Origins of Terahertz Difference Frequency Susceptibility in Midinfrared Quantum Cascade Lasers

Benjamin A. Burnett and Benjamin S. Williams

Phys. Rev. Applied **5**, 034013 — Published 30 March 2016

DOI: 10.1103/PhysRevApplied.5.034013

Origins of terahertz difference frequency susceptibility in mid-infrared quantum cascade lasers

Benjamin A. Burnett* and Benjamin S. Williams
Department of Electrical Engineering and California NanoSystems Institute,
University of California, Los Angeles, Los Angeles, California 90095, USA

(Dated: January 24, 2016)

We present a density matrix-based transport model applicable to quantum cascade lasers which computes both linear and nonlinear optical properties coherently and nonperturbatively. The model is applied to a dual active-region mid-infrared quantum cascade laser which generates terahertz radiation at the difference frequency between two mid-infrared pumps. A new mechanism for terahertz generation is identified as self-detection, ascribed to the beating of current flow following the intensity, associated with stimulated emission. This mechanism peaks at optical rectification but exhibits a bandwidth reaching significantly into the terahertz range, which is primarily limited by the sub-picosecond intersubband lifetimes. A metric is derived to assess the strength of self-detection in candidate active regions through experiment alone, and suggestions are made for improvement of the performance at frequencies below 2 THz.

I. INTRODUCTION

The quantum cascade laser (QCL) has emerged as a leading candidate for a coherent light source in both the terahertz (THz) and mid-infrared (mid-IR) spectral ranges. Mid-IR QCLs were demonstrated first, and have since advanced to watt-level powers in continuous-wave operation at room temperature, enabling widespread commercialization and applications [1, 2]. THz QCLs, on the other hand, have been the greater challenge and despite tremendous effort have reached a limit in operating temperature around 200 K, attributed to thermally-activated LO-phonon scattering and other sources [3–8]. The temperature limitation has been one major hindrance to the commercialization of these devices which could surely find applications in a diverse range of scientific and engineering fields [9, 10].

The only successful approach so far to room-temperature THz output in QCLs has been to harness the power of mid-IR QCLs for nonlinear, rather than direct, generation. Devices have been developed in which two mid-IR QCL active regions share the same cavity, and the THz difference frequency between two mid-IR pumps is generated through a second-order nonlinear susceptibility ($\chi^{(2)}$) that originates within the active region itself. In this way, various groups have demonstrated milliwatt-level peak power, microwatt-level average power, and tunability from 1.7-5.25 THz, all at room temperature [11–15].

Given these successes, it is now worthwhile to take a close look at the origin of the difference-frequency susceptibility. The mechanism is typically described as a resonant interaction between the upper radiative subband and two lower subbands of the active region/injector system (see, for example, the highlighted subbands in Fig. 1). Association with the lasing transition boasts the inherent advantage that population inversion prevents the pump absorption that typically accompanies resonant nonlinearities. However, it is somewhat of an oversimplification to attribute the nonlinearity to only a few subbands, given the large number of subbands in the injector region. A more complete theoretical analysis was undertaken using an Ensemble Monte Carlo method to calculate the steady-state populations under lasing conditions; a perturbative "sum-over-states" (SOS) expression was then applied to calculate $\chi^{(2)}$ considering all possible subband combinations [16–18]. Inclusion of all terms in the SOS expression lent agreement to the notion that $\chi^{(2)}$ is dominated by the resonant processes around the lasing transition, at least for biases near the injection resonance.

However, use of the SOS expression for $\chi^{(2)}$ is still not a complete description for a number of reasons. First, it does not properly treat the effects of permanent dipoles (diagonals of the position operator): these effects are important in QCLs where the state separations are comparable to the dipole elements, and inclusion of these terms in the SOS expression results in an unphysical translational variance. Second, since the SOS expressions are perturbative, they cannot naturally account for high-field effects such as electromagnetically-induced transparency and others that require higher order. Third, as the SOS expression is intended for a finite-sized entity such as an atom or a molecule, it cannot capture the full dynamics in an extended system such as a QCL with a large number of repeated modules that have conduction currents flowing between them. A visualization of the various processes which might contribute to THz generation is given in Fig. 1.

In this paper, we employ a density matrix-based transport model to provide a new look at the origin of the THz difference-frequency susceptibility, taking as an example system a dual active region mid-IR QCL reported in the literature [14]. The model is translationally invariant, nonperturbative, and takes into account the periodic nature of the active region. No *a priori* distinction is made

* bburnett@ucla.edu

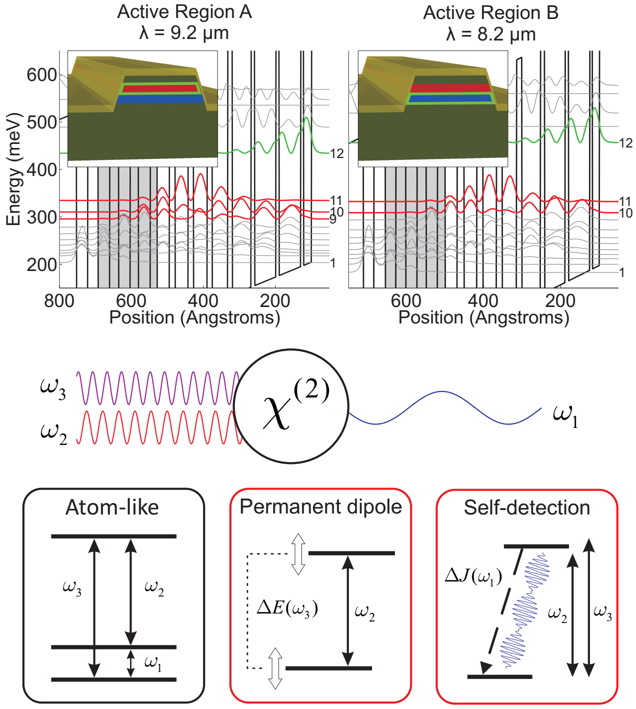

FIG. 1. (Color online) *Top*: Conduction band energy diagram for the two active regions which are combined in series. The states participating in lasing are highlighted. *Bottom*: Three possible mechanisms of difference frequency generation. Atom-like processes involve an intersubband polarization due to off-diagonal dipole matrix elements, enhanced by resonance with the energy structure. Permanent dipole effects are depicted as the modulation of an intersubband transition energy ΔE by an optical field at ω_3, which would modulate the susceptibility of another pump at ω_2 close to resonance, producing a $\chi^{(2)}$. The energy modulation is a first-order Stark effect made possible by spatial separation of the subbands, and does not require resonance of ω_3. Self-detection can be described as an increase in current responding to the intensity beatnote. Mechanisms outlined in red have yet to see thorough theoretical description.

between the displacement current associated with intersubband transitions and the conduction current associated with current flow through the device, and in doing so it is found that both can contribute significantly to THz generation. In particular, we show that an important and previously underappreciated mechanism is the beating of conduction current by stimulated emission across the lasing transition, which becomes dominant over the displacement current contribution for frequencies in the lower THz range ($<$ 2 THz).

The paper is organized as follows. Sec. II describes the method by which the transport and optical properties are calculated using a steady-state density matrix solver and extraction of the velocities. Sec. III gives results and analysis on the active region used in Ref. [14], where it is demonstrated that beating of conduction current is largely responsible for THz generation. Sec. IV summarizes the findings, and details on the steady-state density matrix solver, velocity extraction, and subband electron distributions are given in Appendices.

II. METHOD

Our density matrix solver is adapted from Ref. [19] to compute difference frequency susceptibility. Density matrix transport models have been applied to specific QCL systems since their conception [20–24], but analytic formulations become prohibitively cumbersome for designs consisting of more than 3-4 levels. To address this limitation, generalized density matrix models have recently been presented for modeling of arbitrarily complex designs [19, 25, 26], with the further extensions of coherent optical response and spatial periodicity first made in Ref. [26]. To the best of our knowledge, this work is the first density matrix model for QCLs to coherently include optical fields at more than one frequency.

Each element in the density matrix is an average over the subbands: diagonal elements are therefore the subband population fractions and off-diagonal elements are the coherences between subbands, averaged over the in-plane wavevector. The goal is to solve for the steady-state density matrix of the electronic system (ρ), whose evolution follows a quantum dissipative form:

$$\dot{\rho} = -\frac{i}{\hbar}[H,\rho] + \sum_X C_X^\dagger \rho C_X - \frac{1}{2}\left(C_X^\dagger C_X \rho + \rho C_X^\dagger C_X\right) \\ + \text{pure dephasing...} \tag{1}$$

The first term of the right side is the coherent Liouville-von-Neumann evolution, driven by the Hamiltonian H, which in our case will include subband energy structure, tunneling, and the optical fields. The $\sum_X$ terms are the Lindblad contribution for transitions, where X is a label for each transition process and C_X is the associated "jump" operator. The pure dephasing terms are for processes which reduce coherences but do not alter the subband populations.

A. Hamiltonian and density matrix structure

Following Refs. [19, 26], we assume that the Hamiltonian and density matrix have block periodic form to follow the repetitive structure of a QCL:

$$H = \begin{bmatrix} \ddots & & \vdots & & \iddots \\ & (H_0-\Delta) & (H_1) & (0) & \\ \cdots & (H_{-1}) & (H_0) & (H_1) & \cdots \\ & (0) & (H_{-1}) & (H_0+\Delta) & \\ \iddots & & \vdots & & \ddots \end{bmatrix} \tag{2}$$

$$\rho = \begin{bmatrix} \ddots & & \vdots & & \iddots \\ & (\rho_0) & (\rho_1) & (0) & \\ \cdots & (\rho_{-1}) & (\rho_0) & (\rho_1) & \cdots \\ & (0) & (\rho_{-1}) & (\rho_0) & \\ \iddots & & \vdots & & \ddots \end{bmatrix}. \tag{3}$$

Each term in parentheses is a submatrix of size $N \times N$, where N is the number of states per module. Subscript 0 signifies an intramodule submatrix, and ± 1 refers to the intermodule elements. The module energy difference is accounted for in the matrix $\Delta = E_{mod}\mathbb{1}_N$.

Each submatrix, including the module energy difference, is further decomposed into components at arbitrary frequencies labelled by α:

$$H_p = \sum_\alpha H_p^{(\omega_\alpha)} e^{i\omega_\alpha t} \tag{4}$$

$$\rho_p = \sum_\alpha \rho_p^{(\omega_\alpha)} e^{i\omega_\alpha t} \tag{5}$$

$$\Delta = \sum_\alpha \Delta^{(\omega_\alpha)} e^{i\omega_\alpha t}. \tag{6}$$

The optical field enters into the calculation in an electric dipole sense ($H^{(\omega_\alpha)} = qE(\omega_\alpha)z$, where $E(\omega_\alpha)$ is the optical field and z the position operator). Although this destroys translational invariance in H, application of the Liouville-von Neumann equation will only access *differences* in the diagonals of z, ensuring translational invariance in the complete model. An important consequence of the electric dipole treatment, however, is that the module energy difference fluctuation in Eq. 6 is crucial in treatment of optical nonlinearities. Note that both of $\pm\omega_\alpha$ are included: allowing this in ρ amounts to *not* making any "rotating-wave approximation" as explained in Ref. [26]. A method for solving the entire steady-state is given in Appendix A.

B. Scattering calculations and subband filling

Transitions are added into the steady-state solver through the Lindblad terms in Eq. (1), where the "jump" operator for the transition from state i to f is $C_X = \sqrt{\frac{1}{\tau_{i\to f}}}|f\rangle\langle i|$. The transition time $\tau_{i\to f}$ is averaged over the initial subband electron distribution. In this work, we include transition processes due to interface roughness, LO-phonons, alloy disorder, and ionized impurities. The same processes are calculated for pure dephasing, although their inclusion in the solver is trivial (shown in Appendix A). The scattering calculations were made following Ref. [27].

Some assumptions must be made, which make the model inexact. The usual fitting parameters in QCL simulations are the interface roughness correlation length Λ and average height Δ, for which we found good agreement with experiment using the choices $\Lambda = 25$ nm, $\Delta = 0.8$ Å. In addition, screening must be included in the impurity and LO-phonon calculations, for which we use an isotropic Debye model. Previous studies have found this to be reasonably accurate when the Debye screening length L_D is of the order of or longer than the module length (in our case $L_D \approx 26$ nm with $L_{mod} = 65.5, 69.2$ nm for the two layer sequences studied) [28, 29].

The subband filling statistics in QCLs under operating conditions have been an area of intense study, and are known to have a large impact on the overall transport characteristic. A laser is inherently a nonequilibrium device, and in a QCL equilibrium is broken in more aspects than only the subband populations. Two additional effects are important: the subband electron temperatures T_e tend to be significantly higher than the lattice (phonon) temperature T_L, and subband distributions can often be noticeably nonequilibrium, with hot electrons residing high in the subband, particularly in the lower lasing states [30–32]. In our model, we capture an approximation of both effects by assuming all subbands to be Boltzmann-distributed with $T_e = 500$ K, but with a certain amount of hot electrons superimposed as a Gaussian distribution at higher energy (centered at 140 meV above the subband minimum in light of elastic scattering across the radiative transition). The fraction of hot electrons is made highest in the lower lasing subbands (30%), decreases steadily to zero moving downstream through the injector, and remains at zero for the upper lasing states. More detail is given in Appendix C. This phenomenological scheme is designed to reflect the carrier distributions observed in detailed Monte Carlo simulations which resolve the in-plane k-states [30]. Including these nonequilibium distributions are particularly important to obtain approximate quantitative agreement with experimentally observed current densities within the mid-IR lasers, which is central to the new mechanism of difference susceptibility that we identify.

The results of the scattering calculations give upper state lifetimes near 500 fs and pure dephasing times at the sub-100 fs level.

C. Extraction of transport and optical properties

The steady-state density matrix solution encodes all the known information of the electronic system, and so from it we can extract all the transport and optical properties including current, gain, and nonlinear susceptibility. Because we have an infinitely long chained system with periodic boundary conditions, the polarization is not a uniquely defined quantity as it can in general depend on the boundary positions; therefore, all quantities must be derived from the velocity which is uniquely defined. Supposing we have in general a time-evolution superoperator for the density matrix X ($\dot{\rho} = X\rho$), the

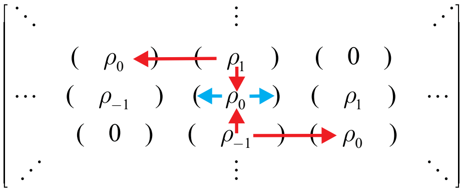

FIG. 2. (Color online) The scheme used to evaluate Eq. (7) to find the total system velocities. Starting/ending points of the arrows correspond to cd/ab elements in Equation 7, respectively. Blue arrows depict terms which move entirely within the intramodule submatrix, and red those which move *into* the intramodule submatrix from outside. Under the assumption that there is no intermodule dipole operator or transitions, the illustrated combinations constitute the fully-representative and nonredundant set.

velocity expectation value is then:

$$\langle v\rangle = Tr(z\dot{\rho}) = Tr(zX\rho) = \sum_{ab,cd} z_{ab}X_{ab,cd}\rho_{cd}. \qquad (7)$$

In other words, we must evaluate the full sum with all possibilities of (density matrix element at *cd*) × (evolution from *cd* to *ab*) × (position element at *ab*). We need a scheme adapted to our periodic system, and so we invoke a requirement for convenience that the module be drawn in such a way that the intermodule dipole matrix elements and transition rates are nonexistent, amounting to a mandate that the module boundary is drawn at the thick tunneling barrier. This location may not identifiable in any QCL system, but is in our case and serves to simplify the mathematics. A visualization of one possible way to evaluate the sum under this assumption is in Fig. 2.

Mathematical details are given in Appendix B, where it is shown how to retrieve the velocities at all frequencies included in the model (Eqs. (4-6)). In simulating a system of two mid-IR pumps at frequencies ω_2 and ω_3 with THz difference frequency ω_1, the parameters of interest (current density J, first-order susceptibility $\chi^{(1)}$, and second-order susceptibility $\chi^{(2)}$) can be extracted as follows:

$$J = N_d q\langle v^{(0)}\rangle \qquad (8)$$

$$\chi^{(1)}(\omega_n) = \frac{N_d q}{i\omega_n \epsilon_0 E_{\omega_n}}\langle v^{(\omega_n)}\rangle \qquad (9)$$

$$\chi^{(2)}(\omega_1 = \omega_3 - \omega_2) = \frac{N_d q}{i\omega_1 \epsilon_0 E_{\omega_3} E_{\omega_2}}\langle v^{(\omega_1)}\rangle. \qquad (10)$$

N_d is the average doping density, E_{ω_n} is the input electric field magnitude at frequency ω_n, and $v^{(\cdots)}$ are the responding velocities at the different frequencies.

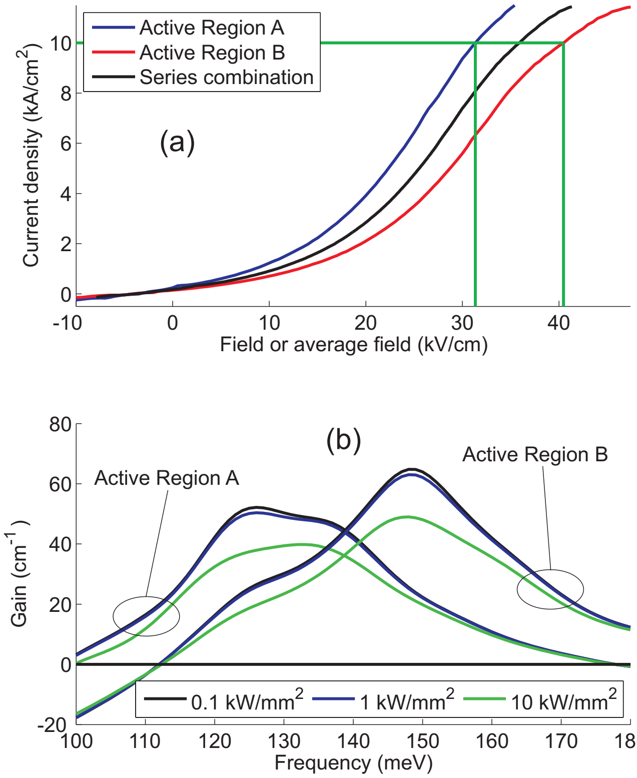

FIG. 3. (Color online) (a) Transport characteristic of both active regions independently and in series combination. Choosing a bias point such that $J = 10$ kA/cm^2 (green line), we find a bias combination of 31.3 and 40.5 kV/cm for the long and short wavelength active regions, respectively (subband energy structures plotted in Fig. 1) (b) Gain computed with increasing intensity for the two active regions separately at the fixed bias fields described above.

III. RESULTS

We choose to model the active region from Vijayraghavan, *et. al.* of dual $\mathrm{In_{0.53}Ga_{0.47}As/In_{0.52}Al_{0.48}As}$ heterostructures [14]. Although the two regions were designed for gain around 8.2 μm (37 THz) and 9.2 μm (33 THz), the transition linewidths were sufficiently broad that it was possible to achieve a large tuning range in the generated THz output from 1.7-5.25 THz by tuning the short wavelength pump in an external cavity setup. The device produced 120 μW of peak power at 4 THz using a dual-period DFG grating cavity, and approximately 15, 45, 15, and 5 μW for 5, 4, 3, and 2 THz, respectively, in the external cavity setup.

A. Transport, bandstructure, and gain

The two active regions are biased in series, and so must draw the same current, which in turn determines the possible bias combinations. Therefore, to choose a pair of biasing points, we must first simulate the transport

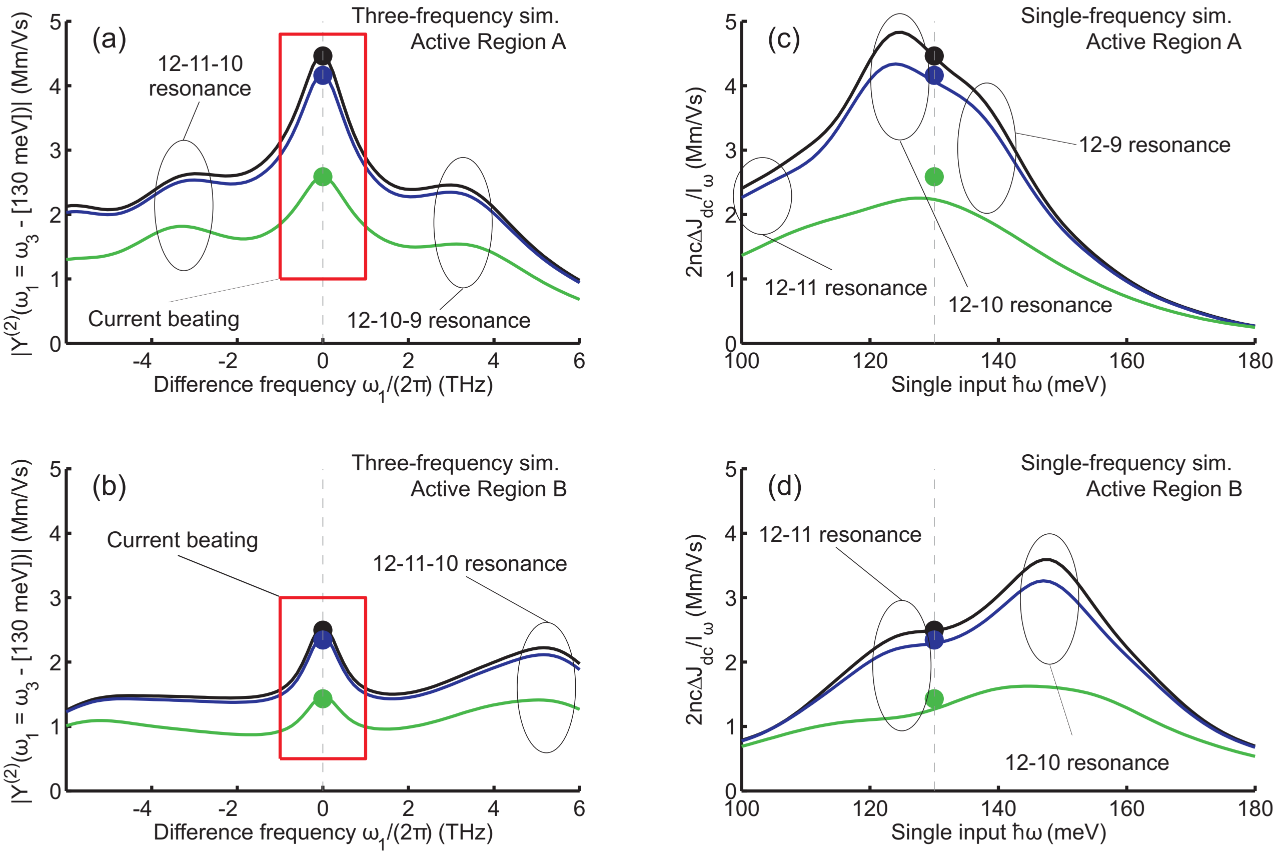

FIG. 4. (Color online) (a,b) Simulated difference frequency current susceptibility $Y^{(2)}(\omega_1 = \omega_3 - \omega_2)$ with increasing pump intensity for (a) Active Region A (long wavelength) and (b) Active Region B (short wavelength). The difference frequency ω_1 is swept with pump ω_2 fixed at 130 meV and the other pump ω_3 swept in conjunction. The black, blue, and green lines correspond to 0.1 kW/mm^2, 1 kW/mm^2, and 10 kW/mm^2, respectively, in each pump. Values at optical rectification are marked by the dots. (c,d) The increase in current as a function of a stimulating frequency as found from a single-frequency simulation for (c) the long wavelength and (d) the short wavelength active regions. (Input intensities are four times the intensity of the single fields in a,b.) Scaling by $2nc$ reproduces the value of $Y^{(2)}$ at optical rectification (denoted by dots), although only for small intensity. Equivalent points in the two simulations along the horizontal axis (single frequency input at 130 meV) are marked by the dashed lines, and resonant processes are denoted on each plot.

characteristics. Bandstructure, tunnel couplings between all pairs of states, and scattering rates are computed at each bias to produce the characteristic shown in Fig. 3a, where we can choose a pair of biases at current density 10 kA/cm^2. It is noted here that the transport characteristic levels off more than was experimentally observed; this is likely due to the fact that leakage to the continuum is not included in our model, which increases with the bias field. Therefore, although it appears in the model that it would be difficult to bias both active regions simultaneously at their highest gain point, in reality the leakage current helps to alleviate this constraint.

Bandstructures calculated using a three-band k.p model at the chosen bias combination are shown in Fig. 1. Wavefunctions are calculated within a single module bounded by adjacent injection barriers, and tunnel couplings are calculated between all possible pairs of states in neighboring modules so as to include any possible injection channels (entering into $H_{1,-1}$ of Eq. (2)). The tunnel couplings were calculated by direct evaluation of the k.p Hamiltonian matrix elements using all of the conduction, light-hole, and split-off wavefunction components which makes for a reliable scheme when nonparabolicity is significant.

Gain for the two active regions with increased intensity (each region treated independently) is shown in Fig. 3b. These simulations include input at only a single frequency, and so neglect cross-saturation due to another pump. The longer wavelength active region exhibits less gain in the model than the shorter, because of the biasing condition explained above. The saturation intensity is realistic: 10 kW/mm^2 amounts to 2 W inside the waveguide with mode area 200 μm^2.

B. Nonlinear susceptibility

Rather than examine the difference frequency susceptibility $\chi^{(2)}$ itself, we instead define a *current susceptibility* $Y^{(2)}$, which is linked to the current, rather than polarization response:

$$Y^{(2)}(\omega_1 = \omega_3 - \omega_2) \equiv \frac{J_{\omega_1}}{\epsilon_0 E_{\omega_2} E_{\omega_3}} = i\omega_1 \chi^{(2)}(\omega_1 = \omega_3 - \omega_2), \tag{11}$$

where J_{ω_1} is the current density response at ω_1 and $E_{\omega_{2,3}}$ are the input electric fields at $\omega_{2,3}$. The actual THz power is then proportional to $|Y^{(2)}|^2$:

$$P_1 = \frac{l_{coh}^2 \left|Y^{(2)}\right|^2 P_2 P_3}{8\epsilon_0 c^3 n_1 n_2 n_3 S_{eff}}, \tag{12}$$

where l_{coh} is the coherence length, P_n are the powers inside the waveguide for each frequency ω_n, n_n are the refractive indices for the same, and S_{eff} is the effective area of interaction [15].

$Y^{(2)}$ is a function of two independent variables, which we could choose to analyze over different lines, as long as the condition $\omega_1 = \omega_3 - \omega_3$ is retained. Figs. 4a,b show $|Y^{(2)}|$ as a function of generated frequency ω_1 with one pump ω_2 fixed at energy 130 meV (9.5 μm). Pump frequency ω_3 is thus swept in conjunction with ω_1 for this scenario. Equal intensities are input in both pumps ω_2, ω_3, while the intensity in the generated frequency is assumed to be negligible. For both active regions, the resonant nonlinearities in the vicinity of 3-5 THz are visible, but they are added to a background which peaks at the optical rectification limit ($\omega_1 = 0, \omega_3 = \omega_2$). The nonzero value of $Y^{(2)}$ at DC generation implies that a steady current is generated, rather than only a polarization; this is the root of the need to analyze $Y^{(2)}$ since $\chi^{(2)}$ exhibits a pole.

Insight into the mechanism behind this peak comes from a single-frequency simulation, shown in Figs. 4c,d. This simulation includes only one optical frequency, which is swept, tracking the increase in DC current ΔJ_{dc}. These functions are, not suprisingly, similar in shape to the gain profile, since the increase in current comes primarily from stimulated emission across the radiative transition. The value of $Y^{(2)}$ at optical rectification can be explained entirely by this mechanism, as shown by the dots connecting equivalent points in the two simulations. At least for vanishing intensity, we see that:

$$\lim_{\omega_3 \to \omega_2} Y^{(2)}(\omega_1 = \omega_3 - \omega_2) = 2nc \frac{\Delta J_{dc}(I_\omega)}{I_\omega}, \tag{13}$$

with n being the refractive index and $\Delta J_{dc}(I_\omega)$ the change in DC current due to intensity I_ω in single pump frequency ω. For fair comparison, we choose $I_\omega = 4I_{\omega_2} = 4I_{\omega_3}$, which is the peak intensity when beating the two pumps. At higher intensity, the right side is found to underpredict the value of $Y^{(2)}$ at optical rectification, meaning that effects at fourth and even higher order begin to have importance. Specifically, this can be interpreted in terms of the harmonics of the beatnote itself; the left-hand side calculates only the first harmonic at ω_1, while the right-hand side would give the complete peak-to-trough distance in the responding current. The fact that the latter underestimates the first means that the higher harmonics work to reduce this distance. Regardless, the quantity of interest is not the peak-to-trough distance but rather the first harmonic itself. Ability to account for this saturation effect highlights the advantage of the nonperturbative treatment used here.

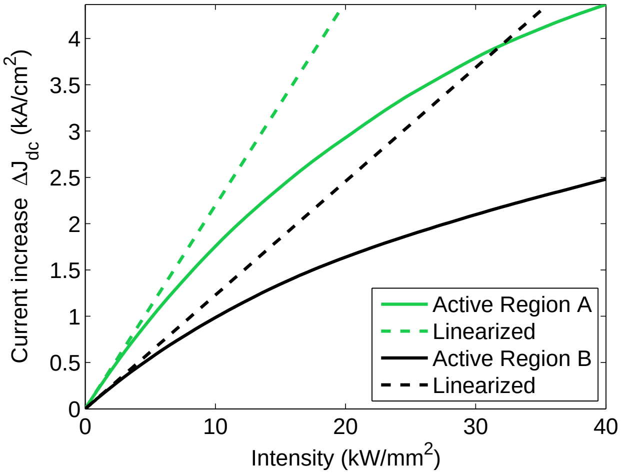

FIG. 5. (Color online) Simulated increase in current density for both active regions in response to a pump at energy 130 meV. The intensity axis is extended to 40 kW/mm^2, the peak intensity when beating two pumps of 10 kW/mm^2 each. Linear expansion of each around zero intensity is given by the dashed lines.

The relationship between current and input intensity at pump energy 130 meV is shown in Figure 5, where the saturation effect, a nonlinearity in $Y^{(2)}$ itself, is clearly evident. The model predicts Active Region A to have less increase in current with intensity than Active Region B, which is an effect of the pump being further from the peak in current stimulation (approximately the same as peak in gain). This is seen also in Figure 4d as compared to Figure 4c, and is also reflected in the reduced height of the optical rectification peak in Figure 4b as compared to 4a.

C. Shortcomings of the perturbative expressions

The nonlinear susceptibility in quantum well active regions has usually been estimated using a perturbative "sum-over-states" (SOS) expression, given as [18]:

$$\chi^{(2)}(\omega_1 = \omega_3 - \omega_2) = \frac{N_d q^3}{\hbar^2 \epsilon_0} \sum_{lmn} z_{ln} z_{nm} z_{ml} (\rho_{ll}^{(0)} - \rho_{mm}^{(0)}) \times \left(\frac{1}{\omega_{nl} - \omega_1 - i\Gamma_{nl}} + \frac{1}{\omega_{nm} + \omega_1 + i\Gamma_{nm}} \right) \times \left(\frac{1}{\omega_{ml} + \omega_2 - i\Gamma_{ml}} + \frac{1}{\omega_{ml} - \omega_3 - i\Gamma_{ml}} \right). \tag{14}$$

The triple sum over state indices l, m, n is within a single module, with ω_{xy} being the resonant frequency between states x and y, z_{xy} the dipole matrix elements, $\rho_{xx}^{(0)}$ the populations at zeroth order (vanishing intensity) and Γ_{xy} the decay rate of density matrix elements at xy. The SOS expression is not meant to handle permanent dipoles (diagonals of the z operator), as it is intended for centrosymmetric atom-like systems, and it can be shown that introduction of these terms does not in general yield a translationally-invariant result. However, the permanent dipoles might in some cases provide a mechanism of intersubband second-order nonlinearity; one classic example is optical rectification in a two-state antisymmetric quantum well system where a case-specific expression had to be derived more carefully [33]. In systems with permanent dipole it is also conceivable that one pump could modulate the energy difference between spatially separated states (a first-order Stark effect). This modulates the first-order susceptibility seen by another pump close to resonance, so that $\chi^{(1)}(\omega_2)$ could be modulated at ω_3 or vice versa. The end result is a second-order nonlinearity only requiring resonance with respect to one of the pump frequencies. To test the hope that the translational variance is small, however, we will additionally consider the result of the SOS expression with permanent dipoles included to attempt to account for permanent dipole effects in the perturbative approach.

Figure 6 displays a comparison between the full calculation of $Y^{(2)}$ and the SOS result both with and without permanent dipoles included in Active Region B this time analyzing: (a) as a function of generated frequency with one pump held at 130 meV, (b) as a function of pump frequency for generation of 4 THz, and (c) as a function of pump frequency for generation of 1 THz. Population inputs to the SOS expression are given from the steady-state solution itself for fair comparison. The phases of $Y^{(2)}$ are given in the insets, where for reference a phase of zero (positive real $Y^{(2)}$) implies velocity in phase with the beating of intensity. We find that while the SOS expression with permanent dipoles included provides a rough estimation of $Y^{(2)}$ in the higher THz range (magnitude comparable and phase within $\pi/8$ at 4 THz), for frequencies in the lower THz the full calculation becomes absolutely necessary. Approaching optical rectification, both SOS expressions yield a vanishing $Y^{(2)}$ since $\chi^{(2)}$ is finite, meaning that any process describable using the sum-over-states has zero efficiency in that limit. It can also be seen from the stark difference in phases that even the SOS expression with permanent dipoles does not include all of the necessary processes; for generation of 1 THz the phase is off by approximately $\pi/2$, and at optical rectification the SOS expressions predict phase at opposite $\pm\pi/2$. This latter implies a DC polarization, when in fact $Y^{(2)}$ has a phase of zero in that limit, corresponding to DC current.

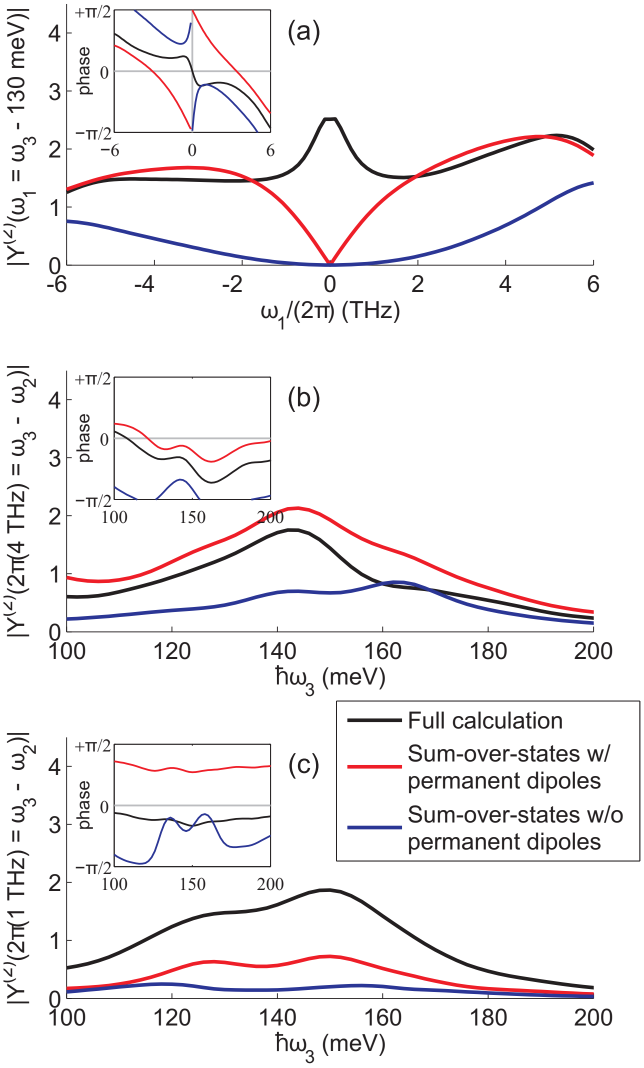

FIG. 6. (Color online) Comparison of $Y^{(2)}(\omega_1 = \omega_3 - \omega_2)$ in Active Region B for the full calculation vs. a sum-over-states for three different scenarios: (a) swept-signal generation with one pump fixed at 130 meV, (b) fixed-signal generation of 4 THz, (c) fixed-signal generation of 1 THz. The magnitudes are shown in the main plots and phases are displayed in the insets. The calculations were done for vanishing intensity. Units for vertical axes are Mm/Vs.

D. Analysis

The effect that we have predicted can be described as the high-frequency tail of self-detection: the addition of two mid-IR waves amounts to a beating of intensity which stimulates current response at the difference frequency. This is associated with the radiative transition, evident in the similar shape to the gain profile with respect to pump frequency (exhibited in all of Figs. 4c,d and 6b,c). Indeed, the increase of current with intensity is experimentally visible in QCLs, most noteably as a discontinuity in the differential conductance at threshold as the onset of stimulated emission decreases the upper state lifetime. The response time associated with this mechanism is linked to the subpicosecond intersubband scattering times, which allows the bandwidth to reach into the THz range.

QCLs have an inherent design advantage in using this effect, since it is tied to the radiative transitions which the pumps are automatically close to resonance with. One might choose to approximate this detection effect by fitting to a simple response model:

$$Y^{(2)}_{detection} \approx \frac{2nc}{1+i\omega\tau}\beta, \tag{15}$$

where τ is a phenomenological response time and β is a coefficient for the current increase with intensity ($\beta = \partial J/\partial I$ for vanishing I). At lasing intensity, however, the coefficient β is reduced because of the saturation of the detection effect, which is tied to gain saturation, exhibited in Figure 4. Nevertheless, we can fit β to the full model for different intensity levels. Figure 7 shows an approximate fit to this model, where the full calculation is compared to a simpler one where the detection ($Y^{(2)}_{detection}$) and SOS ($Y^{(2)}_{SOS}$) contributions are directly superimposed:

$$Y^{(2)} \approx Y^{(2)}_{detection} + Y^{(2)}_{SOS}. \tag{16}$$

Moderate quantitative agreement is found with the simple model, with $Y^{(2)}$ being overpredicted by inclusion of permanent dipoles in the SOS expression but underpredicted by their exclusion. This suggests that permanent dipole effects even not linked to current beating play an important role reaching over the whole THz range, and that the translational variance of the SOS expression in accounting for them presents a significant error. Some error might also be introduced into the simpler model by the fact that β has frequency-dependence which would become important as the higher frequency pump moves further away, or by additional sources including the tunnel couplings which cannot be accounted for in the SOS expression.

Given that the current beating effect contributes significantly to difference frequency generation, it is useful to establish a way to estimate its strength in real devices using commonly measured experimental parameters. One such parameter is the differential conductance discontinuity at threshold ΔG, and another is the "slope efficiency" of the output power P_{out} vs. injection current I_d. Since the key parameter of interest is β, we solve for it at threshold:

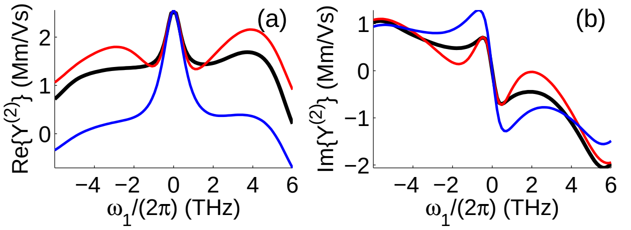

FIG. 7. (Color online) Fitting of the full calculation to a simpler model involving superposition of detection and SOS components. The black, red, and blue lines give results from the full calculation, SOS expression with permanent dipoles, and SOS expression without permanent dipoles. Calculations are performed on Active Region B, for the same fixed-pump scenario as in Figs. 4a,b and 6a. (a) Real and (b) imaginary parts are given for vanishing intensity. The fitted parameter $\beta = 0.125$ ((kA/cm^2)/(kW/mm^2)) for vanishing intensity, although at 10 kW/mm^2 in each pump (not shown) this coefficient is reduced to 0.072 ((kA/cm^2)/(kW/mm^2)). The approximate response time τ is 240 fs.

$$\beta = \frac{\Delta G}{G_0+\Delta G}\frac{1}{S}\frac{A_{mode}}{A_{active}}\frac{T}{2}, \tag{17}$$

G_0 is the differential conductance just below threshold, S is the slope efficiency defined as dP_{out}/dI_d, A_{mode} and A_{active} are the lasing mode and top-down active region areas, respectively, and T is the output facet transmission (approximating that $T/2$ is the ratio of output power to total power inside the waveguide). Since A_{mode}, A_{active}, and T are primarily cavity-related parameters, Eq. (17) provides a useful metric for comparison between different active regions by placement in the same cavity configuration. In a QCL, this expression would be approximately proportional to the population inversion [34], which is intuitively linked to the strength of current beating. It is important to note that the value of β as found by (17) is for vanishing intensity and at threshold, but will likely still be indicative of the strength at more normal operating conditions.

The injection barrier thickness is well known to have large impact on the coherence of the injection process, and hence on ΔG. Fig. 8 displays the effect of an altered barrier width on $Y^{(2)}$ for both active regions. The results suggest that there is some danger in suppression of the current beating by choosing too thick a barrier, and also that there may be some room for improvement by its reduction - at least in the case of Active Region A.

The collective results of this paper suggest a simple strategy for optimizing performance for low THz gener-

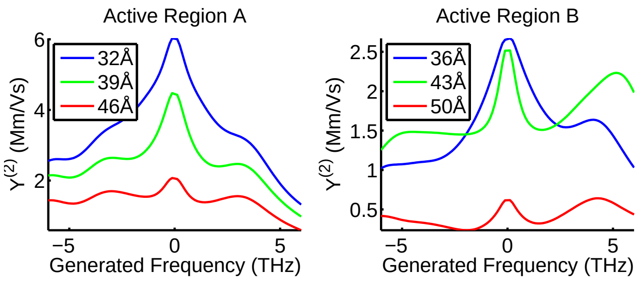

FIG. 8. (Color online) Effect of the injection barrier width (in legends) on $Y^{(2)}$ for vanishing intensity. Green lines correspond to values chosen in Ref. [14].

ation (< 2 THz). We suggest using only a single active region, which should have sufficient mid-IR gain bandwidth when the pumps are closer in frequency. Using a single active region removes the current continuity constraint of using two active regions, and allows operation at the optimum point. Active region A appears to be preferred; that is to say, it is predicted to have a larger $Y^{(2)}$ value when biased at its optimum point, and has more room for improvement by thinning the injection barrier.

IV. CONCLUSIONS

We have presented a density matrix transport model for QCLs which handles both gain and optical nonlinearity coherently and nonperturbatively. Scattering and dephasing processes were carefully accounted for by detailed calculation of the most relevant mechanisms, and care was taken to account for hot subband distributions and nonequilbrium electrons. The model predicts reasonable current levels as compared with experiment and predicts the expected amount of gain for lasing and a reasonable saturation intensity.

The computed nonlinearity exhibits a peak in *efficiency*, as opposed to merely *susceptibility*, at optical rectification, which is ascribed to the increase of DC current proportional to intensity. This increase is identified as a self-detection effect occurring through stimulated emission across the radiative transition, a familiar phenomenon in QCLs which has even seen recent use in THz phase-sensitive imaging systems [35, 36]. We have shown that the high-frequency tail of this effect extends into the THz range because of the subpicosecond intersubband scattering times. The detection itself is highly nonlinear, meaning that fourth- and higher-orders are significant. A sum-over-states expression was found not to accurately reproduce the complete susceptibility, especially for generated frequencies in the lower THz range where the detection is strong. The complete susceptibility is reasonably well-matched by superposition of a fitted detection susceptibility and a sum-over-states expression, although even this is still not exact, since permanent dipoles and other mechanisms play a role in the nonlinearity which is not correctly accounted for by a sum-over-states. Finally, a metric was derived to assess the strength of self-detection for active regions through experimentally-accessible parameters, and suggestions were made for improvement of the performance in the lower THz range (< 2 THz) .

The current beating, or self-detection, effect is large, and our results serve to explain the surprisingly low-frequency THz generation in a device demonstrated in the literature. The prediction of significant current susceptibility $Y^{(2)}$ extending to DC-generation suggests that the eventual low-frequency shutoff is owing more to other factors such as free-carrier absorption, phase matching, and output coupling, which are beyond the scope of this work. Regardless, a route forward to increased conversion efficiency in the lower THz range might aim to exploit this effect. There is no need to attempt to lower the frequency of a resonant nonlinearity whose conversion efficiency will scale downwards as the square of the generated frequency. Specifically, we have pointed out that a simple experimental metric useful for the assessment of such active regions is the relative differential conductance change at the onset of lasing.

Finally, it is possible that the formalism developed in this work is applicable to the study of the recently developed multi-mode QCLs and frequency combs [37–39]. Most directly, it is conceivable that the same detection process predicted here is responsible for the familiar radio-frequency beatnote generation, as a result of current beating from pairs of adjacent frequency lines. Further, a similar model could be extended to study the effect of radio-frequency modulation on such active regions [40, 41], particularly to assess the contribution coming from nonresonant transition energy modulation enabled by permanent dipoles. The formalism presented here could also be readily extended to encompass third order nonlinearities, which would allow the study of comb generation itself.

ACKNOWLEDGMENTS

This work was partially supported by NSF ECCS grants 1002387, 1150071, and 1509801.

Appendix A: Steady-state density matrix solution

We begin by separating the evolution in Eq. (1) into coherent, transition, and dephasing components and writing in the steady-state condition:

$$\dot{\rho} = \dot{\rho}|_{coh} + \dot{\rho}|_{trans} + \dot{\rho}|_{deph} = \sum_n i\omega_n \rho^{(\omega_n)} e^{i\omega_n t}. \tag{A1}$$

Applying block matrix multiplication to the coherent evolution with Eqs. (2,3) as input, we arrive at the general equation for coherent evolution of any submatrix in ρ:

$$\dot{\rho_p}|_{coh} = \sum_q \left[H_{p-q}, \rho_q\right] - p\Delta\rho_p. \tag{A2}$$

Next, Δ and submatrices of H and ρ are expanded into their steady-state harmonics (Eqs. (4-6)), and by isolating in frequency we arrive at the general equation for coherent evolution of any harmonic of any submatrix in ρ:

$$\dot{\rho_p}|_{coh}^{(\omega_m)} = e^{i\omega_m t} \sum_{qn} \left(-\frac{i}{\hbar} \left[H_{p-q}^{(\omega_m-\omega_n)}, \rho_q^{(\omega_n)} \right] - \delta_{pq} q \Delta^{(\omega_m-\omega_n)} \rho_q^{(\omega_n)} \right). \tag{A3}$$

This equation provides some interesting insight, which also aids in writing down the complete solution later: submatrices in ρ connect to other submatrices through the difference submatrix in H, and also frequencies in ρ connect to other frequencies through the difference frequency components in H. Since we are to solve for all elements in each matrix, we apply the vectorization transformation (columnwise conversion of a matrix into a column vector), which has the useful property that $vec\{AB\} = (\mathbb{1}_N \otimes A) vec\{B\} = (B^T \otimes \mathbb{1}_N) vec\{A\}$, for multiplication of two square matrices A and B each having dimension N. Vectorization of Eq. (A3) gives:

$$vec\left\{\dot{\rho_p}|_{coh}^{(\omega_m)}\right\} = e^{i\omega_m t} \sum_{qn} \left[-\frac{i}{\hbar} \left(\mathbb{1}_N \otimes H_{p-q}^{(\omega_m-\omega_n)} - H_{p-q}^{(\omega_m-\omega_n),T} \otimes \mathbb{1}_N \right) - \delta_{pq} q E_{mod}^{(\omega_m-\omega_n)} \right] vec\left\{\rho_{p-q}^{(\omega_n)}\right\}, \tag{A4}$$

with $E_{mod}^{(\omega_m-\omega_n)}$ as the scalar module energy at the different frequencies, including DC.

Next we move to the transition contribution ($\sum_X$ terms in Eq. (1)), where we will work under the assumption that transitions only occur within the module. Each transition process has separate instances inside each module, each of which has its own Lindblad superoperator (as in Eq. (A5) in Ref. [19]) formed with a lone instance of the intramodule submatrix $\bar{C_X}$ at the module position. The result of each can be found using a similar block matrix multiplication tactic as was used for the coherent contribution, and added to yield:

$$\dot{\rho_p}|_{trans} = \sum_X \delta_{p0} \bar{C_X}^\dagger \rho_p \bar{C_X} - \frac{1}{2}\left[\bar{C_X}^\dagger \bar{C_X} \rho_p + \rho_p \bar{C_X}^\dagger \bar{C_X} \right], \tag{A5}$$

which is again a fairly intuitive equation as we see that the transitions can only increase the intramodule elements of ρ, where the population transfer occurs, while the associated dephasing affects all elements. Separation into frequencies is trivial since the jump operators carry no time dependence, and then we can vectorize Eq. (A5), leading to:

$$vec\left\{\dot{\rho_p}|_{trans}^{(\omega_m)}\right\} = e^{i\omega_m t} \sum_x \left[\delta_{p0} \left(\bar{C_X} \otimes \bar{C_X} \right) - \frac{1}{2} \left(\mathbb{1}_N \otimes \bar{C_X^\dagger} \bar{C_X} + \bar{C_X}^\dagger \bar{C_X} \otimes \mathbb{1}_N \right) \right] vec\left\{\rho_p^{(\omega_m)}\right\}. \tag{A6}$$

Finally, the dephasing processes are the simplest to treat. Given matrices of the dephasing times in each submodule named $T_{2,p}$, where $T_{2,0}$ is the intramodule dephasing and $T_{2,\pm 1}$ are the dephasings of $\rho_{1,-1}$, ($T_{2,-1} = T_{2,1}^T$), we have:

$$vec\left\{\dot{\rho_p}|_{trans}^{(\omega_m)}\right\} = -e^{i\omega_m t} vec\left\{T_{2,p}^{\circ(-1)}\right\} \circ vec\left\{\rho_p^{(\omega_m)}\right\}, \tag{A7}$$

with the symbol $\circ$ denoting the Hadamard (elementwise) product and the superscript in $T_{2,p}^{\circ(-1)}$ the Hadamard inverse. Now based on substitution of Eqs. (A4), (A6), and (A7) into the steady-state condition of (A1), we can organize the entire solution by the following:

$$\left[\begin{array}{c|c|c} P(H_0)+Q_1 & P(H_{-1}) & 0 \\ \hline P(H_1) & P(H_0)+Q_0 & P(H_{-1}) \\ \hline 0 & P(H_1) & P(H_0)+Q_{-1} \end{array}\right] \times \left[\begin{array}{c} vec\left\{\rho_{-1}^{(\omega_{\dots})}\right\}\begin{smallmatrix}-\omega_3\\-\omega_2\\-\omega_1\\0\\+\omega_1\\+\omega_2\\+\omega_3\end{smallmatrix} \\ \hline vec\left\{\rho_{0}^{(\omega_{\dots})}\right\}\begin{smallmatrix}-\omega_3\\-\omega_2\\-\omega_1\\0\\+\omega_1\\+\omega_2\\+\omega_3\end{smallmatrix} \\ \hline vec\left\{\rho_{1}^{(\omega_{\dots})}\right\}\begin{smallmatrix}-\omega_3\\-\omega_2\\-\omega_1\\0\\+\omega_1\\+\omega_2\\+\omega_3\end{smallmatrix} \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline 0 \\ \hline 0 \end{array}\right] \tag{A8}$$

$$P(H_p) = \begin{array}{c} \\ -\omega_3 \\ -\omega_2 \\ -\omega_1 \\ 0 \\ +\omega_1 \\ +\omega_2 \\ +\omega_3 \end{array} \begin{array}{ccccccc} -\omega_3 & -\omega_2 & -\omega_1 & 0 & +\omega_1 & +\omega_2 & +\omega_3 \\ \mathcal{O}_p^{(0)} & \mathcal{O}_p^{(-\omega_1)} & \mathcal{O}_p^{(-\omega_2)} & \mathcal{O}_p^{(-\omega_3)} & X & X & X \\ \mathcal{O}_p^{(+\omega_1)} & \mathcal{O}_p^{(0)} & X & \mathcal{O}_p^{(-\omega_2)} & \mathcal{O}_p^{(-\omega_3)} & X & X \\ \mathcal{O}_p^{(+\omega_2)} & X & \mathcal{O}_p^{(0)} & \mathcal{O}_p^{(-\omega_1)} & X & \mathcal{O}_p^{(-\omega_3)} & X \\ \mathcal{O}_p^{(+\omega_3)} & \mathcal{O}_p^{(+\omega_2)} & \mathcal{O}_p^{(+\omega_1)} & \mathcal{O}_p^{(0)} & \mathcal{O}_p^{(-\omega_1)} & \mathcal{O}_p^{(-\omega_2)} & \mathcal{O}_p^{(-\omega_3)} \\ X & \mathcal{O}_p^{(+\omega_3)} & X & \mathcal{O}_p^{(+\omega_1)} & \mathcal{O}_p^{(0)} & X & \mathcal{O}_p^{(-\omega_2)} \\ X & X & \mathcal{O}_p^{(+\omega_3)} & \mathcal{O}_p^{(+\omega_2)} & X & \mathcal{O}_p^{(0)} & \mathcal{O}_p^{(-\omega_1)} \\ X & X & X & \mathcal{O}_p^{(+\omega_3)} & \mathcal{O}_p^{(+\omega_2)} & \mathcal{O}_p^{(+\omega_1)} & \mathcal{O}_p^{(0)} \end{array}$$

$$Q_p = i \begin{bmatrix} +\omega_3 & & & & & & \\ & +\omega_2 & & & & & \\ & & +\omega_1 & & & & \\ & & & 0 & & & \\ & & & & -\omega_1 & & \\ & & & & & -\omega_2 & \\ & & & & & & -\omega_3 \end{bmatrix} \otimes \mathbb{1}_{N^2} + R_p + S_p + D_p \quad \Bigg| \quad \mathcal{O}_p^{(\omega_m)} = -\frac{i}{\hbar}\left(\mathbb{1}_N \otimes H_p^{(\omega_m)} - H_p^{(\omega_m),T} \otimes \mathbb{1}_N\right)$$

$$R_p = -\frac{ip}{\hbar} \begin{array}{c} \\ -\omega_3 \\ -\omega_2 \\ -\omega_1 \\ 0 \\ +\omega_1 \\ +\omega_2 \\ +\omega_3 \end{array} \begin{array}{ccccccc} -\omega_3 & -\omega_2 & -\omega_1 & 0 & +\omega_1 & +\omega_2 & +\omega_3 \\ E_{mod}^{(0)} & E_{mod}^{(-\omega_1)} & E_{mod}^{(-\omega_2)} & E_{mod}^{(-\omega_3)} & X & X & X \\ E_{mod}^{(+\omega_1)} & E_{mod}^{(0)} & X & E_{mod}^{(-\omega_2)} & E_{mod}^{(-\omega_3)} & X & X \\ E_{mod}^{(+\omega_2)} & X & E_{mod}^{(0)} & E_{mod}^{(-\omega_1)} & X & E_{mod}^{(-\omega_3)} & X \\ E_{mod}^{(+\omega_3)} & E_{mod}^{(+\omega_2)} & E_{mod}^{(+\omega_1)} & E_{mod}^{(0)} & E_{mod}^{(-\omega_1)} & E_{mod}^{(-\omega_2)} & E_{mod}^{(-\omega_3)} \\ X & E_{mod}^{(+\omega_3)} & X & E_{mod}^{(+\omega_1)} & E_{mod}^{(0)} & X & E_{mod}^{(-\omega_2)} \\ X & X & E_{mod}^{(+\omega_3)} & E_{mod}^{(+\omega_2)} & X & E_{mod}^{(0)} & E_{mod}^{(-\omega_1)} \\ X & X & X & E_{mod}^{(+\omega_3)} & E_{mod}^{(+\omega_2)} & E_{mod}^{(+\omega_1)} & E_{mod}^{(0)} \end{array} \otimes \mathbb{1}_{N^2}$$

$$S_p = \mathbb{1}_7 \otimes \left\{\sum_X \left[\delta_{p0}\left(\bar{C_X} \otimes \bar{C_X}\right) - \frac{1}{2}\left(\mathbb{1}_N \otimes \bar{C_X}^\dagger \bar{C_X} + \bar{C_X}^\dagger \bar{C_X} \otimes \mathbb{1}_N\right)\right]\right\} \quad \Bigg| \quad D_p = -\mathbb{1}_7 \otimes diag\left\{vec\left\{T_{2,p}^{\circ(-1)}\right\}\right\}.$$

The complete steady state is then soluble after substituting a population sum condition to a single row in (A8). The method is formulated here for a set of three frequencies where $\omega_1 + \omega_2 = \omega_3$, but it is straightforward to generalize to other situations including the single-frequency simulation referred to in Figs. 4c,d and 5.

Appendix B: Evaluating the velocities

The sum in Eq. (7) can be rearranged for interpretation in two different ways: if we choose $\langle v\rangle = \sum_{ab} z_{ab}\left(\sum_{cd} X_{ab,cd}\rho_{cd}\right)$, we recover the original concept of $\langle v\rangle = Tr(z\dot{\rho})$, whereas if we choose instead $\langle v\rangle = \sum_{cd}\rho_{cd}\left(\sum_{ab} z_{ab}X_{ab,cd}\right)$, it appears we have found a velocity operator v and are now using $\langle v\rangle = Tr(v\rho)$. In this spirit, the sum scheme drawn in Fig. 2 can be separated into three parts based on origination from ρ_0, ρ_1, and ρ_{-1}, formally:

$$\langle v \rangle = Tr(v_0\rho_0) + Tr(v_{-1}\rho_1) + Tr(v_1\rho_{-1}). \tag{B1}$$

The first term can be computed directly from the pieces of (A8) and the dipole operator, since all the pieces of the time evolution superoperator X are in place. Even the necessary frequency mixing is already organized. Using the intramodule dipole submatrix Z_0, we can evaluate the first term contributions to (B1) at all frequencies as follows:

$$\begin{bmatrix} A^{(\omega_{\ldots})\begin{smallmatrix}-\omega_3\\-\omega_2\\-\omega_1\\0\\+\omega_1\\+\omega_2\\+\omega_3\end{smallmatrix}} \end{bmatrix} = \Big(P(H_0) + S_0 + D_0\Big) \begin{bmatrix} vec\left\{\rho_0^{(\omega_{\ldots})}\right\}^{\begin{smallmatrix}-\omega_3\\-\omega_2\\-\omega_1\\0\\+\omega_1\\+\omega_2\\+\omega_3\end{smallmatrix}} \end{bmatrix} \quad , \qquad Tr(v_0\rho_0)^{(\omega_{\ldots})} = vec\left\{Z_0\right\}^T A^{(\omega_{\ldots})}. \tag{B2}$$

The second and third terms in Eq. (B1), on the other hand, *cannot* be evaluated in the same approach, because the matrix equation (A8) does not distinguish between destination modules in the evolution pointing from intramodule to intermodule elements. However, since it is seen clearly that evolution of this nature is entirely coherent (always $P(H_{\pm 1})$), we can be sure that the intermodule velocity operators can be constructed entirely from H and the dipole matrix Z using $v = (i/\hbar)[H, Z]$. Applying block matrix multiplication given H in the form of Eq. (2) and a similar Z having only intramodule submatrices, we obtain the off-diagonal velocity operator submatrices:

$$v_{\pm 1} = \frac{i}{\hbar}\left([H_{\pm 1}, Z_0] \mp LH_{\pm 1}\right), \tag{B3}$$

with L as the spatial separation between modules employed in the same fashion as the energy separation in Eq. (2). Since $v_{\pm 1}$ carries no time dependence, evaluation of the second and third traces in (B1) are now straightforward.

Appendix C: Subband electron distributions

To retain simplicity while still capturing the important effects of the nonequilibrium subband distributions, we mandate that the subbands are mostly thermalized to a Boltzmann distribution at 500 K with a certain fraction of hot nonequilibrium electrons. Hot electrons are produced mainly by elastic or LO-phonon scattering across the radiative transition to states high up in the lower lasing subbands, and may exist further down the injector although they will gradually disappear through electron-electron and other intrasubband scattering. Assuming that the hot electrons are produced across the radiative transition, the end result is a bump in the electron distribution at roughly the radiative transition energy above the subband minimum for the lower lasing state; this is approximated in this work as a normal distribution centered at 140 meV. The explicit equation for the subband distribution (fraction of population per unit energy) is then:

$$P_n(E_\parallel) = \frac{(1-f_n)}{k_BT}e^{-E_\parallel/k_BT} + \frac{f_n}{\sqrt{2\pi\sigma^2}}e^{-(E_\parallel-\mu)^2/2\sigma^2}, \tag{C1}$$

where $E_\parallel$ is the in-plane electron kinetic energy, f_n is the hot electron fraction assigned to the subband, $\mu =$ 140 meV is the hot electron energy center, and σ is the standard deviation (chosen as 25 meV). The hot electron fractions are assigned as the following:

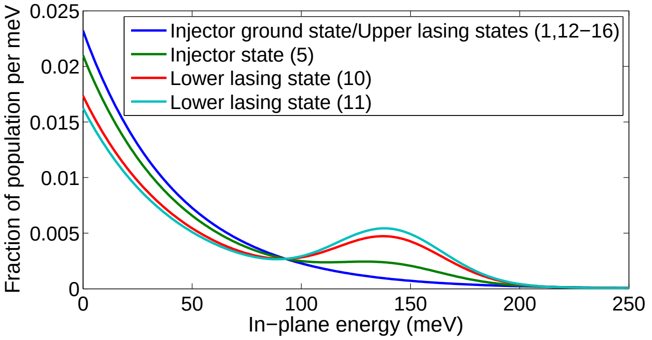

FIG. 9. Assumed electron distributions for selected subbands in Active Region B.

$$f_n = \begin{cases} f_{11}\frac{E_n - E_1}{E_{11}-E_1} & n \le 11 \\ 0 & n \ge 12 \end{cases} \tag{C2}$$

where $f_{11} = 0.3$ is the hot electron fraction for the lower lasing state 11 and E_n are the subband energies. A plot of selected subband distributions for Active Region B is shown in Fig. 9. Our choice of subband electron distribution is in light of other works, most notably the Monte Carlo simulation by Matyas et. al. (see Ref. [30] Fig. 3a), the NEGF results by Lindskog et. al. (see Ref. [31]), and the experimental measurements by Spagnolo et. al. (Ref. [32]).

[1] J. Faist, F. Capasso, D. L. Sivco, C. Sirtori, A. L. Hutchinson, and A. Y. Cho, Science **264**, 553 (1994).

[2] Y. Yao, A. J. Hoffman, and C. F. Gmachl, "Mid-infrared quantum cascade lasers," Nature Photonics **6**, 432–439 (2012).

[3] R. Kohler, A. Tredicucci, F. Beltram, H.E. Beere, E.H. Linfield, A.G. Davies, D.A. Ritchie, R.C. Iotti, and F. Rossi, "Terahertz semiconductor-heterostructure laser," Nature (London) **417**, 156 (2002).

[4] B. S. Williams, "Terahertz quantum-cascade lasers," Nature Photonics **1**, 517 (2007).

[5] R. Nelander and A. Wacker, "Temperature dependence of the gain profile for terahertz quantum cascade lasers," Appl. Phys. Lett. **92**, 081102 (2008).

[6] M. A. Belkin, J. A. Fan, S. Hormoz, F. Capasso, S. P. Khanna, M. Lachab, A. G. Davies, and E. H. Linfield, "Terahertz quantum cascade lasers with copper metal-metal waveguides operating up to 178 K," Optics Express **16**, 3242 (2008).

[7] S. Fathololoumi, E. Dupont, C. W. I. Chan, Z. R. Wasilewski, S. R. Laframboise, D. Ban, A. Matyas, C. Jirauschek, Q. Hu, and H. C. Liu, "Terahertz quantum cascade lasers operating up to 200K with optimized oscillator strength and improved injection tunneling," Optics Express **20**, 3866 (2012).

[8] Y. Chassagneux, Q. J. Wang, S. P. Khanna, E. Strupiechonski, J.-R. Coudevylle, E. H. Linfield, A. G. Davies, F. Capasso, M. A. Belkin, and R. Colombelli, "Limiting factors to the temperature performance of THz quantum cascade lasers based on the resonant-phonon depopulation scheme," IEEE Transactions on Terahertz Science and Technology. **2**, 83 (2012).

[9] P. H. Siegel, "Terahertz technology," IEEE Trans. Microwave Theory Tech. **50**, 910 (2002).

[10] M. Tonouchi, "Cutting-edge terahertz technology," Nature Photonics **1**, 97 (2007).

[11] M. A. Belkin, F. Capasso, A. Belyanin, D. L. Sivco, A. Y. Cho, D. C. Oakley, C. J. Vineis, and G. W. Turner, "Terahertz quantum-cascade laser source based on intracavity difference-frequency generation," Nature Photonics **1**, 288–292 (2007).

[12] Q. Y. Lu, N. Bandyopadhyay, S. Slivken, Y. Bai, and M. Razeghi, "Continuous operation of a monolithic semiconductor terahertz source at room temperature," Appl. Phys. Lett. **104**, 221105 (2014).

[13] K. Vijayraghavan, R. W. Adams, A. Vizbaras, M. Jan, C. Grasse, G. Boehm, M. C. Amann, and M. A. Belkin, "Terahertz sources based on Cerenkov difference-frequency generation in quantum cascade lasers," Appl. Phys. Lett. **100**, 251104 (2012).

[14] K. Vijayraghavan, Y. Jiang, M. Jang, A. Jiang, K. Choutagunta, A. Vizbaras, F. Demmerle, F. Demmerle G. Boehm, M. C. Amann, and M. A. Belkin, "Broadly tunable terahertz generation in mid-infrared quantum cascade lasers," Nature communications **4**, 2021 (2013).

[15] M. A. Belkin and F. Capasso, "New frontiers in quantum cascade lasers: high performance room temperature terahertz sources," Phys. Scr. **90**, 118002 (2015).

[16] C. Jirauschek, A. Matyas, P. Lugli, and M.-C. Amann, "Monte Carlo study of terahertz difference frequency generation in quantum cascade lasers," Optics Express **21**, 6180 (2013).

[17] C. Jirauschek, H. Okeil, and P. Lugli, "Monte Carlo analysis of the terahertz difference frequency generation susceptibility in quantum cascade laser structures," Optics Express **23**, 1670–1678 (2015).

[18] R. W. Boyd, *Nonlinear Optics* (Academic Press, 2007).

[19] B. A. Burnett and B. S. Williams, "Density matrix model for polarons in a terahertz quantum dot cascade laser," Phys. Rev. B **90**, 155309 (2014).

[20] R. F. Kazarinov and R. A. Suris, "Possibility of the amplification of electromagnetic waves in a semiconductor with a superlattice," Sov. Phys. Semicond. **5**, 707 (1971).

[21] C. Sirtori, F. Capasso, J. Faist, A. L. Hutchinson, D. L. Sivco, and A. Y. Cho, "Resonant tunneling in quantum cascade lasers," IEEE J. Quantum Electron. **34**, 1722–1729 (1998).

[22] S. Kumar and Q. Hu, "Coherence of resonant-tunneling transport in terahertz quantum-cascade lasers," Phys. Rev. B **80**, 245316 (2009).

[23] H. Callebaut and Q. Hu, "Importance of coherence for electron transport in terahertz quantum cascade lasers," J. Appl. Phys. **98**, 2005 (2005).

[24] E. Dupont, S. Fathololoumi, and H. C. Liu, "Simplified density-matrix model applied to three-well terahertz quantum cascade lasers," Phys. Rev. B **81**, 205311 (2010).

[25] R. Terazzi and J. Faist, "A density matrix model of transport and radiation in quantum cascade lasers," New J. Phys. **12**, 033045 (2010).

[26] T. V. Dinh, A. Valavanis, L. J. M. Lever, Z. Ikonic, and R. W. Kelsall, "Extended density-matrix model applied to silicon-based terahertz quantum cascade lasers," Phys. Rev. B **85**, 235427 (2012).

[27] T. Noda H. Sakaki T. Unuma, M. Yoshita and H. Akiyama, "Intersubband absorption linewidth in GaAs quantum wells due to scattering by interface roughness, phonons, alloy disorder, and impurities," J. Appl. Phys. **93**, 1586 (2003).

[28] R. Nelander and A. Wacker, "Temperature dependence and screening models in quantum cascade structures," J. Appl. Phys. **1.6**, 063115 (2009).

[29] C. Jirauschek and T. Kubis, "Modeling techniques for quantum-cascade lasers," Appl. Phys. Rev. **1**, 011307 (2014).

[30] A. Matyas, P. Lugli, and C. Jirauschek, "Photon-induced carrier transport in high efficiency midinfrared quantum cascade lasers," J. Appl. Phys. **110**, 013108 (2011).

[31] M. Lindskog, J. M. Wolf, V. Trinite, V. Liverini, J. Faist, G. Maisons, M. Carras, R. Aidam, R. Ostendorf, and A. Wacker, "Comparative analysis of quantum cascade laser modeling based on density matrices and non-equilibrium Green's functions," Appl. Phys. Lett. **105**, 103106 (2014).

[32] V. Spagnolo, G. Scamarcio, H. Page, and C. Sirtori, "Simultaneous measurment of the electronic and lattice temperatures in GaAs/Al0.45Ga0.55As quantum-cascade lasers: Influence on the optical performance," Appl. Phys. Lett. **84**, 3690 (2004).

[33] E. Rosencher and Ph. Bois, "Model system for optical nonlinearities: Asymmetric quantum wells," Phys. Rev.

B **44**, 11315 (1991).

[34] J. Faist, *Quantum Cascade Lasers* (Oxford University Press, 2013).

[35] P. Dean, A. Valavanis, J. Keeley, K. Bertling, Y. L. Lim, R. Alhathlool, S. Chowdhury, T. Taimre, L. H. Li, D. Indjin, S. J. Wilson, A. D. Rakic, E. H. Linfield, and A. G. Davies, "Coherent three-dimensional terahertz imaging through self-mixing in a quantum cascade laser," Appl. Phys. Lett. **103**, 181112 (2013).

[36] F. P. Mezzapesa, L. L. Columbo, G. De Risi, M. Brambilla, M. Dabbicco, V. Spagnolo, and G. Scamarcio, "Nanoscale Displacement Sensing Based on Nonlinear Frequency Mixing in Quantum Cascade Lasers," IEEE J. Sel. Top. Quantum Electron. **21**, 1200908 (2015).

[37] S. Barbieri, J. Alton, C. Baker, T. Lo, H. E. Beere, and D. Ritchie, "Imaging with THz quantum cascade lasers using a Schottky diode mixer," Optics Express **13**, 6497 (2005).

[38] A. Hugi, G. Villares, S. Blaser, H. C. Liu, and J. Faist, "Mid-infrared frequency comb based on a quantum cascade laser," Nature (London) **492**, 229–233 (2012).

[39] D. Burghoff, T.-Y. Kao, N. Han, C.W.I. Chan, X. Cai, Y. Yang, D.J. Hayton, J.-R. Gao, J.L. Reno, and Q. Hu, "Terahertz laser frequency combs," Nature (London) **492**, 229–233 (2012).

[40] S. Barbieri, W. Maineult, S. S. Dhillon, C. Sirtori, J. Alton, N. Breuil, H. E. Beere, and D. A. Ritchie, "13 GHz direct modulation of terahertz quantum cascade lasers," Appl. Phys. Lett. **91**, 143510 (2007).

[41] C. Y. Wang, L. Kuznetsova, V. M. Gkortsas, L. Diehl, F. X. Kartner, M. A. Belkin, A. Belyanin, X. Li, D. Ham, H. Schneider, P. Grant, C. Y. Song, S. Haffouz, Z. R. Wasilewski, H. C. Liu, and F. Capasso, "Mode-locked pulses from mid-infrared Quantum Cascade Lasers," Optics Express **17**, 12929 (2009).