band and valence band, and E_c and E_v are the subband ground-state energies for the conduction and valence bands. [Equations (4) and (5) are also valid for 3D densities with the appropriate redefinition of variables from 2D to 3D.] The bracketed factor represents the difference between recombination and generation. In equilibrium $n = n_o$, $p = p_o$, and the net rate is zero. An np-product greater than equilibrium results in net recombination, whereas an np-product less than equilibrium results in net generation. When considering the impact of drain and source potentials, it is useful to note that

$$\frac{np}{n_o p_o} = \exp\left(\frac{E_{Fn} - E_{Fp}}{kT}\right). \quad (5)$$

For reverse bias, $np \ll n_o p_o$, and Eq. (4) can be written as

$$P_{CHCC}(1,1',2,2') \approx -\frac{n}{N_c}\exp\left(-\frac{E_{2'} - E_c}{kT}\right), \quad (6a)$$

$$P_{HCHH}(1,1',2,2') \approx -\frac{p}{N_v}\exp\left(\frac{E_{2'} - E_v}{kT}\right). \quad (6b)$$

The use of Eq. (6) makes the implicit assumptions that

(i) the carrier probability at the high-energy state $|2'\rangle$ is accurately modeled by non-degenerate statistics,
(ii) the reverse bias is sufficient to make recombination negligible, and
(iii) the joint probability of an electron in the valence band along with a vacant state in the conduction band is sufficiently high such that it does not limit the creation of an electron–hole pair.

Assumption (i) is reasonable in most cases because the high-energy state is typically far from the band-edge, making the use of Maxwell–Boltzmann statistics at $E_{2'}$ appropriate.

Assumption (ii) is valid given the conditions explained in Section III.

Assumption (iii) is justified by revisiting the non-degenerate approximation in Eq. (2): the generation term of P_{CHCC} is given by $f_c'(E_1)f_c'(E_2)f_v(E_{1'})f_c(E_{2'}) \approx f_c(E_{2'})$. In the non-degenerate limit, $f_c' \approx 1$ and $f_v \approx 1$, and the approximation results in little error. As the electron concentration increases towards the degenerate limit, $f_c(E_{2'})$ grows exponentially while $f_c'(E_1)$ and $f_c'(E_2)$ decrease slightly. Even at the extreme when the Fermi level is located at E_1 and E_2 [which assumes that $E_1 = E_2$, shown to be the condition for the most probable transition in Appendix C:], $f_c'(E_1) = f_c'(E_2) = \frac{1}{2}$, leading to only a 4× error; however, at the same time, $f_c(E_{2'})$ has grown exponentially by many orders of magnitude, dominating the characteristics for P. It is therefore reasonable to neglect f_c' and f_v factors for most cases.

B. Matrix element and wavefunctions

The square of the total matrix element for the Auger process is given by[23]

$$\underbrace{|M|^2}_{\text{total}} = \underbrace{|M_{12}|^2 + |M_{21}|^2}_{\text{opposite spins}} + \underbrace{|M_{12} - M_{21}|^2}_{\text{same spins}}, \quad (7)$$

where

$$M_{ij} = \int\int \Psi_{1'}^*(\boldsymbol{r_i})\Psi_{2'}^*(\boldsymbol{r_j})\frac{q^2}{4\pi\epsilon|\boldsymbol{r}_1 - \boldsymbol{r}_2|}\Psi_1(\boldsymbol{r}_1)\Psi_2(\boldsymbol{r}_2)\,d^3\boldsymbol{r}_1 d^3\boldsymbol{r}_2. \quad (8)$$

[The physics community often uses CGS units such that the denominator of Eq. (8) is missing the 4π-factor. Because our work uses SI units (with energy in eV), we include the 4π-factor.] $\Psi_n(\boldsymbol{r})$ is the wavefunction for state $|n\rangle$, q is the elementary charge, and ϵ is the permittivity of the semiconductor. (For the low-energy transitions analyzed in this work, the corresponding frequency is relatively low and use of the static permittivity is reasonable. See Refs. 24 and 25 for a more detailed discussion regarding the use of the static versus optical permittivity.) The bold notation indicates a vector quantity. The middle factor in Eq. (8) emphasizes the electron–electron Coulombic interaction between particles 1 and 2 that causes the Auger transition.

The total matrix element includes coupling due to colliding electrons with the same spins and opposite spins. M_{12} represents the transition from $|1, 2\rangle \rightarrow |1', 2'\rangle$, whereas M_{21} indicates $|1, 2\rangle \rightarrow |2', 1'\rangle$. When the particles have opposite spins, they are distinguishable, and no interference occurs. The contribution of opposite spin collisions to Eq. (7) is $|M_{12}|^2 + |M_{21}|^2$. When the electrons' spins are the same, the M_{12} and M_{21} processes interfere due to the exchange interaction resulting in the $|M_{12} - M_{21}|^2$ term in Eq. (7). (See Refs. 23 and 26 for further explanation.)

The wavefunctions for the conduction and valence bands are given by Bloch functions with confinement along the z-direction (perpendicular to the quantum well)[26]

$$\Psi_{c,\boldsymbol{k}}(\boldsymbol{r}) = \frac{1}{\sqrt{A}}u_{c,\boldsymbol{k}}(\boldsymbol{r})e^{i(k_x x + k_y y)}\psi_c(z) = \frac{1}{\sqrt{A}}\sum_{\boldsymbol{G}} a_{\boldsymbol{k},\boldsymbol{G}}e^{i\left((k_x+G_x)x+(k_y+G_y)y+G_z z\right)}\psi_c(z), \quad (9a)$$

$$\Psi_{v,\boldsymbol{k}}(\boldsymbol{r}) = \frac{1}{\sqrt{A}}u_{v,\boldsymbol{k}}(\boldsymbol{r})e^{i(k_x x + k_y y)}\psi_v(z) = \frac{1}{\sqrt{A}}\sum_{\boldsymbol{G}} b_{\boldsymbol{k},\boldsymbol{G}}e^{i\left((k_x+G_x)x+(k_y+G_y)y+G_z z\right)}\psi_v(z), \quad (9b)$$

where $\boldsymbol{k}$ is the wavevector, $u(\boldsymbol{r})$ is the cell-periodic function with periodicity of the crystal, $e^{i(k_x x+k_y y)}$ and $\psi(z)$ are the envelope wavefunctions along the in-plane and confined dimensions, and the summation is a Fourier series expansion of u in terms of $\boldsymbol{G}$, the set of all reciprocal lattice vectors.

Following the procedure detailed in Appendix A, the Auger matrix element for a quantum well is found to be

$$M_{12}^{QW} = \frac{q^2}{2\epsilon A}\frac{\delta_{\boldsymbol{k}_{\perp 1}-\boldsymbol{k}_{\perp 1'}+\boldsymbol{k}_{\perp 2}-\boldsymbol{k}_{\perp 2'}}}{|\boldsymbol{k}_{\perp 1}-\boldsymbol{k}_{\perp 1'}|}V, \tag{10}$$

where

$$V = \langle u_{1'}|u_1\rangle\langle u_{2'}|u_2\rangle\langle\psi_{1'}|\psi_1\rangle\langle\psi_{2'}|\psi_2\rangle. \tag{11}$$

Here,

$$\langle u_j|u_i\rangle \equiv \int_{\Omega_{cell}} u^*_{j,\boldsymbol{k}_j}(\boldsymbol{r})u_{i,\boldsymbol{k}_i}(\boldsymbol{r})d^3\boldsymbol{r}, \tag{12}$$

$$\langle\psi_j|\psi_i\rangle \equiv \int \psi^*_j(z)\psi_i(z)dz, \tag{13}$$

and Ω_{cell} is the volume of a unit cell. As shown in Appendix C, the most probable transition occurs when $\boldsymbol{k}_1 = \boldsymbol{k}_2$, which gives $M_{12} \approx M_{21}$. Because M_{12} and M_{21} are nearly equal, the squared matrix element of Eq. (7) can be approximated by

$$|M|^2 \approx 2|M_{12}|^2. \tag{14}$$

Physically, Eq. (14) suggests that only collisions between carriers of opposite spins contribute significantly to Auger transitions.

C. Wavefunction overlap

The Auger matrix element depends on the overlap integrals of the cell-periodic function u and the envelope function ψ. The overlap integrals within the same band are taken to be unity,[21] so that Eq. (11) can be simplified to

$$V \approx \langle u_{1'}|u_1\rangle\langle\psi_{1'}|\psi_1\rangle, \tag{15}$$

whereby the numbering convention of Figure 1 is followed, and states $|1\rangle$ and $|1'\rangle$ are defined to be in different bands.

The overlap of the periodic part of the Bloch functions $\langle u_{1'}|u_1\rangle$ is not straightforward to calculate. Burt *et al.*[25] show that the use of the effective-mass sum-rule method, often used to calculate $\langle u_{1'}|u_1\rangle$, is flawed, and instead compute the cell-periodic Bloch function overlap between the Γ and heavy-hole bands using the 15-band $k\cdot p$ and non-local pseudopotential methods. In a subsequent work,[27] the same authors calculate $\langle u_{1'}|u_1\rangle$ for additional III–V and II–VI direct-gap semiconductors. Both studies show strong agreement between the $k\cdot p$ and pseudopotential methods. Pseudopotential calculations[27] give similar results for III–V (InSb, GaSb, GaAs) and II–VI (CdTe, ZnSe) semiconductors, with the overlap for II–VIs approximately half that of III–Vs. For both sets of materials, the cell-periodic Bloch function overlap has a nearly linear dependence on the wavevector difference between the Γ and heavy-hole band ($K = |\boldsymbol{k}_1 - \boldsymbol{k}_{1'}|$) and is given by

$$|\langle u_{1'}|u_1\rangle| \approx c_u K, \tag{16}$$

where c_u equals $\sqrt{2\times10^{-17}}$ cm for III–V and $\sqrt{6\times10^{-18}}$ cm for II–VI semiconductors. Note that Eq. (16) agrees with the work of Verhulst *et al.*,[28] which highlights the zero-coupling between the Γ and heavy-hole band at zone-center (i.e., when $K = 0$).

The calculation of $\langle\psi_{1'}|\psi_1\rangle$ depends on the details of the structure. Specifically, the envelope wavefunction overlap is affected by

(i) *the effective mass of electrons and holes in the quantization direction*: As the effective mass decreases, the particle is less confined to its well and a large fraction of its wavefunction leaks into the band gap.

(ii) *the width of the quantum well*: As the quantum well widens, a larger separation distance between electron and hole wavefunctions occurs when a perpendicular electric field is applied to the quantum well. The wavefunctions separate in real space and $\langle\psi_{1'}|\psi_1\rangle$ decreases considerably. For the same reason, BTBT also decreases considerably as the distance between holes and electrons (i.e., the tunneling distance) increases.

(iii) *the doping profile in the tunneling region*: Doping modifies the energy-band diagram and thus the wavefunctions. Arbitrary doping profiles require a self-consistent calculation of the coupled Poisson–Schrödinger equations to determine the wavefunctions and the overlap integral.

For a structure with low doping concentration, such as the bilayer TFET[29] shown in Figure 3, the energy-band diagram is well-modeled by a uniform electric field across the tunneling junction. The exact wavefunctions for a quantum well with a uniform electric field have been solved analytically and shown by Ref. 30 to be a linear combination of Airy functions. In Figure 5, we calculate the wavefunction overlap $\langle\psi_{1'}|\psi_1\rangle$ for an InAs quantum well with a uniform electric field. $\langle\psi_{1'}|\psi_1\rangle$ at eigenstate energy alignment ($\Delta E = 0$) varies from approximately 0.01 to 0.1 as the quantum-well width decreases from 20 to 10 nm.

D. Auger rate equation

To arrive at a simplified form of the Auger rate equation, the terms derived in Secs. IV A–IV C are substituted into Eq. (1) to give

$$U_{CHCC} \approx \frac{1}{A}\frac{2\pi}{\hbar}\sum_{1,\,1',\,2,\,2'} -\frac{n}{N_c}\exp\left(-\frac{E_{2'}-E_c}{kT}\right) \times 2\left(\frac{q^2}{2\epsilon A}\frac{\delta_{\boldsymbol{k}_{\perp 1}-\boldsymbol{k}_{\perp 1'}+\boldsymbol{k}_{\perp 2}-\boldsymbol{k}_{\perp 2'}}}{|\boldsymbol{k}_{\perp 1}-\boldsymbol{k}_{\perp 1'}|}\right)^2 \times (c_u K)^2\,|\langle\psi_{1'}|\psi_1\rangle|^2\,\delta(E_1 - E_{1'} + E_2 - E_{2'}), \tag{17a}$$

$$U_{HCHH} \approx \frac{1}{A}\frac{2\pi}{\hbar}\sum_{1,\,1',\,2,\,2'} -\frac{p}{N_v}\exp\left(\frac{E_{2'}-E_v}{kT}\right) \times 2\left(\frac{q^2}{2\epsilon A}\frac{\delta_{\boldsymbol{k}_{\perp 1}-\boldsymbol{k}_{\perp 1'}+\boldsymbol{k}_{\perp 2}-\boldsymbol{k}_{\perp 2'}}}{|\boldsymbol{k}_{\perp 1}-\boldsymbol{k}_{\perp 1'}|}\right)^2 \times (c_u K)^2\,|\langle\psi_{1'}|\psi_1\rangle|^2\,\delta(E_1-E_{1'}+E_2-E_{2'}). \tag{17b}$$

Following the procedure in Appendix B, the summation over all states is converted to an integral that evaluates to

$$G_{CHCC} \approx \frac{q^4 {m_c}^3 (kT)^2 {c_u}^2}{4\pi^2\hbar^7\epsilon^2}\frac{(\mu+1)}{(2\mu+1)^2}|\langle\psi_{1'}|\psi_1\rangle|^2 \times \frac{n}{N_c}\exp\left(-\frac{(2\mu+1)}{(\mu+1)}\frac{\Delta E}{kT}\right), \tag{18a}$$

$$G_{HCHH} \approx \frac{q^4 {m_v}^3 (kT)^2 {c_u}^2}{4\pi^2\hbar^7\epsilon^2}\frac{(\mu^{-1}+1)}{(2\mu^{-1}+1)^2}|\langle\psi_{1'}|\psi_1\rangle|^2 \times \frac{p}{N_v}\exp\left(-\frac{(2\mu^{-1}+1)}{(\mu^{-1}+1)}\frac{\Delta E}{kT}\right), \tag{18b}$$

where G is the areal generation rate; ΔE is the energy separation between the first electron and hole subbands; $\mu = m_c/m_v$, and m_c and m_v are the density-of-states masses for the Γ and heavy-hole bands. [When using energy units of [eV], the q^4-factor in Eq. (18) is equal to $(1.602\ldots\times 10^{-19})^2$ [e^2 C^2], where C is Coulombs and e is the elementary unit of charge. Using this prescription, the q^2/ϵ-factor of the Coulombic potential is expressed in units of $\left[\frac{\text{e C}}{\text{F/cm}}\right] = [\text{eV cm}]$.]

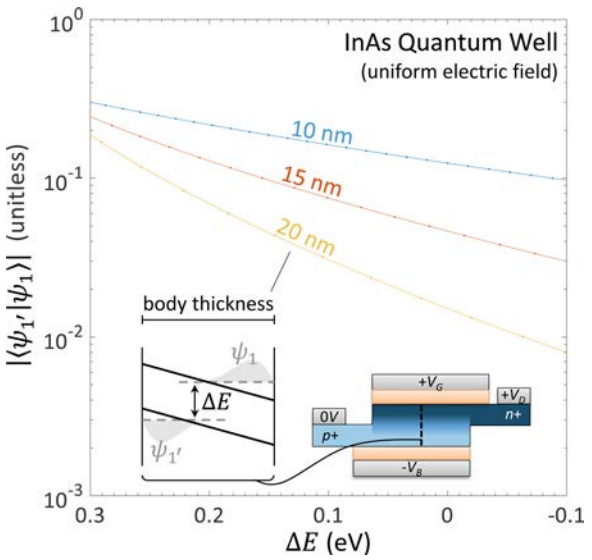

FIG. 5. Overlap integral of the electron and hole envelope wavefunctions as a function of ΔE (the energy separation between the first electron and hole eigenstates) for an InAs quantum well with a uniform electric field. An electric field is applied across the quantum well to decrease ΔE, which increases the distance between electron and hole distributions and hence decreases the envelope wavefunction overlap. Increasing the body thickness also leads to decreased overlap since the carriers are farther apart. The insets show a double-gated bilayer TFET structure with a uniform electric field across the channel. The corresponding energy-band diagram is shown to the left of the structure.

The exponential factor in the generation rate results from the Maxwell–Boltzmann probability of a carrier in the high-energy state $|2'\rangle$. For example, in the case of CHCC generation, there is a high probability for an empty state in the conduction band at $|1\rangle$ and $|2\rangle$ and an electron in the valence band at $|1'\rangle$, but a low probability of an electron at $|2'\rangle$ that ultimately limits the generation rate (see Figure 1). The minimum energy for $|2'\rangle$ that satisfies energy and momentum conservation is

$$\text{CHCC: } E_{2'_{\min}} = \frac{2\mu+1}{\mu+1}\Delta E, \tag{19a}$$

$$\text{HCHH: } E_{2'_{\min}} = \frac{2\mu^{-1}+1}{\mu^{-1}+1}\Delta E. \tag{19b}$$

Equation (19) is derived in Appendix C. By substituting Eq. (19) into (6), the origin of the exponential dependence of the generation rate can be readily understood.

E. Auger generation current

Because Auger generation occurs in the high-field region of a reverse-biased *pn* junction, nearly all the generated carriers are swept out to the contacts where they contribute to device current. With this assumption, the current-density due to generation can be expressed as

$$J_{gen} = qG, \tag{20}$$

where the sign of J_{gen} is defined to correspond with the sign of the drain current I_D.

If the material properties and structural details are known, the Auger generation current can be calculated using Eqs. (18) and (20). We define a generic material to provide an Auger generation current that is relevant to a wide variety of material and doping configurations. The generation current for a specific structure can be determined by multiplying the current for the generic material by relevant scaling factors.

Figure 6 shows a plot of the generation current-density as a function of ΔE for different ratios of the electron and hole mass for CHCC and HCHH processes. The exponential factor of Eq. (18) dominates the current characteristics, giving rise to room-temperature subthreshold swings $(d\Delta E/d\log_{10}(J_{gen}))$ between 30 and 60 meV/decade, depending on the mass ratio. The change in the envelope wavefunction overlap with ΔE (Figure 5) is significantly weaker than the exponential dependence of the Auger rate assuming unity overlap (Figure 6). Therefore, inclusion of a ΔE-dependent wavefunction overlap to Figure 6 will lower the calculated rate, but not significantly change the trends with ΔE.

To convert from ΔE to gate voltage, V_G, the gate efficiency of the device must be known. As described in Ref. 29, the incremental gate-voltage efficiency $(d\Delta E/d(qV_G))$ is poor for materials with a small effective mass. Increasing V_G to decrease ΔE (Figure 4(c)) increases the field across the junction leading to increased quantization that increases ΔE. For example, the expected gate efficiency for a TFET with a 15-nm InAs body and 1-nm effective oxide thickness is

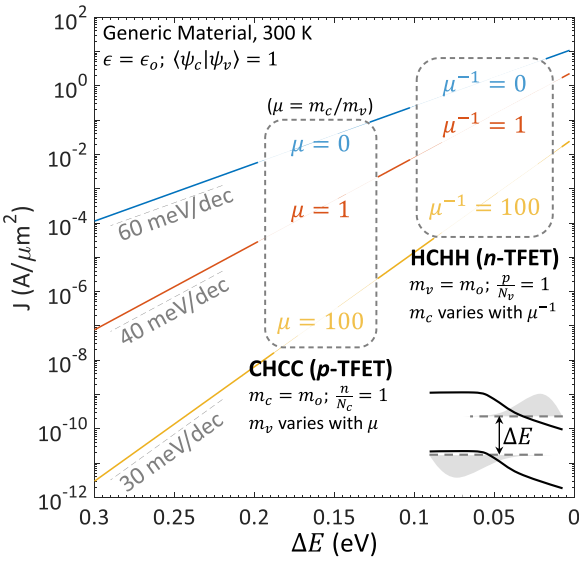

FIG. 6. Auger generation current-density as a function of ΔE (the energy separation between the first electron and hole eigenstates) for a generic material with parameters given in the plot. The curves are valid for both the CHCC and HCHH processes given the interpretation and conditions described in the figure. μ is the ratio of the Γ-band mass to the heavy-hole mass. The dashed lines show swings of 30, 40, and 60 meV/decade. Note that to estimate the conventional gate-voltage-controlled current turn-on swing, the gate efficiency $(d\Delta E/d(qV_G))$ should be included.

~40% (Ref. 29). Such a low gate efficiency will further degrade the subthreshold swing in these devices.

V. COMPARISON OF AUGER AND BTBT PROCESSES

Even though BTBT is a single-particle process while Auger requires two particles, BTBT and Auger transitions are not remarkably different phenomena. Both can be viewed as generation and recombination events, and Fermi's Golden Rule can be used to calculate the transition rate. Changing a device design to increase BTBT likely increases Auger transition rates. We believe that the relationship between Auger and BTBT is instructive for designing devices that minimize Auger and maximize BTBT.

Applying Fermi's Golden Rule to calculate the BTBT rate gives[18]

$$U_{BTBT} = R - G = \frac{1}{A}\frac{2\pi}{\hbar}\sum_{1,\,1'} P(1,1')\,|M|^2\,\delta(E_1 - E_{1'}). \quad (21)$$

The occupation probability function for BTBT (P) is the difference between recombination (where an electron in the conduction band tunnels to the valence band) and generation (where an electron in the valence band tunnels to the conduction band). P is given by

$$P_{BTBT}(1,1') = f_c(E_1)f_v'(E_{1'}) - f_c'(E_1)f_v(E_{1'}) \approx \frac{np}{n_o p_o} - 1; \quad \text{for } E_{Fn} < E_1 = E_{1'} < E_{Fp}, \quad (22)$$

where the approximation assumes Maxwell–Boltzmann statistics. For BTBT in a sufficiently reverse-biased pn junction, $np \ll n_o p_o$ and $P_{BTBT} \approx -1$. [P_{BTBT} is defined for net recombination so a negative value indicates net generation.]

The equations for the occupation probability function for BTBT (22) and Auger (4) are quite similar. The same driving force is seen in both functions, with the Auger function scaled by the probability of a carrier in the high-energy state $|2'\rangle$. This probability factor is crucial to the different dependencies of Auger and BTBT rates.

The matrix element for BTBT tunneling is given by

$$M_{BTBT} = \int \Psi_{1'}^*(\boldsymbol{r})\, q\phi(z)\, \Psi_1(\boldsymbol{r})\, d^3\boldsymbol{r}, \quad (23)$$

where $q\phi(z)$ is the potential energy in the depletion region of the pn junction, and Ψ_1 and $\Psi_{1'}$ are wavefunctions in the valence and conduction band. Following the derivation in Appendix D, the matrix element can be rewritten as

$$M_{BTBT} = (qF) z_{cv} \delta_{\boldsymbol{k}_{\perp 1} - \boldsymbol{k}_{\perp 1'}} \langle\psi_{1'}|\psi_1\rangle, \quad (24)$$

where

$$|z_{cv}|^2 = \frac{\hbar^2}{4 m_r E_G} \quad \text{and} \quad m_r = \frac{m_c m_v}{m_c + m_v}. \quad (25)$$

Here, a constant electric field (F) across the tunneling junction is assumed, $\delta_{\boldsymbol{k}_{\perp 1} - \boldsymbol{k}_{\perp 1'}}$ ensures perpendicular momentum conservation, E_G is the band gap including quantization energy, m_r is the reduced mass, and $z_{cv} = u_{1'}|id/dk_z|u_1$. The expression for z_{cv} in Eq. (25) results from Kane's two-band $k \cdot p$ calculations.[31]

The similarity between the matrix element for BTBT (24) and Auger (10) should be noted. Both require perpendicular momentum conservation and are proportional to the overlap of the envelope wavefunctions of the valence and conduction bands, $\langle\psi_{1'}|\psi_1\rangle$.

The BTBT rate is derived in Appendix E, and the areal generation rate is found to be

$$G_{BTBT} = \frac{(qF)^2}{2\hbar E_G} |\langle\psi_{1'}|\psi_1\rangle|^2; \quad \text{for } \Delta E \le 0. \quad (26)$$

Equation (26) looks quite different from typical expressions for BTBT because most of the details of the tunneling process are hidden in the $\langle\psi_{1'}|\psi_1\rangle$ factor. Using this equation, the *intrinsic* on/off ratio between the BTBT and Auger rates at the point of ideal TFET turn-on ($\Delta E = 0$) is found to be

$$\frac{G_{BTBT}}{G_{CHCC}} = \frac{(qF)^2}{E_G} \frac{2\pi^2\hbar^6\epsilon^2}{q^4 {m_c}^3 (kT)^2 {c_u}^2} \frac{(2\mu + 1)^2}{(\mu + 1)} \frac{N_c}{n}, \quad (27a)$$

$$\frac{G_{BTBT}}{G_{HCHH}} = \frac{(qF)^2}{E_G} \frac{2\pi^2\hbar^6\epsilon^2}{q^4 {m_v}^3 (kT)^2 {c_u}^2} \frac{(2\mu^{-1} + 1)^2}{(\mu^{-1} + 1)} \frac{N_v}{p}. \quad (27b)$$

[N_c, N_v, n, and p are given by their 2D values.] This ratio is plotted as a function of the electric field in Figure 7. The ratio depends only on fundamental constants, material parameters, and the electric field and doping of the structure. It is technology-independent, without the effects of band tails, selection rules, defect states, and other non-idealities. While the turn-on of BTBT is often envisioned as a sharp

jump from near zero current to its on-state value,[18] Eq. (27) quantifies this jump in current and sets limitations to device performance.

For an ideal TFET, an infinitesimal change in gate voltage creates a huge change in drain current, which translates to a huge on/off ratio. Auger generation, however, increases the off-current, thereby reducing the on/off ratio. Equation (27) provides the maximum possible on/off ratio for an infinitesimal change in a TFET's gate voltage. Because we look only at Auger transitions between the first eigenstates of the heavy-hole and Γ band, the actual on/off ratio will be even smaller than the value calculated from Eq. (27). Overall, Auger generation prevents ideal TFET operation because it decreases the on/off ratio, complicating circuit design and increasing off-state power consumption, which reduce the energy-efficiency and limit the potential applications for TFETs.

VI. DISCUSSION

The Arrhenius dependence of Auger-generation current on the eigenstate energy separation is a challenge for researchers seeking a steep switching transistor. The turn-on of BTBT is intrinsically linked to the turn-on of Auger generation, which (unfortunately) is thermally dependent. Efforts to improve the steepness of BTBT may be unproductive if Auger generation dominates the subthreshold characteristics of the device. Therefore, Auger generation (along with other non-ideal effects) must be modeled in the subthreshold region of TFET characteristics to capture its negative impact on device performance and resolve the discrepancy between simulation and experimental results.

Auger generation is difficult to mitigate given that it is an intrinsic phenomenon. Decreasing the wavefunction overlap to decrease Auger generation will also, undesirably, decrease the BTBT current. The best approach to decrease Auger generation is to decrease the carrier concentration of the heavily doped source; however, a reduction in the source doping will decrease the field across the junction, decreasing the tunneling rate. These difficulties require careful device design to minimize the impact of Auger generation.

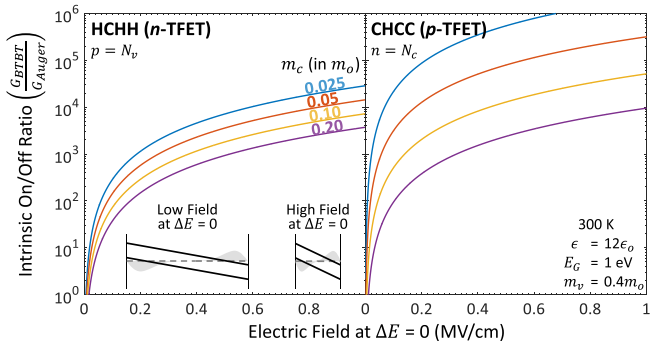

FIG. 7. Intrinsic on/off ratio of the BTBT and Auger rates at eigenstate alignment ($\Delta E = 0$) as a function of the electric field, calculated from Eq. (27). The BTBT rate decreases dramatically as the field decreases, and therefore, the ratio drops. The permittivity (ϵ) and heavy-hole mass (m_v) do not vary significantly among materials; hence, constant values indicated on the plot are used. A 1-eV band gap is also assumed. The ratio depends linearly on $1/E_G$ so decreasing the band gap by half will double the on/off ratio. The CHCC process (dominant in *p*-TFETs with high *n*-doping) gives a much better on/off ratio because the Auger generation rate is much lower for the CHCC process due to the light electron mass. The inset shows the energy-band diagram for two structures with different body thicknesses at $\Delta E = 0$. The thinner structure requires a higher electric field to align the bands, which results in an improved on/off ratio due to increased BTBT at high fields. The electric field at $\Delta E = 0$ will also be dependent on the doping profile and electrostatics of the device, in addition to body thickness.

One straightforward method to reduce Auger generation is to use a *p*-TFET device design with CHCC as the dominant Auger process. As shown in Figure 7, *p*-TFETs yield higher on/off ratios and lower Auger currents compared to *n*-TFETs because CHCC generation depends on the Γ-band mass to the third power ($m_c{}^3$), which is quite small for many materials.

Our work focuses on perpendicular TFETs (i.e., tunneling perpendicular to the gate), but point TFETs (with tunneling parallel to the gate) are also affected by Auger generation. The area over which Auger generation is prevalent is approximately equal to the tunneling area of the device. For a perpendicular TFET, the tunneling area is equal to the gate area. For a point TFET, the tunneling area is less easily defined but can be considered approximately equal to the width of the device multiplied by the thickness over which tunneling occurs. Our approach provides a rough estimate of the Auger current in point TFETs, but a more detailed analysis is needed to address the dependence of Auger generation on device dimensionality.

Auger generation presents a novel method of achieving sub-thermal subthreshold swings—the minimum room-temperature subthreshold swing for an Auger process is 30 mV/decade (assuming perfect gate efficiency), when the mass of one carrier is many times heavier than the other (i.e., either μ or $\mu^{-1} \rightarrow \infty$). For many materials, the hole mass is much heavier than the electron mass which leads to a large μ^{-1}. We propose a new device concept—the Auger FET—that uses the steep slope of Auger generation as a switching mechanism. Such a device may look similar to a perpendicular TFET; however, the structure would be optimized to increase Auger rather than BTBT, such as by increasing the doping concentration of the source. More analysis is needed to characterize the Auger FET's potential for achieving high currents.

ACKNOWLEDGMENTS

This work was supported in part by the Center for Energy Efficient Electronics Science (NSF Award No. 0939514). The authors thank Mathieu Luisier, Roger Lake, and Rebecca Murray for their technical comments regarding the manuscript.

APPENDIX A: DERIVATION OF THE AUGER MATRIX ELEMENT

The final form of the Auger matrix element for a quantum well given in Eq. (10) can be derived from Eq. (8), which gives the standard expression for a matrix element resulting from a Coulombic potential between two particles. Before deriving the matrix element for a quantum well, it is helpful to first analyze the matrix element for a bulk crystal.

1. Auger matrix element for a bulk crystal

The conduction and valence band wavefunctions for a bulk three-dimensional crystal can be written as Bloch functions[26]

$$\Psi_{c,\boldsymbol{k}}(\boldsymbol{r}) = \frac{1}{\sqrt{\Omega}} u_{c,\boldsymbol{k}}(\boldsymbol{r}) e^{i\boldsymbol{k}\cdot\boldsymbol{r}} = \frac{1}{\sqrt{\Omega}} \sum_{\boldsymbol{G}} a_{\boldsymbol{k},\boldsymbol{G}} e^{i(\boldsymbol{G}+\boldsymbol{k})\cdot\boldsymbol{r}}, \quad \text{(A1a)}$$

$$\Psi_{v,\boldsymbol{k}}(\boldsymbol{r}) = \frac{1}{\sqrt{\Omega}} u_{v,\boldsymbol{k}}(\boldsymbol{r}) e^{i\boldsymbol{k}\cdot\boldsymbol{r}} = \frac{1}{\sqrt{\Omega}} \sum_{\boldsymbol{G}} b_{\boldsymbol{k},\boldsymbol{G}} e^{i(\boldsymbol{G}+\boldsymbol{k})\cdot\boldsymbol{r}}. \quad \text{(A1b)}$$

Here, Ω is the volume of crystal, $\boldsymbol{k}$ is a three-dimensional wavevector, and other parameters have the same definitions as in Eq. (9). Substituting the wavefunctions for a bulk crystal into Eq. (8) gives

$$M_{12}^{bulk} = \frac{1}{\Omega^2} \int\int \sum_{\boldsymbol{G}_1,\boldsymbol{G}_{1'},\boldsymbol{G}_2,\boldsymbol{G}_{2'}} a_{\boldsymbol{k}_1,\boldsymbol{G}_1} b^*_{\boldsymbol{k}_{1'},\boldsymbol{G}_{1'}} a_{\boldsymbol{k}_2,\boldsymbol{G}_2} a^*_{\boldsymbol{k}_{2'},\boldsymbol{G}_{2'}} \times e^{i(\boldsymbol{k}_1-\boldsymbol{k}_{1'}+\boldsymbol{G}_1-\boldsymbol{G}_{1'})\cdot\boldsymbol{r}_1} e^{i(\boldsymbol{k}_2-\boldsymbol{k}_{2'}+\boldsymbol{G}_2-\boldsymbol{G}_{2'})\cdot\boldsymbol{r}_2} \times \frac{q^2}{4\pi\epsilon|\boldsymbol{r}_1-\boldsymbol{r}_2|} d^3\boldsymbol{r}_1 d^3\boldsymbol{r}_2. \quad \text{(A2)}$$

Making use of the substitution $\boldsymbol{r}_1 = \boldsymbol{r}_{12} + \boldsymbol{r}_2$, the integral can be rewritten as

$$M_{12}^{bulk} = \frac{q^2}{4\pi\epsilon\Omega^2} \int\int \ldots e^{i(\boldsymbol{k}_1-\boldsymbol{k}_{1'}+\boldsymbol{G}_1-\boldsymbol{G}_{1'})\cdot\boldsymbol{r}_{12}} \times e^{i(\boldsymbol{k}_1-\boldsymbol{k}_{1'}+\boldsymbol{k}_2-\boldsymbol{k}_{2'}+\boldsymbol{G}_1-\boldsymbol{G}_{1'}+\boldsymbol{G}_2-\boldsymbol{G}_{2'})\cdot\boldsymbol{r}_2} \times \frac{1}{|\boldsymbol{r}_{12}|} d^3\boldsymbol{r}_{12} d^3\boldsymbol{r}_2. \quad \text{(A3)}$$

Integrating over $\boldsymbol{r}_2$ and making use of the identity that

$$\lim_{\Omega\to\infty} \frac{1}{\Omega} \int_{\Omega} e^{i\boldsymbol{k}\cdot\boldsymbol{r}} d^3\boldsymbol{r} = \delta_{\boldsymbol{k}}, \quad \text{(A4)}$$

(where $\delta_{\boldsymbol{k}}$ is the Kronecker delta) yields

$$M_{12}^{bulk} = \frac{q^2}{4\pi\epsilon\Omega} \int \ldots \delta_{\boldsymbol{k}_1-\boldsymbol{k}_{1'}+\boldsymbol{k}_2-\boldsymbol{k}_{2'}+\boldsymbol{G}_1-\boldsymbol{G}_{1'}+\boldsymbol{G}_2-\boldsymbol{G}_{2'}} \times e^{i(\boldsymbol{k}_1-\boldsymbol{k}_{1'}+\boldsymbol{G}_1-\boldsymbol{G}_{1'})\cdot\boldsymbol{r}_{12}} \frac{1}{|\boldsymbol{r}_{12}|} d^3\boldsymbol{r}_{12}. \quad \text{(A5)}$$

To evaluate the integral over $\boldsymbol{r}_{12}$, the integral is transformed into spherical coordinates defined such that the vector $\boldsymbol{k}_1 - \boldsymbol{k}_{1'} + \boldsymbol{G}_1 - \boldsymbol{G}_{1'}$ lies along $\theta = 0$. This transformation allows the integral to be expressed as

$$\int e^{i(\boldsymbol{k}_1-\boldsymbol{k}_{1'}+\boldsymbol{G}_1-\boldsymbol{G}_{1'})\cdot\boldsymbol{r}_{12}} \frac{1}{|\boldsymbol{r}_{12}|} d^3\boldsymbol{r}_{12} = \int\int\int e^{i|\boldsymbol{k}_1-\boldsymbol{k}_{1'}+\boldsymbol{G}_1-\boldsymbol{G}_{1'}|\rho_{12}\cos\theta} \frac{1}{\rho_{12}} \rho_{12}^2 \sin\theta d\phi d\theta d\rho_{12} = \frac{4\pi}{|\boldsymbol{k}_1-\boldsymbol{k}_{1'}+\boldsymbol{G}_1-\boldsymbol{G}_{1'}|^2}, \quad \text{(A6)}$$

where $\rho_{12} = |\boldsymbol{r}_{12}|$. Putting this all together gives

$$M_{12}^{bulk} = \frac{q^2}{\epsilon\Omega} \sum_{\boldsymbol{G}_1,\boldsymbol{G}_{1'},\boldsymbol{G}_2,\boldsymbol{G}_{2'}} a_{\boldsymbol{k}_1,\boldsymbol{G}_1} b^*_{\boldsymbol{k}_{1'},\boldsymbol{G}_{1'}} a_{\boldsymbol{k}_2,\boldsymbol{G}_2} a^*_{\boldsymbol{k}_{2'},\boldsymbol{G}_{2'}} \times \frac{\delta_{\boldsymbol{k}_1-\boldsymbol{k}_{1'}+\boldsymbol{k}_2-\boldsymbol{k}_{2'}+\boldsymbol{G}_1-\boldsymbol{G}_{1'}+\boldsymbol{G}_2-\boldsymbol{G}_{2'}}}{|\boldsymbol{k}_1-\boldsymbol{k}_{1'}+\boldsymbol{G}_1-\boldsymbol{G}_{1'}|^2}. \quad \text{(A7)}$$

Given the unlikelihood of Umklapp processes and the requirement of momentum conversation,[26] terms in which $\boldsymbol{G}_1 \neq \boldsymbol{G}_{1'}$ and $\boldsymbol{G}_2 \neq \boldsymbol{G}_{2'}$ are neglected. This allows the matrix element to be written as

$$M_{12}^{bulk} = \frac{q^2}{\epsilon\Omega} \sum_{\boldsymbol{G}_1} a_{\boldsymbol{k}_1,\boldsymbol{G}_1} b^*_{\boldsymbol{k}_{1'},\boldsymbol{G}_1} \sum_{\boldsymbol{G}_2} a_{\boldsymbol{k}_2,\boldsymbol{G}_2} a^*_{\boldsymbol{k}_{2'},\boldsymbol{G}_2} \frac{\delta_{\boldsymbol{k}_1-\boldsymbol{k}_{1'}+\boldsymbol{k}_2-\boldsymbol{k}_{2'}}}{|\boldsymbol{k}_1-\boldsymbol{k}_{1'}|^2}, \quad \text{(A8)}$$

which can be simplified to

$$M_{12}^{bulk} = \frac{q^2}{\epsilon\Omega} \langle u_{1'}|u_1\rangle \langle u_{2'}|u_2\rangle \frac{\delta_{\boldsymbol{k}_1-\boldsymbol{k}_{1'}+\boldsymbol{k}_2-\boldsymbol{k}_{2'}}}{|\boldsymbol{k}_1-\boldsymbol{k}_{1'}|^2}, \quad \text{(A9)}$$

where

$$\langle u_{1'}|u_1\rangle \equiv \langle u_{v,\boldsymbol{k}_{1'}}|u_{c,\boldsymbol{k}_1}\rangle = \int_{\Omega_{cell}} u^*_{v,\boldsymbol{k}_{1'}}(\boldsymbol{r}) u_{c,\boldsymbol{k}_1}(\boldsymbol{r}) d^3\boldsymbol{r} = \sum_{\boldsymbol{G}} a_{\boldsymbol{k}_1,\boldsymbol{G}} b^*_{\boldsymbol{k}_{1'},\boldsymbol{G}}, \quad \text{(A10a)}$$

$$\langle u_{2'}|u_2\rangle \equiv \langle u_{c,\boldsymbol{k}_{2'}}|u_{c,\boldsymbol{k}_2}\rangle = \int_{\Omega_{cell}} u^*_{c,\boldsymbol{k}_{2'}}(\boldsymbol{r}) u_{c,\boldsymbol{k}_2}(\boldsymbol{r}) d^3\boldsymbol{r} = \sum_{\boldsymbol{G}} a_{\boldsymbol{k}_2,\boldsymbol{G}} a^*_{\boldsymbol{k}_{2'},\boldsymbol{G}}. \quad \text{(A10b)}$$

2. Auger matrix element for a quantum well

The Auger matrix element for a quantum well can be derived following a similar procedure as used for the bulk crystal. Substituting the quantum-well wavefunctions from Eq. (9) into (8) gives

$$M_{12}^{QW} = \frac{1}{A^2} \int\int \sum_{\boldsymbol{G}_1,\boldsymbol{G}_{1'},\boldsymbol{G}_2,\boldsymbol{G}_{2'}} a_{\boldsymbol{k}_1,\boldsymbol{G}_1} b^*_{\boldsymbol{k}_{1'},\boldsymbol{G}_{1'}} a_{\boldsymbol{k}_2,\boldsymbol{G}_2} a^*_{\boldsymbol{k}_{2'},\boldsymbol{G}_{2'}} \times e^{i(\boldsymbol{k}_{\perp 1}-\boldsymbol{k}_{\perp 1'}+\boldsymbol{G}_{\perp 1}-\boldsymbol{G}_{\perp 1'})\cdot\boldsymbol{r}_{\perp 1}} \times e^{i(\boldsymbol{k}_{\perp 2}-\boldsymbol{k}_{\perp 2'}+\boldsymbol{G}_{\perp 2}-\boldsymbol{G}_{\perp 2'})\cdot\boldsymbol{r}_{\perp 2}} e^{i(G_{z1}-G_{z1'})\cdot z_1} e^{i(G_{z2}-G_{z2'})\cdot z_2} \times \psi_1(z_1)\psi^*_{1'}(z_1)\psi_2(z_2)\psi^*_{2'}(z_2) \frac{q^2}{4\pi\epsilon|\boldsymbol{r}_1-\boldsymbol{r}_2|} d^3\boldsymbol{r}_1 d^3\boldsymbol{r}_2, \quad \text{(A11)}$$

where $\boldsymbol{k}_\perp$ refers to the two-dimensional in-plane wavevector (perpendicular to the tunneling direction), $\boldsymbol{r}_{\boldsymbol{i}} = x_i + y_i + z_i$, $\boldsymbol{r}_{\perp \boldsymbol{i}} = x_i + y_i$, and other parameters are defined below Eq. (9). Making use of the substitution $\boldsymbol{r}_{\perp 1} = \boldsymbol{r}_{\perp 12} + \boldsymbol{r}_{\perp 2}$, the integral can be rewritten as

$$M_{12}^{QW} = \frac{q^2}{4\pi\epsilon A^2}\int\int \sum_{\boldsymbol{G}_1,\boldsymbol{G}_{1'},\boldsymbol{G}_2,\boldsymbol{G}_{2'}} a_{\boldsymbol{k}_1,\boldsymbol{G}_1} b^*_{\boldsymbol{k}_{1'},\boldsymbol{G}_{1'}} a_{\boldsymbol{k}_2,\boldsymbol{G}_2} a^*_{\boldsymbol{k}_{2'},\boldsymbol{G}_{2'}} \times e^{i(\boldsymbol{k}_{\perp 1}-\boldsymbol{k}_{\perp 1'}+\boldsymbol{G}_{\perp 1}-\boldsymbol{G}_{\perp 1'})\cdot\boldsymbol{r}_{\perp 12}} \times e^{i(\boldsymbol{k}_{\perp 1}-\boldsymbol{k}_{\perp 1'}+\boldsymbol{k}_{\perp 2}-\boldsymbol{k}_{\perp 2'}+\boldsymbol{G}_{\perp 1}-\boldsymbol{G}_{\perp 1'}+\boldsymbol{G}_{\perp 2}-\boldsymbol{G}_{\perp 2'})\cdot\boldsymbol{r}_{\perp 2}} \times e^{i(G_{z1}-G_{z1'})\cdot z_1} e^{i(G_{z2}-G_{z2'})\cdot z_2} \times \psi_1(z_1)\,\psi^*_{1'}(z_1)\,\psi_2(z_2)\,\psi^*_{2'}(z_2) \times \frac{1}{\sqrt{\boldsymbol{r}_{\perp 12}^2+z_{12}^2}} d^2\boldsymbol{r}_{\perp 12} d^2\boldsymbol{r}_{\perp 2} dz_1 dz_2, \tag{A12}$$

where $z_{12} = z_1 - z_2$. Integrating over $\boldsymbol{r}_{\perp 2}$ and making use of the identity that

$$\lim_{A\to\infty}\frac{1}{A}\int_A e^{i\boldsymbol{k}_\perp\cdot\boldsymbol{r}_\perp} d^2\boldsymbol{r}_\perp = \delta_{\boldsymbol{k}_\perp}, \tag{A13}$$

yields

$$M_{12}^{QW} = \frac{q^2}{4\pi\epsilon A}\int \ldots \delta_{\boldsymbol{k}_{\perp 1}-\boldsymbol{k}_{\perp 1'}+\boldsymbol{k}_{\perp 2}-\boldsymbol{k}_{\perp 2'}+\boldsymbol{G}_{\perp 1}-\boldsymbol{G}_{\perp 1'}+\boldsymbol{G}_{\perp 2}-\boldsymbol{G}_{\perp 2'}} \times e^{i(\boldsymbol{k}_{\perp 1}-\boldsymbol{k}_{\perp 1'}+\boldsymbol{G}_{\perp 1}-\boldsymbol{G}_{\perp 1'})\cdot\boldsymbol{r}_{\perp 12}} e^{i(G_{z1}-G_{z1'})\cdot z_1} e^{i(G_{z2}-G_{z2'})\cdot z_2} \times \psi_1(z_1)\,\psi^*_{1'}(z_1)\,\psi_2(z_2)\,\psi^*_{2'}(z_2) \times \frac{1}{\sqrt{\boldsymbol{r}_{\perp 12}^2+z_{12}^2}} d^2\boldsymbol{r}_{\perp 12} dz_1 dz_2. \tag{A14}$$

To evaluate the integral over $\boldsymbol{r}_{\perp 12}$, the integral is transformed into circular coordinates defined such that the vector $\boldsymbol{k}_{\perp 1} - \boldsymbol{k}_{\perp 1'} + \boldsymbol{G}_{\perp 1} - \boldsymbol{G}_{\perp 1'}$ lies along $\theta = 0$. Most of the contribution to the integral over $\boldsymbol{r}_{\perp 12}$ in Eq. (A14) occurs when z_{12} is small. Therefore, the square root factor is approximated as $|\boldsymbol{r}_{\perp 12}|$. The integral over $\boldsymbol{r}_{\perp 12}$ can now be evaluated as follows:

$$\begin{aligned}\int e^{i(\boldsymbol{k}_{\perp 1}-\boldsymbol{k}_{\perp 1'}+\boldsymbol{G}_{\perp 1}-\boldsymbol{G}_{\perp 1'})\cdot\boldsymbol{r}_{\perp 12}}\frac{1}{|\boldsymbol{r}_{\perp 12}|} d^2\boldsymbol{r}_{\perp 12} &= \int\int e^{i|\boldsymbol{k}_{\perp 1}-\boldsymbol{k}_{\perp 1'}+\boldsymbol{G}_{\perp 1}-\boldsymbol{G}_{\perp 1'}|\rho_{12}\cos\theta}\frac{1}{\rho_{12}}\rho_{12} d\theta d\rho_{12} \\ &= \frac{2\pi}{|\boldsymbol{k}_{\perp 1}-\boldsymbol{k}_{\perp 1'}+\boldsymbol{G}_{\perp 1}-\boldsymbol{G}_{\perp 1'}|},\end{aligned} \tag{A15}$$

where $\rho_{12} = |\boldsymbol{r}_{\perp 12}|$. Combining Eqs. (A14) and (A15) and neglecting Umklapp processes gives

$$M_{12}^{QW} = \frac{q^2}{2\epsilon A}\sum_{\boldsymbol{G}_1} a_{\boldsymbol{k}_1,\boldsymbol{G}_1} b^*_{\boldsymbol{k}_1,\boldsymbol{G}_1}\sum_{\boldsymbol{G}_2} a_{\boldsymbol{k}_2,\boldsymbol{G}_2} a^*_{\boldsymbol{k}_2,\boldsymbol{G}_2}\langle\psi_{1'}|\psi_1\rangle\langle\psi_{2'}|\psi_2\rangle \times \frac{\delta_{\boldsymbol{k}_{\perp 1}-\boldsymbol{k}_{\perp 1'}+\boldsymbol{k}_{\perp 2}-\boldsymbol{k}_{\perp 2'}}}{|\boldsymbol{k}_{\perp 1}-\boldsymbol{k}_{\perp 1'}|}. \tag{A16}$$

This expression can then be simplified to Eq. (10) by using the definition of $\langle u_j|u_i\rangle$ provided in Eq. (A10).

APPENDIX B: DERIVATION OF THE AUGER TRANSITION RATE FOR A QUANTUM WELL

To derive the overall Auger transition rate, we begin with the summation in Eq. (1) and the foresight that the matrix element will include a Kronecker delta factor enforcing conservation of momentum. The sum over all states can be transformed into an integral by making use of property that there is one k-state for a reciprocal-space area of $(2\pi)^2/A$

$$\sum_{1,\,1',\,2,\,2'}\delta_{\boldsymbol{k}_{\perp 1}-\boldsymbol{k}_{\perp 1'}+\boldsymbol{k}_{\perp 2}-\boldsymbol{k}_{\perp 2'}} \cdot 2\left(\frac{A}{(2\pi)^2}\right)^3\int d^2\boldsymbol{k}_{\perp 1} d^2\boldsymbol{k}_{\perp 2} d^2\boldsymbol{k}_{\perp 1'}. \tag{B1}$$

The initial factor of 2 accounts for the spin-up and spin-down configurations of the initial states. (The matrix element already accounts for the cases in which the initial states have the same or opposite spins as explained in Section IV B.) The Kronecker delta is used to reduce the sum over four states to an integral over three states since $\boldsymbol{k}_{\perp 2'} = \boldsymbol{k}_{\perp 1} - \boldsymbol{k}_{\perp 1'} + \boldsymbol{k}_{\perp 2}$.

Rewriting Eq. (1) as an integral produces

$$U = \frac{2}{A}\frac{2\pi}{\hbar}\left(\frac{A}{(2\pi)^2}\right)^3\int P(1,1',2,2')2|M_{12}^{QW}(\boldsymbol{k}_{\perp 1}-\boldsymbol{k}_{\perp 1'})|^2 \times \delta(E) d^2\boldsymbol{k}_{\perp 1} d^2\boldsymbol{k}_{\perp 2} d^2\boldsymbol{k}_{\perp 1'}. \tag{B2}$$

Equation (14) has been used to replace the square of the matrix element with $2|M_{12}|^2$. The notation $M_{12}^{QW}(\boldsymbol{k}_{\perp 1}-\boldsymbol{k}_{\perp 1'})$ signifies that M_{12}^{QW} is a function of the wavevector difference between $|1\rangle$ and $|1'\rangle$. $M_{12}^{QW}(\boldsymbol{k}_{\perp 1}-\boldsymbol{k}_{\perp 1'})$ is given by Eq. (10), but the Kronecker delta factor has already been used in the transformation performed in Eq. (B1). For brevity, $\delta(E_1 - E_{1'} + E_2 - E_{2'})$ has been replaced with $\delta(E)$. Substituting Eq. (6a) for $P(1,1',2,2')$ in the above equation gives

$$U \approx -\frac{1}{A}\frac{8\pi}{\hbar}\left(\frac{A}{(2\pi)^2}\right)^3\frac{n}{N_c}Q, \tag{B3}$$

where

$$Q = \int \exp\left(-\frac{E_{2'}-E_c}{kT}\right)|M_{12}^{QW}(\boldsymbol{K})|^2\delta(E) d^2\boldsymbol{K} d^2\boldsymbol{k}_{\perp 2} d^2\boldsymbol{k}_{\perp 1'}, \tag{B4}$$

and $\boldsymbol{K} = \boldsymbol{k}_{\perp 1} - \boldsymbol{k}_{\perp 1'}$. The above expression is valid for the CHCC process under the assumptions detailed in Section IV A.

Much of subsequent procedure follows the work of Ref. 22. It is reproduced here for completeness. To evaluate the integrals of Eq. (B4), $\boldsymbol{K}$ is expressed in terms of polar coordinates (K,θ), and $\boldsymbol{k}_{\perp 1'}$ and $\boldsymbol{k}_{\perp 2}$ are expressed in Cartesian coordinates [denoted $(x_{1'}, y_{1'})$ and (x_2, y_2)] such that $y_{1'}$ and y_2 lie along $\boldsymbol{K}$. The energy for state $|2'\rangle$ can be written as

$$E_{2'} = E_c + \alpha\boldsymbol{k}_{2'}{}^2 \tag{B5a}$$

$$= E_c + \alpha(\boldsymbol{k}_1 - \boldsymbol{k}_{1'} + \boldsymbol{k}_2)^2 \tag{B5b}$$

$$= E_c + \alpha(\boldsymbol{K} + \boldsymbol{k}_2)^2 \tag{B5c}$$

$$= E_c + \alpha(x_2 + (K + y_2))^2, \tag{B5d}$$

where $\alpha = \hbar^2/2m_c$.To provide clarity, these expressions use the abuse of notation that the square of a vector implies the dot product with itself, i.e., $\boldsymbol{k}^2 \equiv \boldsymbol{k}\cdot\boldsymbol{k}$. The statement of energy conservation can be re-expressed as

$$E = E_1 - E_{1'} + E_2 - E_{2'} \tag{B6a}$$

$$= \Delta E + \alpha(\boldsymbol{k_1}^2 + \mu\boldsymbol{k_{1'}}^2 + \boldsymbol{k_2}^2 - \boldsymbol{k_{2'}}^2) \tag{B6b}$$

$$= \Delta E + \alpha((\mu+1)\boldsymbol{k_{1'}}^2 + 2(\boldsymbol{k_1} - \boldsymbol{k_{1'}})(\boldsymbol{k_{1'}} - \boldsymbol{k_2})) \tag{B6c}$$

$$= \Delta E + \alpha((\mu+1)\boldsymbol{k_{1'}}^2 + 2\boldsymbol{K}(\boldsymbol{k_{1'}} - \boldsymbol{k_2})) \tag{B6d}$$

$$= \Delta E + \alpha((\mu+1)({x_{1'}}^2 + {y_{1'}}^2) + 2K(y_{1'} - y_2)). \tag{B6e}$$

By going from (B6b) to (B6c), momentum conservation was used to make the substitution $\boldsymbol{k_{2'}} = \boldsymbol{k_1} - \boldsymbol{k_{1'}} + \boldsymbol{k_2}$. These statements allow Q to be rewritten as

$$Q = \int \exp\left(-\frac{\alpha(x_2 + (K+y_2))^2}{kT}\right)|M_{12}^{QW}(\boldsymbol{K})|^2 \times \delta(E)K d\theta dK dx_{1'} dy_{1'} dx_2 dy_2, \tag{B7}$$

where E is given by (B6e). In this work, M is assumed to be independent of θ, which makes the integration over θ trivial. In reality, M has a θ-dependence that should be accounted for in future work. The integral over x_2 is straightforward

$$\int_{-\infty}^{\infty} \exp\left(-\frac{\alpha {x_2}^2}{kT}\right)dx_2 = 2\int_0^{\infty} \exp\left(-\frac{\alpha {x_2}^2}{kT}\right)dx_2 = \sqrt{\frac{\pi kT}{\alpha}}. \tag{B8}$$

This gives

$$Q = 2\pi\sqrt{\frac{\pi kT}{\alpha}}\int \exp\left(-\frac{\alpha(K+y_2)^2}{kT}\right)|M_{12}^{QW}(K)|^2 \times \delta(E)KdKdx_{1'}dy_{1'}dy_2. \tag{B9}$$

The integral over $x_{1'}$ is evaluated using the property that

$$\delta(g(x)) = \sum_i \frac{\delta(x - x_i)}{|g'(x_i)|}, \quad \{x_i \,|\, g(x_i) = 0\}, \tag{B10}$$

to give

$$\int_{-\infty}^{\infty} \delta\big(\alpha(\mu+1){x_{1'}}^2 - D\big)dx_{1'} = \frac{1}{\sqrt{D\alpha(\mu+1)}}, \tag{B11}$$

where

$$-D = \alpha(\mu+1){y_{1'}}^2 + 2\alpha K(y_{1'} - y_2) + \Delta E. \tag{B12}$$

For the integral over $x_{1'}$ to be non-zero, $D > 0$ since $x_{1'}$ is real. Q is now given by

$$Q = 2\pi\sqrt{\frac{\pi kT}{\alpha}}\frac{1}{\sqrt{\alpha(\mu+1)}}\int \frac{1}{\sqrt{D}} \times \exp\left(-\frac{\alpha(K+y_2)^2}{kT}\right)|M_{12}^{QW}(K)|^2 KdKdy_{1'}dy_2. \tag{B13}$$

Next, the integral over $y_{1'}$ is evaluated under the condition that $D > 0$. This requires $y_{1'}$ to be within the range of

$${y_{1'}}_{\min}^{\max} = \frac{-2\alpha K \pm \sqrt{4\alpha^2K^2 - 4\alpha(\mu+1)(\Delta E - 2\alpha K y_2)}}{2\alpha(\mu+1)}. \tag{B14}$$

Using these limits, the integral can now be evaluated

$$\int_{y_{1'\min}}^{y_{1'\max}} \frac{1}{i\sqrt{-D}}dy_{1'} = \frac{\ln(-1)}{i\sqrt{\alpha(\mu+1)}} = \frac{\pi}{\sqrt{\alpha(\mu+1)}}. \tag{B15}$$

Q can now be written as

$$Q = 2\pi^2\sqrt{\frac{\pi kT}{\alpha}}\frac{1}{\alpha(\mu+1)} \times \int \exp\left(-\frac{\alpha(K+y_2)^2}{kT}\right)|M_{12}^{QW}(K)|^2 KdKdy_2. \tag{B16}$$

Since $y_{1'}$ must be real, y_2 must satisfy

$$y_2 \geq y_{2\min} = \frac{\Delta E}{2\alpha K} - \frac{K}{2(\mu+1)}, \tag{B17}$$

which is found from evaluating the square root term in Eq. (B14). The variable substitution

$$u = \sqrt{\frac{\alpha}{kT}}(K + y_2), \tag{B18}$$

is used to compute the integral over y_2

$$\int_{u_{\min}=\sqrt{\frac{\alpha}{kT}}(K+y_{2\min})}^{\infty} \exp(-u^2)\sqrt{\frac{kT}{\alpha}}du = \sqrt{\frac{kT}{\alpha}}\frac{\sqrt{\pi}}{2}\operatorname{erfc}\left(\sqrt{\frac{\alpha}{kT}}(K + y_{2\min})\right), \tag{B19}$$

where erfc is the complimentary error function. Q can now be given as

$$Q = \frac{\pi^3 kT}{\alpha^2(\mu+1)} \times \int_0^{\infty} \operatorname{erfc}\left(\sqrt{\frac{\alpha}{kT}}\left(\frac{\Delta E}{2\alpha K} + \frac{(2\mu+1)}{2(\mu+1)}K\right)\right)|M_{12}^{QW}(K)|^2 KdK. \tag{B20}$$

The complimentary error function is peaked at the K-value where its argument is a minimum, which occurs when

$$K = K_o \equiv \sqrt{\frac{(\mu+1)}{(2\mu+1)}\frac{\Delta E}{\alpha}}. \tag{B21}$$

For $\Delta E \gg kT$, the complimentary error function varies much more quickly than $|M_{12}^{QW}(K_o)|^2$, allowing the matrix element factor to be moved outside the integral to produce

$$Q \approx \frac{\pi^3 kT}{\alpha^2(\mu+1)} |M_{12}^{QW}(K_o)|^2 \times \int_0^\infty \mathrm{erfc}\left(\sqrt{\frac{\alpha}{kT}}\left(\frac{\Delta E}{2\alpha K} + \frac{(2\mu+1)}{2(\mu+1)}K\right)\right) K dK. \quad \text{(B22)}$$

The integration over K can be performed using Eq. (4.3.34) from Ref. 32 to find

$$Q \approx \frac{\pi^3 (kT)^2}{\alpha^3} \frac{(\mu+1)}{(2\mu+1)^2} |M_{12}^{QW}(K_o)|^2 \exp\left(-\frac{(2\mu+1)}{(\mu+1)}\frac{\Delta E}{kT}\right). \quad \text{(B23)}$$

While the above equation is only strictly true for $\Delta E \gg kT$ due to the approximation made in Eq. (B22), numerical calculations suggest that Eq. (B23) overestimates the exact integral of Eq. (B20) by less than a factor of 2, even when $\Delta E = kT$. Plugging Q back into Eq. (B3) gives

$$U \approx -A^2 \frac{{m_c}^3 (kT)^2}{\pi^2 \hbar^7} \frac{(\mu+1)}{(2\mu+1)^2} |M_{12}^{QW}(K_o)|^2 \frac{n}{N_c} \times \exp\left(-\frac{(2\mu+1)}{(\mu+1)}\frac{\Delta E}{kT}\right). \quad \text{(B24)}$$

Substituting the expression for $M_{12}^{QW}(K_o)$ and using the linear relationship between $\langle u_{1'}|u_1\rangle$ and K given by Eq. (16) produces

$$U \approx -\frac{{m_c}^3 (kT)^2}{\pi^2 \hbar^7} \frac{(\mu+1)}{(2\mu+1)^2} \left(\frac{q^2}{2\epsilon}\right)^2 {c_u}^2 |\langle \psi_{1'}|\psi_1\rangle|^2 \frac{n}{N_c} \times \exp\left(-\frac{(2\mu+1)}{(\mu+1)}\frac{\Delta E}{kT}\right). \quad \text{(B25)}$$

In this expression, the $1/K$-dependence of the matrix element cancels with the K-dependence of the cell-periodic Bloch function overlap. This expression can then be simplified to the form given by Eq. (18a) in the paper. A similar procedure can be used to derive the net recombination rate for the HCHH process.

APPENDIX C: DERIVATION OF THE MINIMUM ENERGY FOR STATE $|2'\rangle$

As discussed in Section IV, the probability of a carrier in the high-energy state $|2'\rangle$ limits the overall Auger generation rate. The probability is highest when $|2'\rangle$ is at its minimum energy ($E_{2'\mathrm{min}}$). The following derivation is performed in one dimension for simplicity but can be expanded to multiple dimensions arriving at the same results. To find $E_{2'\mathrm{min}}$, it is useful to begin with the equations for energy and momentum conservation

$$E_1 - E_{1'} + E_2 - E_{2'} = 0 \quad \text{(C1)}$$

$$k_1 - k_{1'} + k_2 - k_{2'} = 0. \quad \text{(C2)}$$

The subscripts refer to the states shown in Figure 1, and by convention, $|2'\rangle$ is defined to be the high-energy state of the Auger process. For the CHCC process, Eq. (C1) can be rewritten as

$$E_{2'} = E_1 - E_{1'} + E_2 = \Delta E + \frac{\hbar^2}{2m_c}\left[{k_1}^2 + \mu {k_{1'}}^2 + {k_2}^2\right], \quad \text{(C3)}$$

where $\mu = m_c/m_v$. Rearranging Eq. (C2) as $k_2 = -k_1 + k_{1'} + k_{2'}$ and substituting into Eq. (C3) gives

$$E_{2'} = \Delta E + \frac{\hbar^2}{2m_c}[2{k_1}^2 + (\mu+1){k_{1'}}^2 + {k_{2'}}^2 - 2k_1k_{1'} - 2k_1k_{2'} + 2k_{1'}k_{2'}]. \quad \text{(C4)}$$

In the above expression, k_1 and $k_{1'}$ are the only independent variables because $k_{2'}$ can be rewritten in terms of $E_{2'}$. The minimum energy can be found when the gradient of Eq. (C4) is zero

$$\nabla E_{2'} = 0 = \frac{\partial E_{2'}}{\partial k_1}\widehat{k_1} + \frac{\partial E_{2'}}{\partial k_{1'}}\widehat{k_{1'}}, \quad \text{(C5a)}$$

$$\Rightarrow \begin{cases} \dfrac{\partial E_{2'}}{\partial k_1} = 0 = 4k_1 - 2k_{1'} - 2k_{2'} \\ \dfrac{\partial E_{2'}}{\partial k_{1'}} = 0 = -2k_1 + 2(\mu+1)k_{1'} + 2k_{2'}. \end{cases} \quad \text{(C5b)}$$

Solving this linear system of equations gives

$$k_1 = \frac{\mu}{2\mu+1}k_{2'}; \quad k_{1'} = \frac{-1}{2\mu+1}k_{2'}, \quad \text{(C6)}$$

when state $|2'\rangle$ is at its minimum energy. Combining the above with Eq. (C2) gives

$$k_2 = k_1, \quad \text{(C7)}$$

which shows that the wavevectors for $|1\rangle$ and $|2\rangle$ must be equal for the minimum energy condition. Substituting Eq. (C6) into (C4) and simplifying the expression gives

$$E_{2'_{\mathrm{min}}} = \frac{2\mu+1}{\mu+1}\Delta E, \quad \text{(C8)}$$

valid for the CHCC process, and a similar procedure can be used to find the minimum energy for the HCHH process.

APPENDIX D: DERIVATION OF THE BTBT MATRIX ELEMENT

To derive the BTBT matrix element in Eq. (24), we begin with the standard expression in Eq. (23). Assuming a constant electric field (F) across the tunneling junction, the potential $\phi(z)$ can be given as $\phi(z) = \phi_o + Fz$. The constant ϕ_o term results in zero coupling since $\langle \Psi_{1'}|\phi_o|\Psi_1\rangle = \phi_o\langle \Psi_{1'}|\Psi_1\rangle$ and the eigenstates of the system are orthogonal to one another. This allows the BTBT matrix element of Eq. (23) to be written as

$$M_{BTBT} = \frac{qF}{A}\int \left(\Psi_{1',k_{1'}}^{QW}(r)\right)^* z \left(\Psi_{1,k_1}^{QW}(r)\right) d^3\boldsymbol{r} \quad \text{(D1a)}$$

$$= \frac{qF}{A}\int \left(e^{i\boldsymbol{k}_{\perp 1'}\cdot\boldsymbol{r}_\perp}\psi_{1'}(z)u_{v,\boldsymbol{k}_{1'}}(\boldsymbol{r})\right)^* z\left(e^{i\boldsymbol{k}_{\perp 1}\cdot\boldsymbol{r}_\perp}\psi_1(z)u_{c,\boldsymbol{k}_1}(\boldsymbol{r})\right)d^3\boldsymbol{r}, \tag{D1b}$$

with the same definitions as in Eq. (9). Here, $\hat{z}$ is the tunneling direction, $\perp$ indicates directions perpendicular to tunneling, and the integral is over the entire volume of the device.

The confined envelope wavefunction, $\psi(z)$, can be expressed as a Fourier series of plane waves,

$$\psi_n(z) = \sum_{k_z} f_n(k_z)e^{ik_z z}, \tag{D2}$$

where $f_n(k_z)$ are the Fourier coefficients for state n. This allows Eq. (D1a) to be expressed as

$$M_{BTBT} = \frac{qF}{A}\sum_{k_{z1},k_{z1'}} f^*_{1'}(k_{z1'})f_1(k_{z1})\int \left(e^{i\boldsymbol{k}_{1'}\cdot\boldsymbol{r}}u_{v,\boldsymbol{k}_{1'}}(\boldsymbol{r})\right)^* \times z\left(e^{i\boldsymbol{k}_1\cdot\boldsymbol{r}}u_{c,\boldsymbol{k}_1}(\boldsymbol{r})\right)d^3\boldsymbol{r} \tag{D3a}$$

$$= \frac{qF}{A}\sum_{k_{z1},k_{z1'}} f^*_{1'}(k_{z1'})f_1(k_{z1})\int \left(\Psi^{bulk}_{1',\boldsymbol{k}_{1'}}(\boldsymbol{r})\right)^* \times z\left(\Psi^{bulk}_{1,\boldsymbol{k}_1}(\boldsymbol{r})\right)d^3\boldsymbol{r}, \tag{D3b}$$

where $\Psi^{bulk}_{n,\boldsymbol{k}}$ is the Bloch wavefunction for state n and wavevector $\boldsymbol{k}$ in a bulk crystal. Blount[33] shows that the integral in Eq. (D3b) is given by

$$\int \left(\Psi^{bulk}_{1',\boldsymbol{k}_{1'}}(\boldsymbol{r})\right)^* z\left(\Psi^{bulk}_{1,\boldsymbol{k}_1}(\boldsymbol{r})\right)d^3\boldsymbol{r} = \frac{-i\partial}{\partial k_{z1}}\int \left(\Psi^{bulk}_{1',\boldsymbol{k}_{1'}}(\boldsymbol{r})\right)^* \left(\Psi^{bulk}_{1,\boldsymbol{k}_1}(\boldsymbol{r})\right)d^3\boldsymbol{r} + \int e^{i(\boldsymbol{k}_1-\boldsymbol{k}_{1'})\cdot\boldsymbol{r}}u^*_{v,\boldsymbol{k}_{1'}}(\boldsymbol{r})\frac{i\partial u_{c,\boldsymbol{k}_1}(\boldsymbol{r})}{\partial k_{z1}}d^3\boldsymbol{r}, \tag{D4}$$

which can be proven by expanding the partial derivative of the first term using the chain rule and cancelling opposing terms. The first term of Eq. (D4) evaluates to zero because the wavefunctions of different bands are orthogonal. This leaves only the second term. The cell-periodic part of the Bloch function varies much quicker than the envelope wavefunctions, allowing the expression to be written as

$$\int e^{i(\boldsymbol{k}_1-\boldsymbol{k}_{1'})\cdot\boldsymbol{r}}u^*_{v,\boldsymbol{k}_{1'}}(\boldsymbol{r})\frac{i\partial u_{c,\boldsymbol{k}_1}(\boldsymbol{r})}{\partial k_{z1}}d^3\boldsymbol{r} = \sum_{\boldsymbol{R_m}} e^{i(\boldsymbol{k}_1-\boldsymbol{k}_{1'})\cdot\boldsymbol{R_m}}\int_{\Omega_{cell}} u^*_{v,\boldsymbol{k}_{1'}}(\boldsymbol{r})\frac{i\partial u_{c,\boldsymbol{k}_1}(\boldsymbol{r})}{\partial k_{z1}}d^3\boldsymbol{r}, \tag{D5}$$

where $\boldsymbol{R_m}$ is the set of all lattice positions and Ω_{cell} is the volume a unit cell. Using two-band $k\cdot p$ theory, Kane[31] shows that

$$z_{cv} \equiv \frac{1}{\Omega_{cell}}\int_{\Omega_{cell}} u^*_{v,\boldsymbol{k}_{1'}}(\boldsymbol{r})\frac{i\partial u_{c,\boldsymbol{k}_1}(\boldsymbol{r})}{\partial k_{z1}}d^3\boldsymbol{r} = \frac{i\hbar}{2\sqrt{m_r E_G}}, \tag{D6}$$

where E_G is the band gap including the quantization energy and m_r is the reduced mass equal to

$$m_r = \frac{m_c m_v}{m_c + m_v}. \tag{D7}$$

Combining Eqs. (D4) through (D6) gives

$$\int (\Psi^{bulk}_{1',\boldsymbol{k}_{1'}}(\boldsymbol{r}))^* z(\Psi^{bulk}_{1,\boldsymbol{k}_1}(\boldsymbol{r}))d^3\boldsymbol{r} = z_{cv}\Omega_{cell}\sum_{\boldsymbol{R_m}} e^{i(\boldsymbol{k}_1-\boldsymbol{k}_{1'})\cdot\boldsymbol{R_m}} \tag{D8a}$$

$$= z_{cv}\int e^{i(\boldsymbol{k}_1-\boldsymbol{k}_{1'})\cdot\boldsymbol{r}}d^3\boldsymbol{r}. \tag{D8b}$$

Substituting Eq. (D8b) into (D3b) yields

$$M_{BTBT} = \frac{(qF)z_{cv}}{A}\sum_{k_{z1},k_{z1'}} f^*_{1'}(k_{z1'})f_1(k_{z1})\int e^{i(\boldsymbol{k}_1-\boldsymbol{k}_{1'})\cdot\boldsymbol{r}}d^3\boldsymbol{r} \tag{D9a}$$

$$= \frac{(qF)z_{cv}}{A}\int e^{i(\boldsymbol{k}_{\perp 1}-\boldsymbol{k}_{\perp 1'})\cdot\boldsymbol{r}}d^2\boldsymbol{r}_\perp \int \sum_{k_{z1},k_{z1'}} \left(f_{1'}(k_{z1'})e^{ik_{z1'}z}\right)^* \times \left(f_1(k_{z1})e^{ik_{z1}z}\right)dz \tag{D9b}$$

$$= (qF)z_{cv}\delta_{\boldsymbol{k}_{\perp 1}-\boldsymbol{k}_{\perp 1'}}\int \psi^*_{1'}(z)\psi_1(z)dz \tag{D9c}$$

$$= (qF)z_{cv}\delta_{\boldsymbol{k}_{\perp 1}-\boldsymbol{k}_{\perp 1'}}\langle\psi_{1'}|\psi_1\rangle. \tag{D9d}$$

Equations (A13) and (D2) are used to go from (D9b) to (D9c). The final form of the BTBT matrix element in Eq. (D9d) is quite simple as it hides most of the tunneling physics in the $\langle\psi_{1'}|\psi_1\rangle$ factor. This simplified form is especially convenient for comparison to Auger processes, as shown in Section V.

APPENDIX E: DERIVATION OF THE BTBT RATE FOR A QUANTUM WELL

To calculate the BTBT transition rate, we begin by converting the sum over all states of Eq. (21) to an integral using the property that

$$\sum_{1,\,1'}\delta_{\boldsymbol{k}_{\perp 1}-\boldsymbol{k}_{\perp 1'}} \;\rightarrow\; 2\left(\frac{A}{(2\pi)^2}\right)\int d^2\boldsymbol{k}_{\perp 1}. \tag{E1}$$

The initial factor of two accounts for the spin degeneracy. Rewriting Eq. (21) using (E1) produces

$$U_{BTBT} = \frac{1}{A}\frac{2\pi}{\hbar}2\left(\frac{A}{(2\pi)^2}\right)\int P(1,1')\,|M|^2\,\delta(E_1 - E_{1'})d^2\boldsymbol{k}_{\perp 1}. \tag{E2}$$

Using the values for P and M given in Section V, the areal generation rate is found to be

$$G_{BTBT} = \frac{1}{A}\frac{2\pi}{\hbar}2\left(\frac{A}{(2\pi)^2}\right)(qF)^2|z_{cv}|^2|\langle\psi_{1'}|\psi_1\rangle|^2 \times \int \delta(E_1 - E_{1'})d^2\boldsymbol{k}_{\perp 1}. \tag{E3}$$

Assuming spherical bands, the integral can be rewritten as

$$\int d^2\boldsymbol{k}_{\perp 1} \rightarrow \int 2\pi k_{\perp 1} dk_{\perp 1}, \tag{E4}$$

to give

$$G_{BTBT} = \frac{2}{\hbar}(qF)^2 |z_{cv}|^2 |\langle \psi_{1'} | \psi_1 \rangle|^2 \int \delta(E_1 - E_{1'}) k_{\perp 1} dk_{\perp 1}. \tag{E5}$$

The energy difference in the Dirac delta function can be expressed as

$$E_1 - E_{1'} = (E_{z1} + E_{\perp 1}) - (E_{z1'} - E_{\perp 1'}) \tag{E6a}$$

$$= \Delta E + \frac{\hbar^2 k_{\perp 1}^2}{2m_c} + \frac{\hbar^2 k_{\perp 1'}^2}{2m_v} \tag{E6b}$$

$$= \Delta E + \frac{\hbar^2 k_{\perp 1}^2}{2m_r} \equiv \Delta E + E_{\perp}. \tag{E6c}$$

Using the variable substitution that

$$dE_{\perp} = \frac{\hbar^2}{m_r} k_{\perp 1} dk_{\perp 1}, \tag{E7}$$

the integral of Eq. (E5) can be expressed as

$$\int \frac{m_r}{\hbar^2} \delta(\Delta E + E_{\perp}) dE_{\perp} = \frac{m_r}{\hbar^2}; \quad \text{for } \Delta E \leq 0. \tag{E8}$$

Putting this all together, we get

$$G_{BTBT} = \frac{2m_r}{\hbar^3}(qF)^2 |z_{cv}|^2 |\langle \psi_{1'} | \psi_1 \rangle|^2; \quad \text{for } \Delta E \leq 0. \tag{E9}$$

Equation (26) can be then found by substituting the expression from (D4) for z_{cv}.

[1]K. Bernstein, R. K. Cavin, W. Porod, A. Seabaugh, and J. Welser, Proc. IEEE **98**, 2169 (2010).
[2]D. E. Nikonov and I. A. Young, Proc. IEEE **101**, 2498 (2013).
[3]D. E. Nikonov and I. A. Young, IEEE J. Explor. Solid-State Comput. Devices Circuits **1**, 3 (2015).
[4]H. Lu and A. Seabaugh, IEEE J. Electron Devices Soc. **2**, 44 (2014).
[5]S. Mookerjea, D. Mohata, T. Mayer, V. Narayanan, and S. Datta, IEEE Electron Device Lett. **31**, 564 (2010).
[6]T. Yu, U. Radhakrishna, J. L. Hoyt, and D. A. Antoniadis, IEEE Int. Electron Devices Meet. **2015**, 22.4.1–22.4.4.
[7]A. M. Walke, A. Vandooren, R. Rooyackers, D. Leonelli, A. Hikavyy, R. Loo, A. S. Verhulst, K.-H. Kao, C. Huyghebaert, G. Groeseneken, V. R. Rao, K. K. Bhuwalka, M. M. Heyns, N. Collaert, and A. V.-Y. Thean, IEEE Trans. Electron Devices **61**, 707 (2014).
[8]U. E. Avci, B. Chu-Kung, A. Agrawal, G. Dewey, V. Le, R. Rios, D. H. Morris, S. Hasan, R. Kotlyar, J. Kavalieros, and I. A. Young, IEEE Int. Electron Devices Meet. **2015**, 34.5.1–34.5.4.
[9]J. T. Teherani, W. Chern, S. Agarwal, J. L. Hoyt, and D. A. Antoniadis, in *2015 Fourth Berkeley Symposium on Energy Efficient Electronic Systems E3S* (2015), pp. 1–3.
[10]X. Zhao, A. Vardi, and J. A. del Alamo, IEEE Int. Electron Devices Meet. **2014**, 25.5.1–25.5.4.
[11]G. Dewey, B. Chu-Kung, J. Boardman, J. M. Fastenau, J. Kavalieros, R. Kotlyar, W. K. Liu, D. Lubyshev, M. Metz, N. Mukherjee, P. Oakey, R. Pillarisetty, M. Radosavljevic, H. W. Then, and R. Chau, IEEE Int. Electron Devices Meet. **2011**, 33.6.1–33.6.4.
[12]M. Kim, Y. Wakabayashi, R. Nakane, M. Yokoyama, M. Takenaka, and S. Takagi, IEEE Int. Electron Devices Meet. **2014**, 13.2.1–13.2.4.
[13]M. A. Kinch, *Fundamentals of Infrared Detector Materials (SPIE Tutorial Text Vol. TT76)* (SPIE Publications, 2007).
[14]M. A. Kinch, M. J. Brau, and A. Simmons, J. Appl. Phys. **44**, 1649 (1973).
[15]W. E. Tennant, J. Electron. Mater. **39**, 1030 (2010).
[16]W. E. Tennant, D. Lee, M. Zandian, E. Piquette, and M. Carmody, J. Electron. Mater. **37**, 1406 (2008).
[17]H. Yuan, M. Meixell, J. Zhang, P. Bey, J. Kimchi, and L. C. Kilmer, Proc. SPIE **8353**, 835309 (2012).
[18]S. Agarwal and E. Yablonovitch, e-print arXiv:1109.0096.
[19]M. Takeshima, J. Appl. Phys. **43**, 4114 (1972).
[20]N. K. Dutta and R. J. Nelson, J. Appl. Phys. **53**, 74 (1982).
[21]C. Smith, R. A. Abram, and M. G. Burt, J. Phys. C: Solid State Phys. **16**, L171 (1983).
[22]C. Smith, R. A. Abram, and M. G. Burt, Superlattices Microstruct. **1**, 119 (1985).
[23]B. K. Ridley, *Quantum Processes in Semiconductors*, 4th ed. (Oxford University Press, New York, 2000).
[24]A. Haug, J. Phys. Chem. Solids **49**, 599 (1988).
[25]M. G. Burt, S. Brand, C. Smith, and R. A. Abram, J. Phys. C: Solid State Phys. **17**, 6385 (1984).
[26]A. R. Beattie and P. T. Landsberg, Proc. R. Soc. London, Ser. A **249**, 16 (1959).
[27]S. Brand, M. G. Burt, C. Smith, and R. A. Abram, in *Proceedings of the 17th International Conference Phys. Semiconductors*, edited by J. D. Chadi and W. A. Harrison (Springer, New York, 1985), pp. 1013–1016.
[28]A. S. Verhulst, D. Verreck, M. A. Pourghaderi, M. V. de Put, B. Sorée, G. Groeseneken, N. Collaert, and A. V.-Y. Thean, Appl. Phys. Lett. **105**, 43103 (2014).
[29]J. T. Teherani, S. Agarwal, E. Yablonovitch, J. L. Hoyt, and D. A. Antoniadis, IEEE Electron Device Lett. **34**, 298 (2013).
[30]C. T. Giner and J. López Gondar, Phys. BC **138**, 287 (1986).
[31]E. Kane, J. Phys. Chem. Solids **12**, 181 (1960).
[32]E. W. Ng and M. Geller, J. Res. Natl. Bur. Stand. **73B**, 1 (1968).
[33]E. I. Blount, in *Solid State Physics*, edited by F. S. Turnbull and D. Turnbull (Academic Press, 1962), pp. 305–373.