APPLIED PHYSICS REVIEWS

Modeling techniques for quantum cascade lasers

Christian Jirauschek[1] and Tillmann Kubis[2]

[1]Institute for Nanoelectronics, Technische Universität München, D-80333 Munich, Germany
[2]Network for Computational Nanotechnology, Purdue University, 207 S Martin Jischke Drive, West Lafayette, Indiana 47907, USA

Quantum cascade lasers are unipolar semiconductor lasers covering a wide range of the infrared and terahertz spectrum. Lasing action is achieved by using optical intersubband transitions between quantized states in specifically designed multiple-quantum-well heterostructures. A systematic improvement of quantum cascade lasers with respect to operating temperature, efficiency, and spectral range requires detailed modeling of the underlying physical processes in these structures. Moreover, the quantum cascade laser constitutes a versatile model device for the development and improvement of simulation techniques in nano- and optoelectronics. This review provides a comprehensive survey and discussion of the modeling techniques used for the simulation of quantum cascade lasers. The main focus is on the modeling of carrier transport in the nanostructured gain medium, while the simulation of the optical cavity is covered at a more basic level. Specifically, the transfer matrix and finite difference methods for solving the one-dimensional Schrödinger equation and Schrödinger-Poisson system are discussed, providing the quantized states in the multiple-quantum-well active region. The modeling of the optical cavity is covered with a focus on basic waveguide resonator structures. Furthermore, various carrier transport simulation methods are discussed, ranging from basic empirical approaches to advanced self-consistent techniques. The methods include empirical rate equation and related Maxwell-Bloch equation approaches, self-consistent rate equation and ensemble Monte Carlo methods, as well as quantum transport approaches, in particular the density matrix and non-equilibrium Green's function formalism. The derived scattering rates and self-energies are generally valid for n-type devices based on one-dimensional quantum confinement, such as quantum well structures. © 2014 AIP Publishing LLC. [http://dx.doi.org/10.1063/1.4863665]

TABLE OF CONTENTS

GLOSSARY OF SYMBOLS

$\hat{1} =$ Unity operator
$1D =$ One-dimensional
$3D =$ Three-dimensional
$A =$ Spectral function
$a =$ Power loss coefficient
$a_m =$ Mirror or outcoupling power loss coefficient
$a_w =$ Waveguide power loss coefficient
$CPU =$ Central processing unit
$C =$ Ideal midpoint of a rough interface of two adjacent materials
$c =$ Vacuum speed of light
$D_{TO} =$ Optical deformation potential field component $\parallel \mathbf{u}$
$D^R =$ Retarded environmental Green's function
$D^< =$ Lesser environmental Green's function
$d_{ij} =$ Optical dipole matrix element for transition from level i to j
$\mathbf{E} =$ Electric field vector
$E_\xi =$ Electric field component ($\xi = x, y, z$)
$\hat{\mathbf{E}} =$ Electric field amplitude vector
$\hat{E}_\xi =$ Component of $\hat{\mathbf{E}}$ ($\xi = x, y, z$)
$E =$ Energy
$E' =$ Energy above conduction band edge, $E' = E - V_c$
$E_i/E_{i\mathbf{k}} =$ Eigenenergy of state $|i\rangle/|i\mathbf{k}\rangle$
$\tilde{E}_i =$ Energy of state i relative to conduction band edge
$E_i^F =$ Quasi Fermi level of subband i
$E_p =$ Bias drop over a single QCL period
$E(x) =$ Complete elliptical integral of the second kind
$EMC =$ Ensemble Monte Carlo
$e =$ Elementary charge
$f =$ Distribution function
$f_i =$ Carrier distribution function in subband i

$f_i^{FD} =$ Fermi-Dirac distribution function
$f_i^{MB} =$ Maxwell-Boltzmann distribution function
$G^R =$ Retarded electron Green's function
$G^< =$ Lesser electron Green's function
$g =$ Power gain coefficient
$g_{th} =$ Threshold power gain coefficient
$\mathbf{H} =$ Magnetic field vector
$\hat{\mathbf{H}} =$ Magnetic field amplitude vector
$\hat{H} =$ Total Hamilton operator
$\hat{H}_0 =$ Free electron Hamilton operator
$H_\xi =$ Magnetic field component ($\xi = x, y, z$)
$\hat{H}_\xi =$ Component of $\hat{\mathbf{H}}$ ($\xi = x, y, z$)
$\hbar =$ Reduced Planck constant
$I =$ Optical intensity
$\mathcal{J} =$ Imaginary part
$I_0 =$ Modified Bessel function
$I_m =$ Optical intensity in mode m
$J =$ Current density
$\hat{J} =$ Current density operator
$\mathbf{k} =$ In-plane wave vector, $\mathbf{k} = [k_x, k_y]^T$
$\mathbf{k}' =$ Final state in-plane wave vector, $\mathbf{k}' = [k_x', k_y']^T$
$k_0 =$ Vacuum wave number ω/c
$k_{21} =$ Wave number $\omega_{21} n_0/c$
$k_n =$ Wave number (of ψ or $\hat{H}_y$) in segment n
$\tilde{k}_n = k_n/m_n^*$ (for ψ); $k_n/\epsilon_r^{(n)}$ (for $\hat{H}_y$)
$k_B =$ Boltzmann constant
$L =$ Resonator length
$L_p =$ Length of a QCL period
$\mathcal{L} =$ Lorentzian lineshape function
$\mathcal{L}_{ij} =$ Lorentzian lineshape function for the transition from subband i to j
$LA =$ Longitudinal acoustic
$LO =$ Longitudinal optical

$m^{\|}=$ In-plane effective mass
$m^{*}=$ Γ valley effective mass
$m^{*}=$ Effective mass in the growth direction
$N_{e}=$ Number of electrons
$N_{p}=$ Number of periods
$N_{Ph}=$ Mode independent phonon occupation number
$N_{Q}=$ Phonon occupation number of a mode with wave vector $\mathbf{Q}$
NEGF $=$ Nonequilibrium Green's function
$n=$ Electron density
$n_{0}=$ Refractive index
$n_{D}=$ Donor concentration
$n_{eff}=$ Effective refractive index β/k_{0}
$n^{s}=$ Total sheet density per QCL period
$n_{i}^{s}=$ Electron sheet density of level i
$n_{i}^{E}=$ Density of states per unit area and energy in subband i, $n_{i}^{E}=m_{i}^{\|}/(\pi\hbar^{2})$
$P=$ Probability to find a single impurity
Pr $=$ Principal value integral
$\mathbf{Q}=$ Phonon wave vector $[\mathbf{q}, q_{z}]^{T}$
$\mathbf{q}=$ In-plane component of $\mathbf{Q}$, exchanged wave vector
$q_{s}=$ Inverse screening length
QCL $=$ Quantum cascade laser
$R=$ Facet reflectance
$\Re=$ Real part
RPA $=$ Random phase approximation
$\mathbf{r}=$ In-plane position vector $(x, y)^{T}$
$s=$ Numerical grid spacing
$S=$ In-plane cross section area
$S_{g}=$ Gain medium cross section area
$T=$ Facet transmittance
$T_{e}=$ Electron temperature
$T_{i}=$ Electron temperature in subband i
$T_{L}=$ Lattice temperature
TA $=$ Transverse acoustic
TO $=$ Transverse optical
t $=$ Time
$\mathbf{u}=$ Displacement vector for lattice vibrations
$V=$ Potential
$\tilde{V}=$ Electrostatic potential energy
$V_{c}=$ Conduction band edge potential
$V_{ii_{0}jj_{0}}^{b}=$ Bare Coulomb matrix elements
$V_{ii_{0}jj_{0}}^{s}=$ Screened Coulomb matrix element
$V_{imp}=$ Impurity scattering potential
$V_{IR}=$ Interface roughness potential
$V_{o}=$ Conduction band offset
$\hat{v}=$ Velocity operator
$v_{s}=$ Longitudinal sound velocity
$V_{j\mathbf{k}',i\mathbf{k}}=$ Matrix elements for elastic scattering processes
$V_{j\mathbf{k}',i\mathbf{k}}^{\pm}=$ Matrix elements for emission/absorption processes
$W_{i\mathbf{k},j\mathbf{k}'}=$ Transition rate from an initial state $|i\mathbf{k}\rangle$ to a final state $|j\mathbf{k}'\rangle$
$W_{i\mathbf{k},j\mathbf{k}'}^{\pm}=$ Emission/absorption rate from an initial state $|i\mathbf{k}\rangle$ to a final state $|j\mathbf{k}'\rangle$
$W_{ij}^{opt}=$ Stimulated transition rate from subband i to j
$\mathbf{x}=$ Position vector $[x, y, z]^{T}$
$\hat{x}=$ Position operator
$x=$ In-plane coordinate in light propagation direction
$x=$ Alloy mole fraction

$y=$ In-plane coordinate
$z=$ Coordinate in growth direction
$\alpha=$ Optical field amplitude absorption coefficient
$\alpha'=$ Nonparabolicity parameter
$\beta=$ Complex propagation constant
$\beta'=$ Nonparabolicity parameter
$\Gamma=$ Overlap factor
$\tilde{\Gamma}=$ Scattering rate
$\gamma_{E}=$ Energy relaxation rate
$\gamma_{i,j}=$ Optical linewidth of transition from level i to j
$\delta\hat{V}=$ Photonic perturbation potential
$\Delta=$ Standard deviation of interface roughness
$\Delta_{21}=$ Inversion, $\Delta_{21}=\rho_{22}-\rho_{11}$
$\Delta_{21}^{eq}=$ Equilibrium inversion
$\Delta_{n}=$ Length of nth segment; thickness of nth layer
$\Delta_{z}=$ Uniform grid spacing
$\epsilon=$ Permittivity, $\epsilon=\epsilon_{0}\epsilon_{r}$
$\epsilon_{0}=$ Vacuum permittivity
$\epsilon_{r}=$ Dielectric constant
$\epsilon_{r,0}=$ Static dielectric constant
$\epsilon_{r,\infty}=$ Dielectric constant at very high frequencies
$\phi=$ Potential drop per period
$\Phi=$ Electrostatic potential
$\eta_{21}=$ Slowly varying envelope function of ρ_{21}
$\Lambda=$ Interface roughness correlation length
$\mu=$ Chemical potential
$\mu_{0}=$ Vacuum permeability
$\Xi=$ Deformation potential
$\rho=$ Space charge
$\rho_{c}=$ Density of crystal
$\rho_{ij}=$ Density matrix element
$\rho_{N}=$ Number density of crystal
$\Sigma^{R}=$ Retarded self-energy
$\Sigma^{<}=$ Lesser self-energy
$\sigma=$ Complex conductance
$\tau_{i}=$ Lifetime of level i
$\tau_{ji}=$ Inverse transition rate from subband j to i
$\psi_{i}=$ Wave function of subband i
$\Omega_{ij}=$ Rabi frequency for transition from level i to j
$\Omega_{c}=$ Crystal volume
$\omega=$ Angular frequency
$\omega_{D}=$ Debye frequency
$\omega_{ij}=$ Resonance frequency for transition from subband i to j, $(E_{i}-E_{j})/\hbar$
$\omega_{LO}=$ LO phonon frequency at $Q=0$
$\omega_{TO}=$ TO phonon frequency at $Q=0$
$[\hat{H},\hat{x}]_{-}=$ Commutator of Hamilton operator $\hat{H}$ and position operator $\hat{x}$, $\hat{H}\hat{x}-\hat{x}\hat{H}$

I. INTRODUCTION

The quantum cascade laser (QCL) is a fascinating device, combining aspects from different fields such as nanoelectronics and quantum engineering, plasmonics, as well as subwavelength and nonlinear photonics. Since its first experimental realization in 1994,[1] rapid progress has been achieved with respect to its performance, in particular the output power and the frequency range covered. The development of innovative types of QCLs and subsequent design

optimization has gone hand in hand with detailed modeling, involving more and more sophisticated simulation tools. The advancement of simulation methods is driven by an intrinsic motivation to describe the underlying physical processes as exactly and realistically as possible. In addition to scientific motivations, improved agreement with experimental data and enhanced predictive power is desired. Furthermore, the extension and improvement of QCL simulation methods is largely driven by experimental progress, such as innovative designs.

The QCL is a special type of semiconductor laser where the optical gain medium consists of a multiple quantum well heterostructure, giving rise to the formation of quantized electron states. These so-called subbands assume the role of the laser levels in the QCL. To date, lasing has only been obtained for n-type QCLs, i.e., designs using intersubband transitions in the conduction band. Well established material systems are InGaAs/InAlAs on InP substrate, and GaAs/AlGaAs on GaAs.[2,3] Furthermore, antimonide QCLs have been demonstrated, e.g., near- and mid-infrared QCLs using InAs/AlSb[4,5] and aluminum-free InGaAs/GaAsSb QCLs operating in the mid-infrared[6] and terahertz[7] regime. Since all these materials (apart from AlSb) have direct bandgaps, the development of simulation methods and tools for QCLs has focused on the conduction band Γ valley where the lasing transitions take place.

In contrast to conventional semiconductor lasers where lasing occurs due to electron-hole recombination between conduction and valence band, the emission wavelength of QCLs is not determined by the bandgap of the material. Rather, by adequately designing the nanostructured gain medium, the device can be engineered to lase at a certain frequency. In this way, the mid/far infrared and terahertz regimes become also accessible, important for applications such as chemical and biological sensing, imaging, and special communication applications. In addition, the QCL offers the typical advantages of semiconductor devices, such as compactness, reliability, and potentially low production costs.[8] The basic working principle of the QCL is illustrated in Fig. 1. Lasing is obtained due to stimulated emission between the upper and the lower laser level, here resulting in mid-infrared radiation at a wavelength of 5 μm. The depopulation level is separated from the lower laser level by the longitudinal optical (LO) phonon energy for InGaAs of $\sim$30 meV, enabling efficient depopulation of the lower laser level. The injector level, located between two adjacent periods, ensures electron transport and injection into the upper laser level of the next period. The optical field confinement and beam shaping is provided by a specifically designed resonator, which is frequently based on plasmonic effects and can exhibit subwavelength structuring for a further performance improvement.

The QCL operating principle was devised in 1971 by Kazarinov and Suris.[10] However, due to the experimental challenges involved, it took more than two decades until QCL operation was experimentally demonstrated at Bell Labs by Capasso and his co-workers and Faist *et al*.[1] Since then, considerable progress has been achieved, e.g., with respect to operating temperature,[11] output power,[12] and the

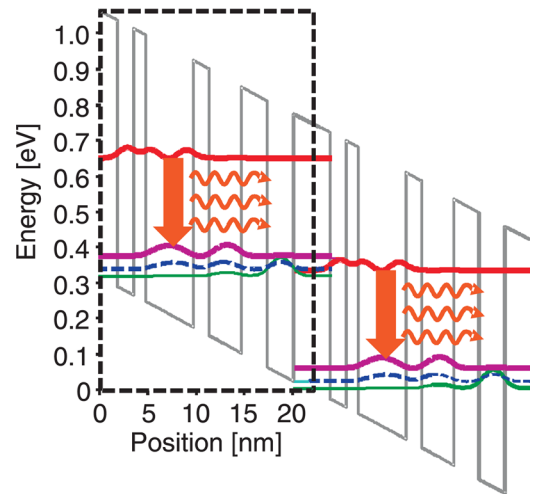FIG. 1. Conduction band profile and probability densities for a mid-infrared QCL[9] lasing at 5 μm. Only the relevant energy states are displayed. The upper and lower laser levels are indicated by bold solid lines. Furthermore, the depopulation (dashed line) and injector (thin solid line) state probability densities are shown. The rectangle denotes a single QCL period.

available frequency range.[5,13] Meanwhile, QCLs cover a spectral range from 2.6 μm to above 400 μm (obtained by applying an external magnetic field).[5,13] At room temperature, continuous wave output powers of several W can now routinely be obtained in the mid-infrared region,[12] and widely tunable designs are available.[14] Furthermore, at cryogenic temperatures wall-plug efficiencies of 50% and above have been demonstrated,[9,15] and up to 27% at room temperature.[12] Commercial QCLs and QCL-based systems are already available from various companies.[8]

On the other hand, terahertz QCLs, first realized in 2002,[16] still require cryogenic operating temperatures. Lasing up to $\sim$200 K has been obtained,[17] and by applying a strong external magnetic field, an operating temperature of 225 K has been demonstrated.[18] A long standing goal is to achieve operation at room temperature or at least in the commercially available thermoelectric cooler range ($\sim$240 K). For a further enhancement of the operating temperature, it is crucial to improve the gain medium design and reduce the resonator loss.[17,19] For systematic QCL optimization and exploration of innovative designs, detailed modeling is essential, requiring advanced simulation approaches which enable realistic QCL simulations.

Besides its function as the optical gain medium, the QCL quantum well heterostructure can also exhibit a strong nonlinear optical response. Like the laser transition itself, such nonlinearities are based on quantized electron states, and thus can be custom-tailored for specific applications and optimized to exhibit extremely high nonlinear susceptibilities. This has led to a new paradigm for QCL-based THz generation at room temperature: Here, difference frequency mixing is used, where the THz radiation corresponds to the difference in frequency of two detuned mid-infrared beams, generated by highly efficient mid-infrared QCLs.[20] The optical nonlinearity is either implemented into the QCL separately in a specifically designed heterostructure,[21] or, more commonly, directly integrated into the gain medium.[20] With such an approach,

THz generation at room temperature was obtained at power levels of up to $120\,\mu$W.[22,23] Furthermore, broadband frequency tunability has been demonstrated, which is important for various applications such as spectroscopy.[23,24] Here, the main goal is to push the available room temperature output power to a few mW, as required for most technical applications. The modeling of such terahertz QCL sources is particularly demanding, requiring the coupled simulation of two mid-infrared QCLs and a careful modeling of the nonlinear susceptibility and the associated frequency conversion process in the heterostructure.[25] Artificial optical nonlinearities have also been used to extend QCL operation towards shorter wavelengths, which cannot be reached directly since the energy spacing of the quantized laser levels is limited by the quantum well depth of the gain medium. Based on a frequency doubling structure,[26] a room temperature QCL source operating at $2.7\,\mu$m could be demonstrated.[27]

The ongoing development and optimization of the QCL heterostructure is accompanied by improvements of the optical cavity. For example, changes in the design and material of the plasmonic resonator structure have played a crucial role for increasing the operating temperature of THz QCLs.[17,19] Also, the performance of difference frequency THz sources has largely benefited from special cavity designs, such as the surface emitting Cherenkov waveguide scheme to obtain increased output power and efficiency.[23,28] Thus, a quantitative modeling and reliable numerical design optimization and exploration must also include the optical cavity. Moreover, various types of resonators with periodic subwavelength structuring have been developed in the THz and mid-infrared regime. This includes distributed feedback structures to enforce single mode lasing and distributed Bragg reflectors for enhanced facet reflectivity.[29–31] A further example are surface emission schemes based on one- and two-dimensional photonic crystal structures,[29,32,33] offering tailorable emission properties and improved beam quality. The simulation of such subwavelength-structured cavities requires advanced electromagnetic modeling, e.g., based on coupled mode theory[34] or even full finite difference time domain simulations of Maxwell's equations.[33,35]

The goal of this review is to give a detailed survey and discussion of the modeling techniques used for QCL simulation, ranging from basic empirical approaches to advanced self-consistent simulation methods. The focus is here on the modeling of the heterostructure gain medium. Also the simulation of the optical cavity will be covered for simple resonator waveguide structures. As mentioned above, the modeling of complex cavities such as photonic crystal structures constitutes an advanced electromagnetic modeling task, which is beyond the scope of this paper. The review is organized as follows: In Sec. II, the numerical solution of the one-dimensional Schrödinger equation is discussed, providing the eigenenergies and wave functions of the energy states in the QCL heterostructure. Furthermore, the inclusion of space charge effects by solving the Schrödinger-Poisson equation system is treated. Section III covers the modeling of the optical resonator, where we focus on a basic waveguide resonator structure. In Sec. IV, an overview and classification of the different carrier transport models is given which are

commonly used for the theoretical description of the QCL gain medium. Section V contains a discussion of empirical modeling approaches for the gain medium, relying on experimental or empirical input parameters, namely, empirical rate equation and related Maxwell-Bloch equation approaches, which are discussed in Secs. V A and V B, respectively. Advanced self-consistent methods, which only require well known material parameters as input, are covered in Secs. VI–VIII: Semiclassical approaches such as the self-consistent rate equation model (Sec. VI E) and the ensemble Monte Carlo method (Sec. VI F), as well as quantum transport approaches such as the density matrix method (Sec. VII) and the non-equilibrium Green's function formalism (Sec. VIII). In this context, also the transition rates (Secs. VI A–VI D) and self-energies (Sec. VIII D) are derived for the relevant scattering processes in QCLs. The paper is concluded in Sec. IX with an outlook on future trends and challenges.

II. SCHRÖDINGER-POISSON SOLVER

A careful design of the quantized states in the QCL heterostructure is crucial for the development and optimization of experimental QCL devices. In particular, the lasing frequency is determined by the energy difference between the upper and lower laser level. Furthermore, a careful energy alignment of the levels is necessary to obtain an efficient injection into the upper and depopulation of the lower laser level. For example, for the structure shown in Fig. 1, the depopulation level is separated from the lower laser level by the InGaAs LO phonon energy of $\sim$30 meV to obtain efficient depopulation, while the injector level is aligned with the upper laser level of the next period.[9] Also careful engineering of the wave functions is important, determining the strength of both the optical and nonradiative transitions. Moroever, various carrier transport simulation approaches, such as ensemble Monte Carlo (EMC), require the quantized energy levels as input. The eigenenergies and wave functions are determined by solving the stationary Schrödinger equation or, if space charge effects are taken into account, the Schrödinger-Poisson equation system.

A. Schrödinger equation

The QCL heterostructure consists of alternately grown thin layers of different semiconductor materials, resulting in the formation of quantum wells and barriers in the conduction band. Thus, the potential V and effective masses vary in the growth direction, here denoted by z, whereas x and y refer to the in-plane directions. The QCL heterostructure is usually treated in the framework of the Ben Daniel-Duke model which only considers the conduction band,[36] where the lasing transitions and carrier transport take place. Within this approximation, the stationary Schrödinger equation is given by

$$0 = \left\{ -\frac{\hbar^2}{2} \left[\frac{1}{m^{\parallel}(z)} \left(\partial_x^2 + \partial_y^2 \right) + \partial_z \frac{1}{m^*(z)} \partial_z \right] \right.$$
$$\left. + V(z) - E \right\} \psi_{3D}(x, y, z), \tag{1}$$

where ψ_{3D} and E denote the wave function and eigenenergy, respectively. Furthermore, $m^{\parallel}$ refers to the in-plane effective mass and m^* is the effective mass in the growth direction, i.e., perpendicular to epitaxial layers. For bound states, the wave function is commonly normalized, i.e., $\int\int\int_{-\infty}^{\infty} |\psi_{3D}|^2 dxdydz = 1$. Since V and the effective masses only depend on the z coordinate, we can make the ansatz

$$\psi_{3D}(x, y, z) = S^{-1/2} \psi_{\mathbf{k}}(z) \exp(ik_x x + ik_y y). \tag{2}$$

Here, S is the in-plane cross section area and $\mathbf{k} = [k_x, k_y]^T$ denotes the in-plane wave vector, where T indicates the transpose. The factor $S^{-1/2}$ is added in Eq. (2) to obtain the normalization condition $\int |\psi_{\mathbf{k}}|^2 dz = 1$. Insertion of Eq. (2) in Eq. (1) yields the Ben Daniel-Duke model

$$\left\{ \frac{\hbar^2}{2} \frac{k_x^2 + k_y^2}{m^{\parallel}(z)} - \frac{\hbar^2}{2} \partial_z \frac{1}{m^*(z)} \partial_z + V(z) - E_{\mathbf{k}} \right\} \psi_{\mathbf{k}}(z) = 0, \tag{3}$$

where the wave function $\psi_{\mathbf{k}}(z)$ and energy $E_{\mathbf{k}}$ depend on the in-plane electron motion, i.e., on $\mathbf{k}$. Decoupling can be obtained by neglecting the z dependence of the in-plane effective mass $m^{\parallel}$, yielding the one-dimensional (1D) Schrödinger equation in its usual form

$$\left[-\frac{\hbar^2}{2} \partial_z \frac{1}{m^*(z)} \partial_z + V(z) - E \right] \psi(z) = 0. \tag{4}$$

The eigenenergy E due to 1D electron confinement in z direction is related to the total energy $E_{\mathbf{k}}$ by $E_{\mathbf{k}} = E + E_{\mathrm{kin}}$, where

$$E_{\mathrm{kin}} = \hbar^2(k_x^2 + k_y^2)/(2m^{\parallel}) = \hbar^2 \mathbf{k}^2/(2m^{\parallel}) \tag{5}$$

is the kinetic electron energy due to the free in-plane motion of the electrons. The effective mass values not only depend on the material composition[37] but also on the lattice temperature and doping level.[38] The latter effects tend to play a secondary role in QCLs and are thus usually neglected. For strained QCL structures, the effective masses are additionally affected by the lattice mismatch between the different semiconductor materials, resulting in modified values $m^{\parallel} \neq m^*$.[39]

B. Boundary conditions

Strictly speaking, the quantum states in the QCL heterostructure are not bound, since the electron energy E exceeds the barrier potential for large values of z. Thus, the electrons will tunnel out of the multiple quantum well system after a limited time. This situation is illustrated in Fig. 2. If the electron remains in the quantum well system for a considerable amount of time, the concept of quasi-bound or quasi-
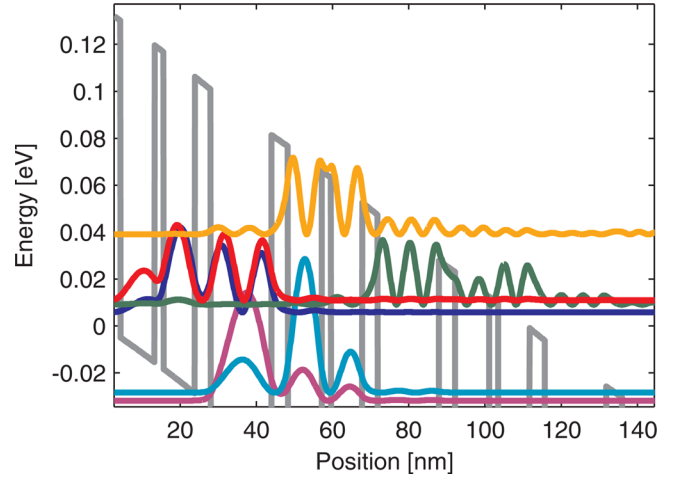

FIG. 2. Quasi-bound states in the conduction band of a QCL heterostructure.

stationary states can be used.[40] Such a state can again be described by Eq. (4), however the eigenenergy is now a complex quantity, $E = E_0 - i\hbar\gamma_q/2$. The meaning of the imaginary term $-i\hbar\gamma_q/2$ for quasi-stationary states can be understood by considering that the time dependence of a stationary wave function ψ is given by $\exp(-iEt/\hbar)$. Consequently, for a complex eigenenergy E the density probability $|\psi|^2$ of the electron in the quantum well system decays as $\exp(-\gamma_q t)$, thus γ_q is the probability density decay rate of the quasi-bound state, and $\tau_q = \gamma_q^{-1}$ can be associated with the lifetime of the particle in the quantum well.[40]

Some numerical approaches have been developed to solve Eq. (4) for quasi-bound states.[41,42] However, the relevant states which are considered for the design of a QCL structure, such as the upper and lower laser levels and injector states, are typically strongly bound to obtain optimum performance. Since the electron lifetime in these states is governed by scattering processes and resonant tunneling rather than τ_q, they are usually treated as bound states. This can be done by restricting the simulation to a finite simulation window containing a limited number of periods, and imposing artificial boundary conditions $\psi = 0$ at the borders, as illustrated in Fig. 3. The simulation window has to be chosen sufficiently large so that the portion of the wave function which lies inside the quantum wells, i.e., which significantly deviates from 0, is not markedly affected. For the simulation, we assume that the QCL heterostructure is periodic, where a period is defined by a sequence of multiple barriers and wells, as illustrated in Fig. 1. Thus, for a biased QCL structure, it is sufficient to compute the eigenenergies and corresponding wave functions for a single energy period of width E_p, corresponding to the bias drop over a QCL period (see Fig. 3). The solutions of the other periods can then simply be obtained by shifts in position by multiples of the period length L_p, and corresponding shifts in energy. For example, the solutions in the right-neighboring period are given by $\psi_{i'}(z) = \psi_i(z - L_p)$, $E_{i'} = E_i - E_p$, where ψ_i and E_i are the wave function and eigenenergy of the ith solution of Eq. (4) obtained in the central period.

For the laser design and simulation, usually only the relevant laser subbands such as the upper and lower laser level
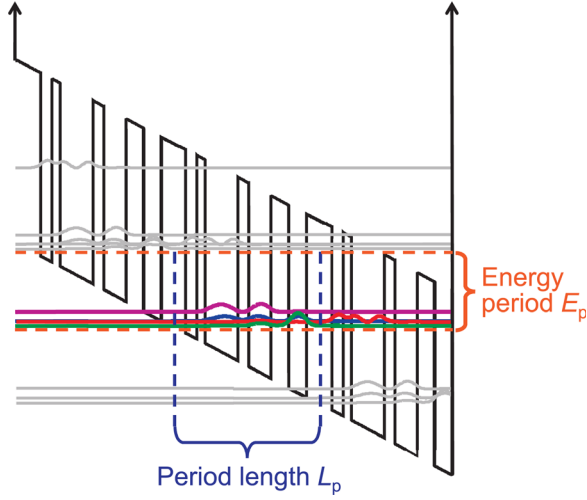
FIG. 3. Finite simulation window with artificial boundary conditions $\psi = 0$. The central QCL period of the simulated structure is indicated, as well as the energy interval used for determining the bound states, corresponding to an energy period. The states in the other periods are found by adequate shifts in position and energy.
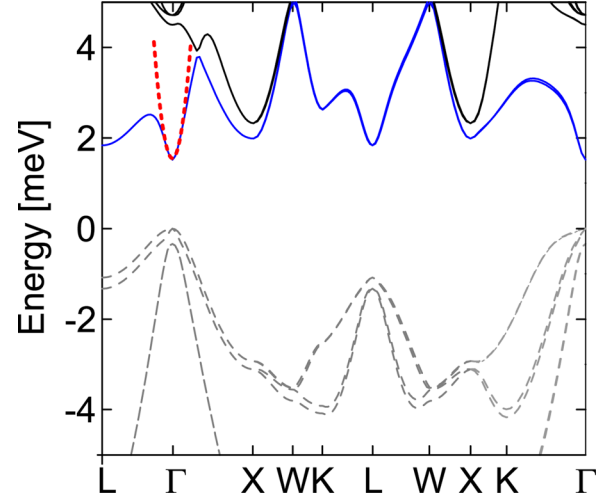

FIG. 4. Band structure of gallium arsenide (GaAs), obtained with NEMO5 in 20 band empirical tight binding.[44] Shown are the valence bands (dashed lines), the conduction band (solid line), and the parabolic dispersion relation used for the Γ valley (dotted line).

and injector states are considered. These typically correspond to the strongly bound states of the QCL heterostructure. Furthermore, the artificial boundary conditions illustrated in Fig. 3 give rise to spurious solutions of Eq. (4), which are not localized in the quantum wells. A systematic selection of the relevant subbands can be achieved by considering only the most strongly bound levels in each period. In this context, the energy of state i relative to the conduction band edge

$$\tilde{E}_i = E_i - \int V|\psi_i|^2 \mathrm{d}z \qquad (6)$$

is a useful quantity.[43] An automated selection of the relevant subbands can then be implemented by choosing the states with the lowest energies $\tilde{E}_i$ in each period, considering all subbands which contribute significantly to the carrier transport.

C. Nonparabolicity

In Fig. 4, the band structure of GaAs as a typical III–V semiconductor material is displayed. Shown is the valence band (dashed lines), consisting of heavy hole, light hole, and split-off bands, and the conduction band (solid line). As pointed out above, only the conduction band is considered in the Ben Daniel-Duke model, introduced in Sec. II A. The conduction band has three minima, referred to as the Γ, L, and X (or Δ) valley according to their position in **K** space. For direct bandgap semiconductors typically used in QCLs, the Γ valley is the lowest minimum. Thus, the theoretical treatment is usually restricted to that valley, although under certain conditions transitions to the X or L valleys can also affect QCL operation.[45–48] Furthermore, Eqs. (1)–(5) assume a parabolic dispersion relation between energy and wave number (dotted line in Fig. 4). This approximation only holds close to the Γ valley minimum, i.e., it breaks down for high-lying energy levels. The deviation from the parabolic

dispersion, i.e., the nonparabolicity, scales roughly inversely with the semiconductor bandgap.[49] This effect thus plays a role especially for mid-infrared QCLs with lasing transition energies of up to a few 100 meV, which are frequently based on low-bandgap semiconductors such as InGaAs or InGaSb. A detailed treatment of nonparabolicity effects in QCLs has been performed by using **k·p** theory.[45–47,50] A related strategy, which we will discuss in the following, is to consider nonparabolicity in Eq. (4) by implementing an energy dependent mass, derived from **k·p** theory. This method involves additional approximations, but is less complex than full **k·p** calculations and can easily be incorporated into the transfer matrix approach, which is widely used to solve Eq. (4). According to Ekenberg,[51] we obtain the energy dependent effective masses

$$m^*(E') = \frac{m^*}{2\alpha'E'}[1 - (1 - 4\alpha'E')^{1/2}], \qquad (7a)$$

$$m^{\|}(E') = m^*[1 + (2\alpha' + \beta')E']. \qquad (7b)$$

Here, α' and β' are nonparabolicity parameters, which are defined in the framework of a 14-band **k·p** calculation.[51,52] Approximately, we have $\alpha' = \left(E_g + \Delta_{so}/3\right)^{-1}$,[49] where E_g and Δ_{so} are the bandgap energy and the energy difference between the light-hole and split-of valence bands, respectively. Furthermore, m^* is the Γ valley effective mass. In the QCL heterostructure, the electron kinetic energy in the conduction band is given by $E'(z) = E - V(z)$. Since the semiconductor material changes along the growth direction, also m^*, α', and β' are z dependent. This approach is valid for moderate energies $E' \ll 1/\alpha'$. The fact that Eq. (7a) even breaks down for $\alpha'E' > 1/4$ can cause numerical problems for large values of z close to the right simulation boundary, where ψ is already close to 0 but E' can assume relatively large values (see Fig. 3). This issue can be avoided by using Eq. (7a) for $E' < 0$, and its second order Taylor expansion

$$m^*(E') = m^*(1 + \alpha'E'), \tag{8}$$

for $E' > 0$, which is the widely used lowest order implementation of nonparabolicity.[49,51] For an energy dependent effective mass, the Hamiltonian in Eq. (4) is not Hermitian; thus, the obtained wave functions are in general not orthogonal.[53] Also other approaches are available for implementing nonparabolicity effects, for example based on the Kane model.[53]

In self-consistent simulations, each scattering process is evaluated based on the corresponding Hamiltonian, which depends on the in-plane effective masses of the subbands involved. Thus, if we want to include corrections due to nonparabolicity also for the scattering processes,[54,55] then Eq. (7b) should be used to determine $m^{\parallel}(E')$, which is however z dependent. This problem can be avoided by defining an average effective mass for the ith subband

$$m_i^{\parallel} = \int m^*(z)\{1 + [2\alpha'(z) + \beta'(z)][E_i - V(z)]\}|\psi_i(z)|^2 dz, \tag{9}$$

where the wave function is assumed to be normalized, $\int |\psi_i|^2 dz = 1$, and E_i denotes the corresponding subband eigenenergy. This lowest order implementation of nonparabolicity thus yields different, albeit constant in-plane masses for each subband.

D. Numerical solution

Numerical approaches for solving the one-dimensional effective mass Schrödinger equation [Eq. (4)] are required to be robust. Also computational efficiency is crucial, especially for QCL design and optimization tasks where many simulations have to be performed. Furthermore, a straightforward implementation is desirable. Widely used numerical approaches include the transfer matrix method[41,43,56,57] and finite difference scheme.[58,59] Both methods have their strengths and shortcomings. In particular, effects such as nonparabolicity can be included more easily into the transfer matrix approach. A further advantage is the exact treatment of the potential steps between barriers and wells in the QCL heterostructure. On the other hand, this method can exhibit numerical instabilities for multiple or extended barriers due to an exponential blowup caused by roundoff errors.[58] This issue can however be overcome, for example by using a somewhat modified approach, the scattering matrix method.[60] In Fig. 5, the transfer matrix and finite difference schemes are illustrated.

1. Transfer matrix approach

The transfer matrix approach uses the fact that analytical solutions of Eq. (4) are available for constant or linear potential sections and for potential steps.[61] An arbitrary potential can then be treated by approximating it in terms of piecewise constant or linear segments, respectively. In the first case, the solution is given by complex exponentials,[41,56] while the approximation by linear segments gives rise to Airy function solutions.[41] The Airy function approach provides an exact
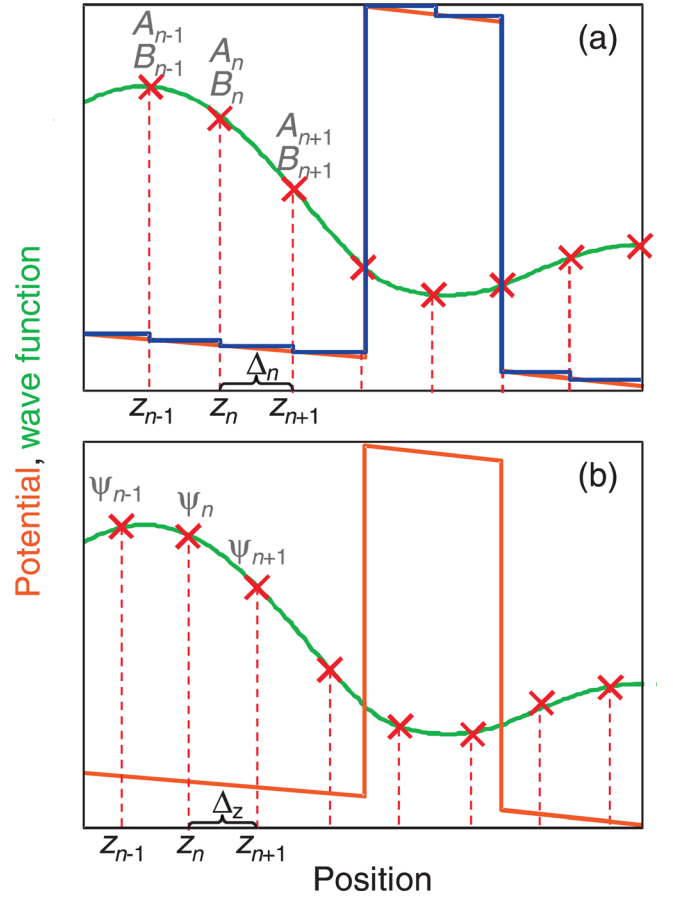
FIG. 5. Numerical solution of the one-dimensional effective mass Schrödinger equation: (a) Transfer matrix approach and (b) finite difference method.

solution for structures which consist of segments with constant effective masses and piecewise linear potentials, such as a biased QCL heterostructure if we neglect nonparabolicity and space charge effects. On the other hand, Airy functions are much more computationally expensive than exponentials, and also prone to numerical overflow for segments with nearly flat potential.[62] Thus, great care has to be taken to avoid numerical problems and to evaluate the Airy functions efficiently.[63]

In the following, we focus on the exponential transfer matrix scheme, illustrated in Fig. 5(a). We start by dividing the structure into segments which can vary in length, but should be chosen so that all band edge discontinuities occurring between wells and barriers are located at the border of a segment, i.e., do not lie within a segment. The potential and effective mass in each segment n with $z_n \leq z < z_n + \Delta_n = z_{n+1}$ are approximated by constant values, e.g., $V_n = V(z_n)$, $m_n^* = m^*(z_n)$, resulting in a jump $V_n \rightarrow V_{n+1}$, $m_n^* \rightarrow m_{n+1}^*$ at the border between the segments n and $n+1$.[41] Nonparabolicity can straightforwardly be implemented by using Eq. (7a) for m^*; then $m_n^* = m_n^*(E)$ depends on the eigenenergy E. The solution of Eq. (4) in segment n is given by

$$\psi(z) = A_n \exp[ik_n(z - z_n)] + B_n \exp[-ik_n(z - z_n)]. \tag{10}$$

Here, $k_n = \sqrt{2m_n^*(E - V_n)}/\hbar$ is the wave number (for $E < V_n$, we obtain $k_n = i\kappa_n = i\sqrt{2m_n^*(V_n - E)}/\hbar$).[41] The

matching conditions for the wave function at a potential or effective mass discontinuity read

$$\psi(z_0+) = \psi(z_0-),$$
$$[\partial_z\psi(z_0+)]/m^*(z_0+) = [\partial_z\psi(z_0-)]/m^*(z_0-), \quad (11)$$

where z_0+ and z_0- denote the positions directly to the right and left of the discontinuity.[61] Using Eqs. (10) and (11), the amplitudes A_{n+1} and B_{n+1} can be related to A_n and B_n by

$$\begin{pmatrix} A_{n+1} \\ B_{n+1} \end{pmatrix} = T_{n,n+1} \begin{pmatrix} A_n \\ B_n \end{pmatrix}, \quad (12)$$

where the transfer matrix is with $\tilde{k}_n = k_n/m_n^*$ given by

$$T_{n,n+1} = T_{n\to n+1}T_n(\Delta_n)$$
$$= \begin{pmatrix} \dfrac{\tilde{k}_{n+1}+\tilde{k}_n}{2\tilde{k}_{n+1}}e^{ik_n\Delta_n} & \dfrac{\tilde{k}_{n+1}-\tilde{k}_n}{2\tilde{k}_{n+1}}e^{-ik_n\Delta_n} \\ \dfrac{\tilde{k}_{n+1}-\tilde{k}_n}{2\tilde{k}_{n+1}}e^{ik_n\Delta_n} & \dfrac{\tilde{k}_{n+1}+\tilde{k}_n}{2\tilde{k}_{n+1}}e^{-ik_n\Delta_n} \end{pmatrix}. \quad (13)$$

Here,

$$T_n(\Delta_n) = \begin{pmatrix} e^{ik_n\Delta_n} & 0 \\ 0 & e^{-ik_n\Delta_n} \end{pmatrix} \quad (14)$$

corresponds to the transfer matrix for a flat potential obtained from Eq. (10), and

$$T_{n\to n+1} = \frac{1}{2\tilde{k}_{n+1}}\begin{pmatrix} \tilde{k}_{n+1}+\tilde{k}_n & \tilde{k}_{n+1}-\tilde{k}_n \\ \tilde{k}_{n+1}-\tilde{k}_n & \tilde{k}_{n+1}+\tilde{k}_n \end{pmatrix} \quad (15)$$

is the potential step matrix for the interface between the segments n and $n+1$ at $z_0 = z_{n+1}$, derived from Eq. (11).[61] For a structure divided into N segments, the relation between the amplitudes at the left and right boundaries of the structure, A_0, B_0 and A_N, B_N, can be obtained by multiplying the matrices for all segments

$$\begin{pmatrix} A_N \\ B_N \end{pmatrix} = T_{N-1,N}T_{N-2,N-1}\ldots T_{0,1}\begin{pmatrix} A_0 \\ B_0 \end{pmatrix}$$
$$= \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix}\begin{pmatrix} A_0 \\ B_0 \end{pmatrix}. \quad (16)$$

This equation must be complemented by appropriate boundary conditions. As described in Sec. II B, for the QCL heterostructure we can restrict our simulation to a limited number of periods (see Fig. 3) and assume $\psi = 0$ at the boundaries of our simulation window, corresponding to $A_0 + B_0 = 0$ and $A_N + B_N = 0$. The left boundary condition can, for example, be enforced by setting $A_0 = 1$, $B_0 = -1$. The right boundary condition $A_N + B_N = 0$ can then only be satisfied if the energy dependent matrix elements $T_{11}(E)$, $T_{12}(E)$, $T_{21}(E)$, and $T_{22}(E)$ in Eq. (16) assume certain values. The corresponding energies E are the eigenenergies of the bound states.[41] Numerically, these eigenenergies are found by the so-called shooting method. Here, the wave function at

the right boundary is computed from Eq. (16) as a function of energy, $\psi_N(E) = A_N(E) + B_N(E)$, and the eigenenergies are given by $\psi_N(E) = 0$. For the periodic QCL heterostructure, it is sufficient to restrict the simulation to a single energy period (see Fig. 3). In practice, Eq. (16) can only be solved for a limited number of discrete energy points $E_m = E_0 + m\Delta_E$ with sufficiently close spacing Δ_E. The eigenenergies are located in intervals $E_m..E_{m+1}$ with $\psi_N(E_m)\psi_N(E_{m+1}) < 0$, and can be determined more accurately by choosing a finer energy grid in the corresponding intervals, or preferably by applying a root-finding algorithm such as the bisection method.[64]

The accuracy of the transfer matrix scheme can generally be improved by replacing Eq. (13) with a symmetric transfer matrix

$$T_{n,n+1} = T_{n+1}\left(\frac{\Delta_n}{2}\right)T_{n\to n+1}T_n\left(\frac{\Delta_n}{2}\right)$$
$$= \begin{pmatrix} \dfrac{\tilde{k}_{n+1}+\tilde{k}_n}{2\tilde{k}_{n+1}}e^{ik_n^+\Delta_n} & \dfrac{\tilde{k}_{n+1}-\tilde{k}_n}{2\tilde{k}_{n+1}}e^{-ik_n^-\Delta_n} \\ \dfrac{\tilde{k}_{n+1}-\tilde{k}_n}{2\tilde{k}_{n+1}}e^{ik_n^-\Delta_n} & \dfrac{\tilde{k}_{n+1}+\tilde{k}_n}{2\tilde{k}_{n+1}}e^{-ik_n^+\Delta_n} \end{pmatrix}, \quad (17)$$

where $k_n^{\pm} = (k_n \pm k_{n+1})/2$ and $k_n = \sqrt{2m_n^*(E - V_n)}/\hbar$, $\tilde{k}_n = k_n/m_n^*$.[43] Here, the band edge discontinuities occurring between wells and barriers must be treated separately by using the corresponding transfer matrix Eq. (15).

2. Finite difference method

The finite difference method works by converting Eq. (4) to a finite difference equation. As illustrated in Fig. 5(b), a spatial grid with uniform spacing Δ_z is introduced, and the wave function $\psi(z)$, potential $V(z)$, and effective mass $m^*(z)$ are represented by the corresponding values ψ_n, V_n, and m_n^* on the grid points z_n. First order derivatives are approximated by $\partial_z\psi(z_n + \Delta_z/2) \approx \Delta\psi_{n+1/2} = (\psi_{n+1} - \psi_n)/\Delta_z$. Consequently, the term $\partial_z(m^*)^{-1}\partial_z\psi$ in Eq. (4) at $z = z_n$ can be expressed as $(\Delta\psi_{n+1/2}/m_{n+1/2}^* - \Delta\psi_{n-1/2}/m_{n-1/2}^*)/\Delta_z$. Using linear interpolation $m_{n+1/2}^* = (m_n^* + m_{n+1}^*)/2$, we obtain the discretized form of Eq. (4)[58,59]

$$-s_n\psi_{n-1} + d_n\psi_n - s_{n+1}\psi_{n+1} = E\psi_n, \quad (18)$$

with

$$s_n = \frac{\hbar^2}{\Delta_z^2(m_{n-1}^* + m_n^*)}, \quad (19a)$$

$$d_n = \frac{\hbar^2}{\Delta_z^2}\left(\frac{1}{m_{n-1}^* + m_n^*} + \frac{1}{m_n^* + m_{n+1}^*}\right) + V_n. \quad (19b)$$

Here, the grid should be chosen so that the band edge discontinuities occur halfway between two adjacent grid points,[58] as illustrated in Fig. 5(b). We again assume that the wave function is zero at the boundaries of our simulation window, $\psi_0 = \psi_N = 0$. Writing Eq. (18) in matrix form yields an eigenvalue equation of the form $(\mathbf{H} - E\mathbf{I})\psi = \mathbf{0}$, where ψ is

the wave function vector with $\psi = [\psi_1, \psi_2, ..., \psi_{N-1}]^T$, $\mathbf{I}$ represents the identity matrix of size $N - 1$, and $\mathbf{H}$ is the Hamiltonian matrix with the non-zero elements $H_{n,n} = d_n$, $H_{n,n-1} = -s_n$, $H_{n,n+1} = -s_{n+1}$. This equation can be solved by using a standard eigenvector solver for a tridiagonal matrix problem.[64] However, if nonparabolicity is taken into account, this is not directly possible since the effective masses m_n^* depend on E, and a modified numerical scheme must be employed.[65] Alternatively, the shooting method discussed in Sec. II D 1 can be employed.

E. Schrödinger-Poisson equation system

To lowest order, electron-electron interaction can be considered by self-consistently solving Eq. (4) together with the Poisson equation[57,66]

$$e^{-1}\partial_z\big[\epsilon(z)\partial_z\tilde{V}(z)\big] = e\Big[n_D(z) - \sum_i n_i^s |\psi_i(z)|^2\Big]. \quad (20)$$

From a quantum mechanical point of view, this corresponds to a mean-field treatment of the electron-electron interaction referred to as Hartree approximation, representing the lowest order of a perturbation expansion in the electron-electron interaction potential. In Eq. (20), $\epsilon(z)$ is the permittivity which varies with semiconductor composition and thus is also periodic, and e denotes the elementary charge. The right hand side of Eq. (20) corresponds to the space charge ρ in the QCL heterostructure due to the positively charged donors with concentration $n_D(z)$ and the electrons, where n_i^s is the electron sheet density of level i with wave function $\psi_i(z)$. This charge distribution in the structure gives rise to space charge effects, resulting in an additional electrostatic potential energy $\tilde{V}(z)$ which causes conduction band bending.[67]

The total potential V in Eq. (4) is then given by $V = V_0 + \tilde{V}$. Here, $V_0 = V_c - E_p z/L_p$, where V_c is the unbiased conduction band profile due to the varying material composition, thus describing the wells and barriers, and the term $-E_p z/L_p$ results from the applied bias. Since the energy drop across a period is given by the external bias, $V(z_0) - V(z_0 + L_p) = E_p$, we have $\tilde{V}(z_0) = \tilde{V}(z_0 + L_p)$. Due to the charge neutrality in each period, $\int_{z_0}^{z_0+L_p} \rho dz = 0$, we furthermore obtain $\partial_z\tilde{V}(z_0) = \partial_z\tilde{V}(z_0 + L_p)$, i.e., $\tilde{V}$ has the periodicity of V_c. Thus, we can restrict the solution of Eq. (20) to a single QCL period $z\epsilon[z_0, z_0 + L_p]$ and assume the boundary conditions $\tilde{V}(z_0) = \tilde{V}(z_0 + L_p) = 0$.

Equation (20) can, for example, be solved by applying the finite difference method. In analogy to Eqs. (18) and (19), we obtain

$$\tilde{s}_n\tilde{V}_{n-1} - \tilde{d}_n\tilde{V}_n + \tilde{s}_{n+1}\tilde{V}_{n+1} = \rho_n, \quad (21)$$

with

$$\rho_n = e\Big[n_{D,n} - \sum_i n_i^s |\psi_{i,n}|^2\Big] \quad (22)$$

and

$$\tilde{s}_n = \frac{1}{2e\Delta_z^2}(\epsilon_{n-1} + \epsilon_n),$$
$$\tilde{d}_n = \frac{1}{2e\Delta_z^2}(\epsilon_{n-1} + 2\epsilon_n + \epsilon_{n+1}). \quad (23)$$

Equation (21) is then solved over a single QCL period $z\epsilon[z_0, z_0 + L_p]$, with the grid points z_n, $n = 0...P$ which should coincide with the grid used for solving the Schrödinger equation Eq. (4). Applying the boundary conditions $\tilde{V}_0 = \tilde{V}_P = 0$, Eq. (21) can be written as a matrix equation $\mathbf{M}\tilde{\mathbf{V}} = \rho$, where $\tilde{\mathbf{V}}$ and ρ represent vectors with elements $\tilde{V}_n$ and ρ_n, respectively, with $n = 1...(P - 1)$. $\mathbf{M}$ is a matrix with the non-zero elements $M_{n,n} = -\tilde{d}_n$, $M_{n,n-1} = \tilde{s}_n$, $M_{n,n+1} = \tilde{s}_{n+1}$. This equation can be efficiently solved using an algorithm for tridiagonal equation systems.[64]

While n_i^s in Eq. (22) can in principle only be determined by detailed carrier transport simulations, simpler and much faster approaches are often adopted, e.g., for design optimizations of experimental QCL structures over an extended parameter range. Frequently, Fermi-Dirac statistics is applied,[57,66] with

$$n_i^s = \frac{m_i^{\parallel}}{\pi\hbar^2} k_B T_L \ln\big\{1 + \exp\big[(\mu - \tilde{E}_i)/(k_B T_L)\big]\big\}. \quad (24)$$

Here, we assume that the electron distribution is described by the lattice temperature T_L. Furthermore, μ is the chemical potential, k_B denotes the Boltzmann constant, and $m_i^{\parallel}$ is the effective mass associated with the ith subband, see Eq. (9). If nonparabolicity effects are neglected, m_i^* can often be approximated by the value of the well material. In Eq. (24), the energy $\tilde{E}_i$ defined in Eq. (6) is used instead of the eigenenergy E_i, to correctly reflect the invariance properties of the biased structure.[43] Especially, this guarantees that the obtained results do not depend on the choice of the elementary period in the heterostructure. The chemical potential μ is found from the charge neutrality condition within one period, i.e.,

$$n^s = \int_{z_0}^{z_0+L_p} n_D dz = \sum_i n_i^s. \quad (25)$$

This is done recursively by first determining a lower and an upper boundary value for μ where $\sum_i n_i^s < n^s$ and $\sum_i n_i^s > n^s$, respectively, and then finding the exact μ, e.g., by using the bisection method.[64]

The total potential in the Schrödinger equation (4) is given by $V = V_0 + \tilde{V}$, where $\tilde{V}$ has to be obtained by solving the Poisson equation (20). On the other hand, the wave functions ψ_i in Eq. (20) must be determined from Eq. (4). In practice, this is done by iteratively solving the Schrödinger and Poisson equations,[57] initially assuming $\tilde{V} = 0$, until the results for $\tilde{V}$, ψ_i, and E_i converge. If the n_i^s are obtained from self-consistent carrier transport simulations instead of using Eq. (24), then the carrier transport simulation and the numerical solution of the Schrödinger-Poisson system have to be performed iteratively until convergence is obtained. Such simulations are then referred to as self-self-consistent approaches.[67,68]
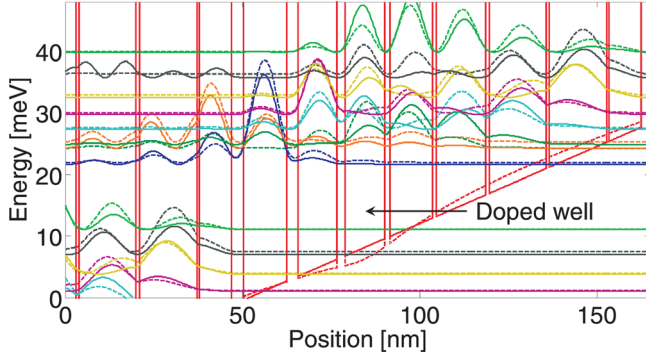
FIG. 6. Conduction band profile and probability densities for a bound-to-continuum terahertz QCL.[69] Shown are the results without considering space charge effects (solid lines) and for space charge effects taken into account, assuming thermally occupied subbands (dashed lines).

FIG. 7. Typical waveguide resonator geometry of a QCL.

In Fig. 6, simulation results for a terahertz QCL[69] obtained by solving the Schrödinger equation (solid lines) and the Schrödinger-Poisson system (dashed lines) are compared. Although the populations obtained by Eq. (24) usually deviate considerably from detailed carrier transport simulations, this simplified approach can already give improved results, as compared to completely neglecting space charge effects.[68] The reason is that the positively charged donors are usually localized in a relatively small section, e.g., a single quantum well, while the electrons are more or less distributed across the whole QCL period. This leads to considerable conduction band bending due to space charge effects, as illustrated in Fig. 6. While a more exact determination of n_i^s yields a somewhat different electron distribution, the overall result for $\tilde{V}$ will be similar as for the simplified model based on Eq. (24).

The method shown above is not the only one for solving the Schrödinger-Poisson system self-consistently under the condition of global charge neutrality. A common alternative approach is to keep the chemical potential fixed and solve the Poisson equation with Neumann boundary conditions. These boundary conditions allow the Poisson potential to self-adjust the density and maintain the global charge neutrality. Typically, this requires a nonlinear realization of the Poisson equation to include an explicit potential dependence of the charge density. Such a nonlinear Poisson equation can then effectively be solved with iterative methods such as the predictor-corrector approach.[70]

III. OPTICAL RESONATOR MODELING

The resonator spatially confines the radiation field, furthermore providing optical outcoupling, beam shaping and frequency selection. While most theoretical work has focused on the carrier transport in the gain medium, there has also been progress in the modeling of the cavity. For example, the resonator loss, which is crucial for the temperature performance of THz QCLs,[17,19] has been extracted from finite element simulations of the resonator.[71] In Fig. 7, a typical waveguide resonator geometry of a QCL is sketched, and the used coordinate system is shown for reference. The propagation direction of the optical field is denoted by x, y refers
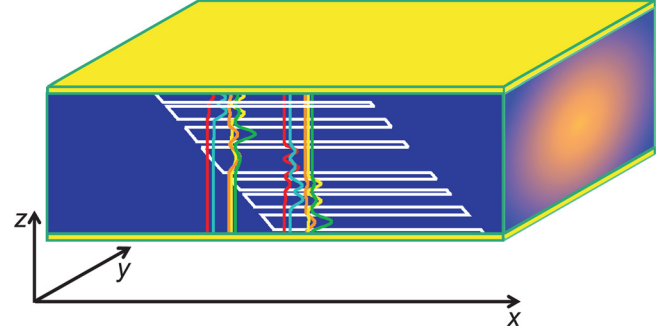
to the lateral direction, and z indicates the growth direction of the heterostructure. The resonator consists of materials with different permittivities to obtain waveguiding and optical outcoupling. Thus, for optical cavity simulations, the material permittivities must be known, which in general depend on the frequency, doping level, and temperature. In this context, often the Drude model is employed with adequately chosen fitting parameters.[71] For intersubband optical transitions, only the dipole matrix element in the z direction where quantum confinement occurs is nonzero, see Sec. V B. Thus, only resonator modes with an electric field component along the z direction are amplified. For this reason, surface emission, i.e., outcoupling through the xy-plane, can only be obtained by using special outcoupling schemes.[33,72]

A. Maxwell's equations

The propagation of the electric and magnetic field vectors $\mathbf{E}(\mathbf{x}, t)$ and $\mathbf{H}(\mathbf{x}, t)$ is generally described by Maxwell's equations. Here, $\mathbf{x} = [x, y, z]^T$ and t are the position vector and time, respectively. In the following, monochromatic fields are considered. We use the physics convention, i.e., we assume a harmonic time dependence $\propto \exp(-i\omega t)$, where ω denotes the angular frequency. The fields are then expressed as $\mathbf{H}(\mathbf{x}, t) = \Re\{\hat{\mathbf{H}}(\mathbf{x})\exp(-i\omega t)\}$, $\mathbf{E}(\mathbf{x}, t) = \Re\{\hat{\mathbf{E}}(\mathbf{x})\exp(-i\omega t)\}$, where $\hat{\mathbf{H}}$ and $\hat{\mathbf{E}}$ denote the complex amplitude vectors. All equations can easily be converted to the engineering convention [expressing the time dependence as $\propto \exp(j\omega t)$] by making the formal substitution $i \rightarrow -j$. For optical frequencies, typically the permeability is given by its vacuum value μ_0. Furthermore, it is here assumed that the dielectric constant can be described by a position dependent complex scalar $\epsilon_r(\mathbf{x})$. Under these conditions, Maxwell's equations simplify to

$$\nabla \times \hat{\mathbf{H}} = -i\omega\epsilon_0\epsilon_r\hat{\mathbf{E}}, \tag{26a}$$

$$\nabla \times \hat{\mathbf{E}} = i\omega\mu_0\hat{\mathbf{H}}. \tag{26b}$$

$\nabla\mathbf{B} = \mu_0\nabla(\mathbf{H}) = 0$ is then automatically fulfilled, as can be seen by taking the divergence of Eq. (26b). The equation $\nabla\mathbf{D} = \nabla(\epsilon_0\epsilon_r\mathbf{E}) = \rho$ delivers the charge density ρ and is not needed to compute $\mathbf{E}$ and $\mathbf{H}$. Eqs. (26) and the corresponding boundary conditions define an eigenvalue problem, which yields the electromagnetic resonator modes. General numerical approaches for solving Maxwell's equations include the finite element[73] and finite difference time domain[35] method,

which also have been applied to the simulation of QCL cavities.[71,74]

B. Two-dimensional waveguide model

By eliminating $\hat{\mathbf{E}}$ from Eqs. (26), a wave equation for $\hat{\mathbf{H}}$ can be derived.[75] In many cases, a waveguide geometry is used which does not depend on the longitudinal x direction, i.e., $\epsilon_r = \epsilon_r(y, z)$. Optical outcoupling is then obtained through the cleaved semiconductor facets which serve as partly transparent mirrors. Since the resonator length of up to a few mm is large as compared to the transverse resonator dimensions, the computation of the transverse mode profile in the yz-plane can be decoupled from the propagation coordinate and reduces to a 2D problem. The facet transmittance is then calculated based on the obtained transverse field distribution. For a constant waveguide geometry in propagation direction x, we can assume a field dependence $H_{y,z}(\mathbf{x}, t) = \Re\{\hat{H}_{y,z}(y, z)\exp(i\beta x - i\omega t)\}$ with the complex propagation constant β, and analogously for the electric field. The wave equation then reduces to two coupled differential equations for the transverse field components

$$(\partial_p^2 + \partial_q^2)\hat{H}_p + \frac{\partial_q \epsilon_r}{\epsilon_r}(\partial_p \hat{H}_q - \partial_q \hat{H}_p) = (\beta^2 - k_0^2 \epsilon_r)\hat{H}_p, \quad (27)$$

with $k_0 = \omega\sqrt{\mu_0 \epsilon_0} = \omega/c$, $p = y, z$, and $q = z, y$.[75] The longitudinal component H_x is obtained from $\nabla\mathbf{H} = 0$. Furthermore, the electric field components can be calculated from Eq. (26a). The solution of Eq. (27), along with the corresponding boundary condition of vanishing fields for $y^2 + z^2 \to \infty$, constitutes an eigenvalue problem, and the corresponding solutions $\hat{H}_{y,z}$ and β correspond to the waveguide modes. As mentioned above, only resonator modes with an electric field component along the z direction are amplified. This is fulfilled for the transverse magnetic (TM) modes, which are characterized by a vanishing magnetic field in the propagation direction, i.e., $H_x = 0$.

A waveguide mode is frequently characterized using three parameters, the overlap (or field confinement) factor Γ, the waveguide loss coefficient a_w, and the mirror or outcoupling loss coefficient a_m, with the total power loss coefficient $a = a_w + a_m$. Based on these quantities, the threshold gain is given by $g_{th} = a/\Gamma$. Furthermore, for QCL simulations including the optical cavity field, these parameters enter the simulation of the QCL gain medium to describe the properties of the waveguide mode.[76,77] Γ, a_w, and a_m can be obtained from the mode solutions of Eq. (27). The waveguide loss arises from the absorption in the waveguide layers and is given by $a_w = 2\Im\{\beta\}$. The overlap factor corrects for the fact that the mode only partially overlaps with the gain medium. To reflect the fact that the gain medium only couples to the E_z component, the overlap factor is defined as[78]

$$\Gamma = \frac{\iint_{S_g} |\hat{E}_z|^2 \mathrm{d}y\mathrm{d}z}{\iint_{-\infty}^{\infty} |\hat{\mathbf{E}}|^2 \mathrm{d}y\mathrm{d}z}, \quad (28)$$

where we integrate over the gain medium cross section area S_g in the enumerator.

Strictly speaking, the calculation of the facet transmission constitutes a full 3D problem, since the facets introduce an abrupt change in x direction. However, since the QCL resonator length is large as compared to its transverse dimensions, the computation of the transverse mode profile in the yz-plane can be decoupled from the x coordinate, as mentioned above. Only for sufficiently wide transverse waveguide dimensions, the facet reflectance R can be estimated from Fresnel's formula

$$R = |n_{\mathrm{eff}} - 1|^2 / |n_{\mathrm{eff}} + 1|^2, \quad (29)$$

with the effective refractive index defined as $n_{\mathrm{eff}} = \beta/k_0$. In general, modal effects lead to an increased reflectance.[71] Various methods have been developed to extract R from the mode solutions provided by Eq. (27).[79–82] The description of the outcoupling loss by a distributed coefficient a_m is obtained from $R = \exp(-a_m L)$ with the resonator length L, yielding

$$a_m = -\ln(R)/L. \quad (30)$$

If one facet is reflection coated and the light is outcoupled only at one side, we obtain $a_m = -\ln(R)/(2L)$.

C. One-dimensional waveguide slab model

If the waveguide width in lateral y direction significantly exceeds its thickness, the waveguide calculations can be reduced to a 1D problem with $\epsilon_r = \epsilon_r(z)$, corresponding to the simulation of a slab waveguide structure.[83] This applies for example to typical THz metal-metal waveguide resonators, where the vertical z dimension of around $10\,\mu m$ is often significantly smaller than the lateral y dimension $(\approx 25 - 200\,\mu m)$.[71] For TM modes, in addition to $H_x = 0$ we have approximately $H_z \approx 0$, and the y component is given by $H_y(x, z) = \Re\{\hat{H}_y(z)\exp(i\beta x - i\omega t)\}$, where β denotes the complex propagation constant. The one-dimensional wave equation is then obtained from Eqs. (26) or Eq. (27) as

$$\epsilon_r \partial_z(\epsilon_r^{-1}\partial_z \hat{H}_y) = (\beta^2 - \epsilon_r k_0^2)\hat{H}_y. \quad (31)$$

The electric field amplitude is with Eq. (26a) given by

$$\hat{E}_z = -\frac{\beta}{\omega \epsilon_0 \epsilon_r}\hat{H}_y, \quad (32a)$$

$$\hat{E}_x = -\frac{i}{\omega \epsilon_0 \epsilon_r}\partial_z \hat{H}_y. \quad (32b)$$

From Eq. (31) we see that both H_y and $(1/\epsilon_r)\partial_z H_y$ must be continuous, corresponding to the continuity of the field components H_y and, with Eq. (32b), E_x parallel to the layers. Equation (31) can be brought into the same form as the one-dimensional Schrödinger equation (4) by associating $\hat{H}_y(z)$, $\epsilon_r(z)$, and $\epsilon_r^{-1}\beta^2 - k_0^2$ with $\psi(z)$, $-2m^*(z)/\hbar^2$, and $E - V(z)$, respectively. Thus, we can use the transfer matrix method introduced in Sec. II D 1 also for solving Eq. (31).

We start by dividing the waveguide in z direction into layers n, $z_n \leq z < z_n + \Delta_n = z_{n+1}$, with constant relative permittivities $\epsilon_r^{(n)}$. Although the gain medium itself consists of different layers, it can be described by its total layer thickness and a single effective dielectric constant since the individual layers are so thin that the electromagnetic wave cannot resolve the structure. For TM modes, the inverse effective dielectric constant is given by $\epsilon_{r,eff}^{-1} = (\Delta_b \epsilon_{r,b}^{-1} + \Delta_w \epsilon_{r,w}^{-1})/(\Delta_b + \Delta_w)$, where Δ_b and Δ_w denote the total thickness of the barriers and wells in the gain medium, respectively, and $\epsilon_{r,b}$, $\epsilon_{r,w}$ are the corresponding dielectric constants.[84,85] In analogy to Eq. (10) for solving the Schrödinger equation, we can write the solution of Eq. (31) in layer n as[83]

$$\hat{H}_y^{(n)} = A_n \exp\left[ik_n(z - z_n)\right] + B_n \exp\left[-ik_n(z - z_n)\right], \quad (33)$$

with $k_n = (\epsilon_r^{(n)} k_0^2 - \beta^2)^{1/2}$. The propagation through the segment n is then described by the matrix $T_n(\Delta_n)$ defined in Eq. (14). As mentioned above, Eq. (31) implies the continuity of H_y and $(1/\epsilon_r)\partial_z H_y$ across layer boundaries. These matching conditions between two layers can be expressed in terms of matrix Eq. (15) by choosing $\tilde{k}_n = k_n/\epsilon_r^{(n)}$. There is, however, one fundamental difference between Eqs. (4) and (31): Equation (31) generally has complex eigenvalues β^2 since ϵ_r becomes complex for materials with optical loss or gain. Consequently, the shooting method described in Sec. II D 1 is not applicable, and a complex root finding algorithm has to be used.

The transmittance $T = 1 - R$ through the facet can be approximately computed from the mode profile $\hat{H}_y(z)$. For TM polarization, we obtain from the boundary value method[79]

$$T = \left[\int_{-\infty}^{\infty} \frac{\Re\{\beta\}\Re\{\epsilon_r(z)\} + \Im\{\beta\}\Im\{\epsilon_r(z)\}}{|\epsilon_r(z)|^2} |\hat{H}_y(z)|^2 dz\right]^{-1}$$
$$\times 8\pi|\beta|^2 k_0 \int_{-k_0}^{k_0} \frac{\sqrt{1 - k_z^2/k_0^2}|\Phi(k_z)|^2|\Phi'(k_z)|^2}{|k_0\sqrt{1 - k_z^2/k_0^2}\Phi(k_z) + \beta\Phi'(k_z)|^2} dk_z, \quad (34)$$

with the Fourier transforms

$$\Phi(k_z) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{H}_y(z)\exp(-ik_z z)dz,$$
$$\Phi'(k_z) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{\hat{H}_y(z)}{\epsilon_r(z)}\exp(-ik_z z)dz. \quad (35)$$

Several extensions of the one-dimensional mode calculations presented above are available. For high aspect ratios of the waveguide cross section, the 2D mode calculation can be approximately reduced to 1D problems in y and z directions, respectively, by using the effective refractive index method.[86] A generalization of the transfer matrix scheme for solving the two-dimensional wave equation, Eq. (27), is the film mode matching method, which is especially efficient for waveguides with a rectangular geometry.[87] Waveguides with a periodic structure in propagation direction, such as a grating for surface emission, can be treated with coupled mode theory.[34]

IV. OVERVIEW AND CLASSIFICATION OF CARRIER TRANSPORT MODELS

Depending on the intended goals, carrier transport models with varying degrees of complexity have been used for the QCL gain medium simulation, ranging from simple rate equation approaches to fully quantum mechanical descriptions. The central task is to determine the optical gain, which is proportional to the population inversion between the upper and the lower laser levels, and the current through the heterostructure as a function of the applied bias voltage. Most simulation approaches require the eigenenergies and wave functions of the energy levels in the QCL heterostructure as an input, which are determined by solving the Schrödinger equation or the Schrödinger-Poisson equation system (see Sec. II).

In Fig. 8, various theoretical descriptions of the gain medium with different levels of complexity are illustrated. The most basic model of laser gain media in general is the rate equation approach,[88] which is also frequently applied to QCLs.[77,89–94] The electron transport is simply described by transition rates between the relevant energy levels in the gain medium. In the case of the QCL, these are the electron subbands, and the corresponding transitions are referred to as
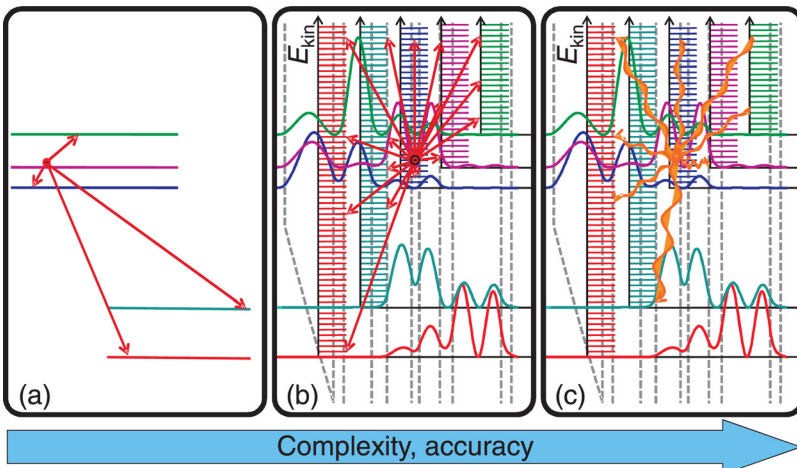

FIG. 8. Illustration of various carrier transport models with different levels of complexity and accuracy: (a) Rate equation approach; (b) ensemble Monte Carlo method; (c) 3D density matrix/NEGF approach.

intersubband transitions. The application of the rate equation model to the QCL is illustrated in Fig. 8(a). In addition to the transitions induced by the optical lasing field, nonradiative transitions occur due to various scattering mechanisms, such as the interaction of the electrons with phonons, other electrons or defects in the semiconductor lattice. In the simplest case, these scattering rates are experimental or empirical input parameters.[89,90] In more advanced implementations, the transition rates are computed based on the corresponding Hamiltonian,[91,92] only relying on well known material parameters such as the effective mass. Thus, no specific experimental input is required, and no free "fitting" parameters are available. A direct numerical solution is here not possible since the transition rates depend on the *a priori* unknown electron populations. Rather, the steady state solution must be found by simulating the temporal evolution of the system until convergence is reached, or by using an iterative scheme. Such advanced carrier transport methods are generally referred to as self-consistent approaches. Apart from providing an intuitive description, rate equation models are numerically very efficient, and thus are frequently employed for the design and optimization of experimental QCL structures.[94] An extension are the Maxwell-Bloch equations, which include the carrier-light interaction based on the density matrix formalism.[95] This approach is used to investigate coherence effects between the laser field and the gain medium, relevant in particular for the pulse formation in mode locked QCLs.[96–98]

In a quantum well heterostructure such as the QCL gain medium, the electrons are only confined in the growth direction. Thus, energy quantization occurs in one dimension, while the electrons can still freely move in two dimensions. Due to this free in-plane motion in the quantum well, the electrons have kinetic energy in addition to the eigenenergy of the corresponding quantized level, as illustrated in Fig. 8(b). For a more detailed modeling and an improved understanding of the carrier transport in QCLs, the in-plane motion of the electrons must be taken into account. Such three-dimensional (3D) simulation approaches also consider intrasubband transitions occurring between different kinetic energies within a subband, and yield the electron distribution in each subband in addition to the level populations. On the other hand, the complexity increases significantly as compared to one-dimensional (1D) descriptions such as above discussed rate equation approach, since transitions are now characterized by the initial and final subbands as well as the corresponding kinetic energies. An example for a self-consistent 3D approach is the EMC method. Here, the intrasubband processes are fully taken into account, and the scattering rates are self-consistently evaluated based on the corresponding Hamiltonian. Combining versatility and reliability with relative computational efficiency, the EMC approach has been widely used for the analysis, design, and optimization of QCLs.[46,47,99–110]

In both the rate equation and EMC approach, the carrier transport is described by scattering-induced transitions of discrete electrons between the quantized energy levels, corresponding to a hopping transport model. Such methods are referred to as semiclassical, since the energy quantization in

the QCL heterostructure is considered, but quantum coherence effects and quantum mechanical dephasing are not included. Various quantum transport simulation approaches have been developed which take into account quantum correlations between the energy levels. One example is the density matrix method. Its 1D version can be seen as a generalization of the rate equation approach, and is frequently used for the analysis and optimization of THz QCLs.[17,111–113] In its 3D form[114,115] illustrated in Fig. 8(c), it corresponds to a generalization of semiclassical approaches based on the Boltzmann transport equation such as EMC.[114] Also the nonequilibrium Green's function (NEGF) method, considered the most general quantum transport approach, has been applied to the simulation of QCL structures.[116–124] Quantum transport approaches are numerically much more demanding than their semiclassical counterparts. On the other hand, quantum coherence effects can play a pronounced role especially in THz QCLs where the energetic spacing between the quantized levels is relative small,[115,125] while they are less relevant in mid-infrared QCLs.[114]

In Table I, a classification of the different simulation approaches covered in this review is given. Here, we differentiate between semiclassical schemes based on hopping transport between the quantized energy levels, and quantum transport approaches taking into account quantum correlations. We divide the methods into empirical approaches relying on empirical or experimental input parameters, and self-consistent schemes which evaluate the transition rates based on the corresponding Hamiltonian. Furthermore, we distinguish between 1D modeling techniques only considering the subband populations and intersubband transitions, and 3D approaches also taking into account the electron distribution in the subbands and the intrasubband dynamics.

V. EMPIRICAL APPROACHES

The most basic approach for modeling the electron dynamics in a laser is to use experimental or empirical transition rates between the relevant energy levels in the laser gain medium.[88] For QCLs, these levels correspond to the quantized eigenstates of the heterostructure, which have to be found by solving the Schrödinger equation, Eq. (4). The electron dynamics is then described by rate equations.[88] Often, only the nonradiative transitions are considered which occur due to various scattering mechanisms, such as the interaction

TABLE I. Classification of carrier transport modeling techniques. The corresponding section number is given in brackets.

	Semiclassical	Quantum transport
	Empirical	
1D	Rate equations (V A)	1D density matrix (VII)
	Maxwell-Bloch[a] (V B)	
	Self-consistent	
1D	Rate equations (VI E)	1D density matrix (VII)
3D	Monte Carlo (VI F)	3D density matrix (VII)
		NEGF (VIII)

[a]Only the carrier-light interaction is modeled using a density matrix formalism, while scattering is treated based on rate equations.

of the electrons with phonons, other electrons or defects in the semiconductor lattice. Simulations not including the optical cavity field can be used to investigate under which conditions sufficient optical gain is obtained, so that lasing operation can start at all. In this way, parameters such as the threshold current density and maximum operating temperature can be extracted. To investigate the lasing operation itself, including optical output powers, saturation effects or the intrinsic linewidth,[90] the optical field has to be included as well. Modified empirical scattering-rate approaches have been developed for specific types of QCLs.[126] Maxwell-Bloch equations which are a generalization of the rate equation approach can be used to model the coherent interaction between the laser field and the gain medium,[95] e.g., to investigate the formation of optical instabilities in mode-locked QCLs.[96]

A. Empirical rate equations

The rate equations for a laser are given by[88]

$$d_t n_i^s = \sum_{j \neq i} \tau_{ji}^{-1} n_j^s - \tau_i^{-1} n_i^s + \sum_{j \neq i} (-W_{ij}^{\text{opt}} n_i^s + W_{ji}^{\text{opt}} n_j^s). \quad (36)$$

The first two terms contain the relaxation transitions due to scattering, e.g., the interaction of the electrons with phonons, other electrons or defects in the semiconductor lattice. Also spontaneous emission can be included here. The scattering rate from a level j to i is often expressed in terms of an inverse scattering lifetime τ_{ji}^{-1}, where $\tau_i^{-1} = \sum_{j \neq i} \tau_{ij}^{-1}$ indicates the total inverse lifetime of level i. The last sum contains the lasing transitions, where W_{ij}^{opt} are the stimulated optical transition rates for those transitions where an optical field at or near the corresponding frequency $|\omega_{ij}| = |E_i - E_j|/\hbar$ is present.[88] The rates W_{ij}^{opt} are proportional to the optical intensities in the corresponding lasing modes. Typically, only one or a few transitions contribute to lasing. Furthermore, n_i^s is the electron sheet density of subband i, i.e., the electron number divided by the in-plane cross section area S. This quantity is often used in QCL heterostructures where energy quantization occurs in one dimension and the electrons can still freely move in in-plane direction.

Commonly, the QCL heterostructure is designed strictly periodically, as illustrated in Fig. 1. Apart from fabrication tolerances, a periodic model is valid for the central QCL periods far away from the contacts, if effects such as domain formation[127] and local variations of the optical field intensity can be neglected. The sum in Eq. (36) can then be restricted to one representative central period

$$d_t n_i^s = \sum_{j \neq i} \hat{\tau}_{ji}^{-1} n_j^s - \tau_i^{-1} n_i^s + \sum_{j \neq i} (-\hat{W}_{ij}^{\text{opt}} n_i^s + \hat{W}_{ji}^{\text{opt}} n_j^s), \quad (37)$$

with $i, j = 1..N$, where N is the number of subbands in each period. Here, $\hat{\tau}_{ji}^{-1} = \sum_{n \in \mathbb{Z}} \tau_{j,i+nN}^{-1}$ includes the transitions to all equivalent levels in the different periods, and analogously for $\hat{W}_{ji}^{\text{opt}}$. The total sheet density in each period is determined by the doping sheet density

$$n^s = \sum_{i=1}^{N} n_i^s. \quad (38)$$

The steady state solution is obtained by setting $d_t n_i^s = 0$. If we are not interested in the lasing operation itself, but only if sufficient inversion for lasing is obtained, stimulated optical effects can be excluded, $\hat{W}_{ij}^{\text{opt}} = \hat{W}_{ji}^{\text{opt}} = 0$. The subband populations n_i^s ($i = 1..N$) can then be found by solving the linear equation system Eq. (37) with $d_t n_i^s = 0$ and $i = 1..(N-1)$, complemented by Eq. (38) to obtain a linearly independent system and thus a unique solution. To include lasing, this system has to be complemented by equations describing the optical intensities in the lasing modes, since the $\hat{W}_{ij}^{\text{opt}}$ is intensity dependent.

Often the empirical rate equation is restricted to the most important subbands and transitions to obtain compact analytical results. A common model is the three-level system,[89] illustrated in Fig. 9. This description only includes the upper laser level 2, lower laser level 1, and a reservoir level 0 representing the extraction and injector subbands. A fraction η_2 of the current density J is injected into the upper laser level, and η_1 into the lower laser level. Furthermore, only the transition rates τ_{21}^{-1}, τ_{20}^{-1}, and τ_{10}^{-1} are included, with the inverse upper and lower laser level lifetimes $\tau_2^{-1} = \tau_{21}^{-1} + \tau_{20}^{-1}$, $\tau_1^{-1} = \tau_{10}^{-1}$. With Eq. (36), the rate equations for the upper and lower laser level are then obtained as

$$d_t n_2^s = \eta_2 J/e - \tau_2^{-1} n_2^s - \sigma_{21}^{\text{opt}} I (n_2^s - n_1^s),$$
$$d_t n_1^s = \eta_1 J/e - \tau_1^{-1} n_1^s + \tau_{21}^{-1} n_2^s + \sigma_{21}^{\text{opt}} I (n_2^s - n_1^s), \quad (39)$$

where σ_{21}^{opt} is the cross section for stimulated emission from level 2 to 1 and I is the optical intensity of the laser radiation at frequency $\omega_{21} = (E_2 - E_1)/\hbar$. If lasing is neglected, $I = 0$, the steady state population inversion is with $d_t = 0$ in Eq. (39) obtained as[89]

$$n_2^s - n_1^s = \frac{J}{e} \left[\eta_2 \tau_2 \left(1 - \frac{\tau_1}{\tau_{21}} \right) - \eta_1 \tau_1 \right]. \quad (40)$$

The three-level model can be extended to take into account further effects such as thermal backfilling of the lower laser level and backscattering from level 1 to level 2 in terahertz QCLs, which can be modeled by introducing additional lifetimes τ_{01} and τ_{12}, respectively.[90]

B. Maxwell-Bloch equations

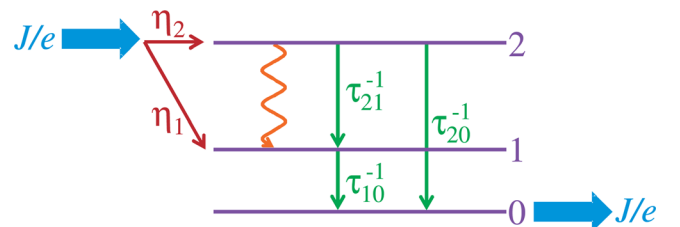Bloch equations are a generalization to the rate equation approach where the interaction of the laser level electrons

FIG. 9. Three-level system model for a QCL.

with the optical field is modeled using a density matrix formalism rather than scattering rates.[95] In this way, optical nonlinearities and coherence effects between the laser field and the gain medium can be considered, while the carrier transport due to nonradiative mechanisms is included by the corresponding lifetimes as for the rate equations. To describe the optical propagation, the Bloch equations are complemented by Maxwell's or related equations, such as the wave equation. This model has, for example, been used to investigate the formation of optical instabilities in QCLs,[96,97] and to study the possibility of self-induced transparency modelocking.[98] The nonlinear optical dynamics in the gain medium also plays an important role for the recently demonstrated QCL-based frequency combs.[128]

The interaction of a classical optical field with a two-level system is in the density matrix formalism described by[95]

$$
\begin{aligned}
\partial_t \rho_{21} &= -i\omega_{21}\rho_{21} - i\hbar^{-1}d_{12}^* E_z \Delta_{21} - \gamma_{21}\rho_{21}, \\
\partial_t \Delta_{21} &= 2i\hbar^{-1}\left(d_{12}^*\rho_{21}^* - d_{12}\rho_{21}\right)E_z - \gamma_E\left(\Delta_{21} - \Delta_{21}^{\mathrm{eq}}\right),
\end{aligned}
\tag{41}
$$

where the asterisk denotes the complex conjugate. $E_z(x,t)$ represents the field component in the growth direction z of the heterostructure since the other components do not interact with the gain medium, and $d_{12} = -e\langle 1|z|2\rangle$ is the corresponding dipole matrix element of the laser transition. Furthermore, $\rho_{ij}(x,t) = \rho_{ji}^*(x,t) = \langle i|\hat{\rho}(x,t)|j\rangle$ are the density matrix elements, and $\Delta_{21}(x,t) = \rho_{22}(x,t) - \rho_{11}(x,t)$ is the inversion. The density matrix can be normalized so that $\rho_{ii} = n_i^{\mathrm{s}}/n^{\mathrm{s}}$ gives the relative population of subband i. Dissipative processes are phenomenologically included by adding decay terms with relaxation rates γ_E and γ_{21}, describing the energy relaxation and dephasing, respectively. $\Delta_{21}^{\mathrm{eq}}$ is the equilibrium inversion which the system approaches for $E_z = 0$. If the gain medium is not homogeneous along the propagation direction, as is the case for self-induced transparency mode locking structures,[98] the parameters d_{12}, γ_E, γ_{21}, ω_{21}, and $\Delta_{21}^{\mathrm{eq}}$ depend on the propagation coordinate x.

Assuming weak nonlinearity and inhomogeneity, the optical field propagation can be described by the wave equation.[95] For propagation in x direction, it is given by

$$
\left(\partial_x^2 - c^{-2}n_0^2\partial_t^2\right)E_z = \epsilon_0^{-1}c^{-2}\partial_t^2 P_z.
\tag{42}
$$

$P_z(x,t)$ is the polarization component in z direction due to the lasing transition, given by $P_z = (n^{\mathrm{s}}/L_{\mathrm{p}})(d_{12}\rho_{21} + d_{12}^*\rho_{21}^*)$. Here, L_{p} is the length of a single QCL period and $n^{\mathrm{s}}/L_{\mathrm{p}}$ thus corresponds to the average electron concentration.[97] Furthermore, n_0 denotes the refractive index of the gain medium material.

E_z and ρ_{21} are typically expressed by their slowly varying envelope functions

$$
\begin{aligned}
E_z(x,t) &= \frac{1}{2}\hat{E}_z(x,t)\exp[i(k_{21}x - \omega_{21}t)] + c.c., \\
\rho_{21}(x,t) &= \eta_{21}(x,t)\exp[i(k_{21}x - \omega_{21}t)],
\end{aligned}
\tag{43}
$$

where $k_{21} = \omega_{21}n_0/c$ and c.c. denotes the complex conjugate. Inserting Eq. (43) into Eq. (41) and neglecting the

rapidly oscillating terms $\propto \exp(\pm 2i\omega_{21}t)$, we obtain the density matrix equations in the rotating wave approximation[95]

$$
\partial_t \eta_{21} = -i(2\hbar)^{-1}d_{12}^*\hat{E}_z\Delta_{21} - \gamma_{21}\eta_{21},
\tag{44a}
$$

$$
\partial_t \Delta_{21} = i\hbar^{-1}\left(\eta_{21}^* d_{12}^*\hat{E}_z - \eta_{21}d_{12}\hat{E}_z^*\right) - \gamma_E\left(\Delta_{21} - \Delta_{21}^{\mathrm{eq}}\right).
\tag{44b}
$$

Inserting Eq. (43) into Eq. (42) yields with the slowly varying amplitude approximation $|\partial_x^2\hat{E}_z| \ll |k_{21}\partial_x\hat{E}_z|$, $|\partial_t^2\hat{E}_z| \ll |\omega_{21}\partial_t\hat{E}_z|$[95,97]

$$
\partial_x\hat{E}_z + n_0 c^{-1}\partial_t\hat{E}_z = i\frac{\omega_{21}\Gamma}{\epsilon_0 c n_0}\frac{n^{\mathrm{s}}}{L_{\mathrm{p}}}d_{12}\eta_{21} - \frac{1}{2}a\hat{E}_z,
\tag{45}
$$

furthermore assuming $|\partial_t^2\eta_{21}|, |\omega_{21}\partial_t\eta_{21}| \ll |\omega_{21}^2\eta_{21}|$. Here, the overlap factor Γ, Eq. (28), has been added to correct for the fact that the optical mode only partially overlaps with the gain medium, and a loss term with the power loss coefficient a has been included. For applying the Maxwell-Bloch equation model to QCLs, typically the simplifying assumption is made that the lower laser level is depopulated very efficiently, $\tau_1 \to 0$, resulting in $n_1^{\mathrm{s}} = 0$ in Eq. (39).[97] The inversion is then directly given by the upper laser level population, $\Delta_{21} = n_2^{\mathrm{s}}$. By comparison of Eq. (44) with Eq. (39), we obtain $\Delta_{21}^{\mathrm{eq}} = \tau_2\eta_2 J/(n^{\mathrm{s}}e)$, $\gamma_E = \tau_2^{-1}$.

The Maxwell-Bloch equations, Eqs. (44) and (45), are a versatile approach to describe the optical dynamics in QCLs, and additional effects can straightforwardly be implemented. For example, the influence of spatial hole burning due to the standing wave modes in a linear cavity has been extensively studied,[96,97,129] and dispersion as well as saturable absorption have been added to the model.[130]

1. Optical gain and transition rates

The Bloch equations Eq. (44) can be used to derive the optical gain coefficient and transition rate associated with stimulated emission and absorption.[95] For a monochromatic electromagnetic field, the stationary solution of Eq. (44a) is obtained by setting $\partial_t\eta_{21} = 0$

$$
\eta_{21} = -\frac{i}{2\hbar\gamma_{21}}d_{12}^*\hat{E}_z\Delta_{21}.
\tag{46}
$$

Multiplying Eq. (45) from left with $\hat{E}_z^*$ and adding the complex conjugate, we obtain with Eq. (46)

$$
\partial_x I + n_0 c^{-1}\partial_t I = \Gamma g I - aI,
\tag{47}
$$

with the power gain coefficient

$$
g = \frac{\omega_{21}}{\hbar\gamma_{21}\epsilon_0 c n_0}\frac{n^{\mathrm{s}}}{L_{\mathrm{p}}}|d_{12}|^2(\rho_{22} - \rho_{11}).
\tag{48}
$$

Here, we have replaced the electric field by the optical intensity $I = \epsilon_0 c n_0 |\hat{E}_z|^2/2$. The optical power inside the resonator is given by $P = IS_{\mathrm{g}}/\Gamma$, where S_{g} is the cross section area of the gain medium and S_{g}/Γ corresponds to the effective area of the waveguide mode. Frequently, I is assumed to change

only slightly along the resonator, i.e., the intensity is averaged over the x coordinate. This assumption is valid especially for the case of moderate output coupling at the facets where the mirror loss can be described by a distributed coefficient a_m, Eq. (30). Equation (47) then simplifies to

$$n_0 c^{-1} \partial_t I = \Gamma g I - a I. \quad (49)$$

The transition rate due to the optical field is with $\partial_t \Delta_{21}|_\text{opt} = 2\partial_t \rho_{22}|_\text{opt} = -2\partial_t \rho_{11}|_\text{opt}$ obtained from Eqs. (44b) and (46) as

$$\partial_t \rho_{22}|_\text{opt} = \frac{1}{2} i\hbar^{-1} (\eta_{21}^* d_{12}^* \hat{E}_z - \eta_{21} d_{12} \hat{E}_z^*)$$
$$= -\frac{1}{\hbar^2 \gamma_{21} \epsilon_0 c n_0} |d_{12}|^2 (\rho_{22} - \rho_{11}) I. \quad (50)$$

The contribution $\propto \rho_{22} I$ leading to a reduction of ρ_{22} is due to stimulated emission, while the contribution $\propto \rho_{11} I$ corresponds to absorption. For a slightly detuned optical field at a frequency $\omega \neq \omega_{21}$, Eqs. (48) and (50) can be adapted by replacing γ_{21}^{-1} with $\pi \mathcal{L}(\omega)$, where $\mathcal{L}(\omega)$ is the Lorentzian lineshape function[95]

$$\mathcal{L}(\omega) = \frac{1}{\pi} \frac{\gamma_{21}}{\gamma_{21}^2 + (\omega - |\omega_{21}|)^2}. \quad (51)$$

Thus, we obtain with the sheet densities $n_1^s = n^s \rho_{11}$, $n_2^s = n^s \rho_{22}$,

$$g = \frac{\pi \omega}{\hbar \epsilon_0 c n_0 L_p} |d_{12}|^2 (n_2^s - n_1^s) \mathcal{L}(\omega), \quad (52)$$

and

$$\partial_t n_2^s|_\text{opt} = -\frac{\pi}{\epsilon_0 c n_0 \hbar^2} |d_{12}|^2 (n_2^s - n_1^s) I \mathcal{L}(\omega). \quad (53)$$

Comparison with Eq. (36) yields for the stimulated optical transition rates

$$W_{ij}^\text{opt} = W_{ji}^\text{opt} = \frac{\pi}{\epsilon_0 c n_0 \hbar^2} |d_{12}|^2 I \mathcal{L}(\omega). \quad (54)$$

VI. SELF-CONSISTENT SEMICLASSICAL APPROACHES

Advanced self-consistent simulation approaches only rely on well known material parameters such as the effective mass, and no further specific experimental or empirical input is required. This also means that no adjustable parameters are available to fit the simulation results to experimental data. These approaches are based on the evaluation of the transitions between the various states due to different scattering mechanisms, including the interaction of the electrons with phonons, impurities, and other electrons. The associated scattering rates are computed based on the corresponding Hamiltonian. A direct numerical solution of the resulting equations is not possible since the transition rates depend on the initially unknown electron populations. Rather, the steady state solution must be self-consistently found by simulating the temporal evolution of the system until convergence is reached, or by using iterative schemes. These methods rely on the subband wave functions and eigenenergies found by solving the Schrödinger or Schrödinger-Poisson equation, as described in Sec. II. The carrier transport is then modeled by transitions of discrete electrons between the quantized energy levels, also referred to as hopping transport. Thus, these methods are called semiclassical, since the energy quantization in the QCL heterostructure is considered, but quantum coherence effects and dephasing mechanisms are not included. Formally, semiclassical carrier transport descriptions can be derived from the more general density matrix formalism by neglecting the contribution of the off-diagonal matrix elements, i.e., only considering the diagonal elements corresponding to the occupations of the states.[114,131] In the following, the most relevant scattering mechanisms in the QCL will be discussed. Furthermore, the self-consistent rate equation approach and the EMC method will be described, which are the two most widely used advanced semiclassical QCL simulation schemes.

A. Scattering mechanisms and transition rates

A transition of a carrier from one state to another due to a perturbation is referred to as a scattering process. The perturbation is described by a corresponding potential V, which can be static or time dependent.[61] This results in different classes of scattering processes, illustrated in Fig. 10 for the case of one-dimensional electron confinement as in QCL heterostructures. Elastic scattering, shown in Fig. 10(a), occurs for time constant potentials. Here, the carrier energy is conserved. Relevant mechanisms in QCLs include impurity, interface roughness, and alloy scattering. For potentials with harmonic time dependence, $V \propto \cos(\omega_0 t)$, the carrier energy is changed by $\mp \hbar \omega_0$, corresponding to the case of emission
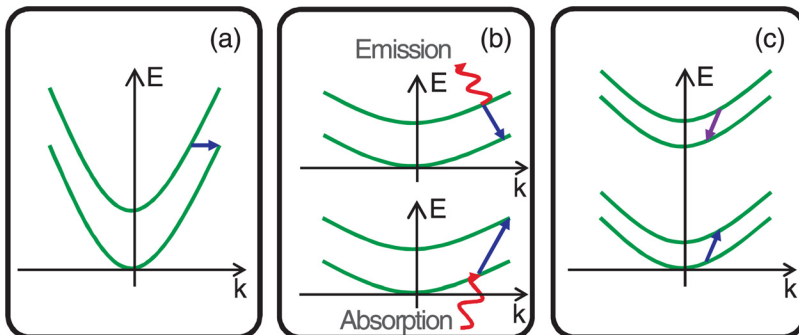

FIG. 10. Different classes of scattering processes: (a) elastic scattering; (b) inelastic scattering; (c) carrier-carrier scattering.
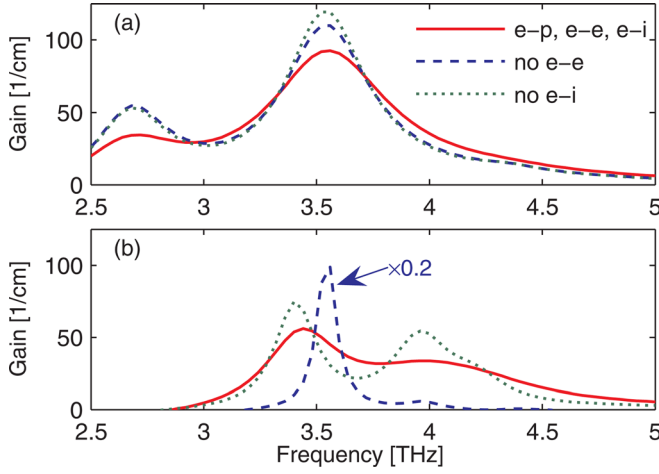
FIG. 11. Simulation results for the spectral gain vs. frequency, as obtained by evaluating electron-phonon (e-p), electron-impurity (e-i), and electron-electron (e-e) scattering (solid curves) and neglecting e-e (dashed curves) or e-i (dotted curves) scattering. (a) 3.4 THz resonant phonon depopulation structure; (b) 3.5 THz bound-to-continuum structure. Reprinted with permission from J. Appl. Phys. **105**, 123102 (2009). Copyright 2009 AIP Publishing LLC.[107]

and absorption, respectively. This is referred to as inelastic scattering [see Fig. 10(b)]. An important inelastic mechanism is optical phonon scattering, and also the interaction with photons can be viewed as an inelastic scattering process. A special case is intercarrier process illustrated in Fig. 10(c) such as electron-electron scattering, where two electrons are involved in the scattering event. Figure 11 shows the influence of various scattering mechanisms on the spectral gain obtained from an EMC simulation for two different terahertz QCL designs.[107]

QCLs are n-type devices which are typically based on direct bandgap semiconductors. Thus, in the following, we restrict our treatment of scattering to the conduction band Γ valley where also the lasing transitions occur. However, we note that under certain conditions transitions to other valleys can affect QCL operation.[45–48] Scattering causes an electron transition from an initial state $|i\mathbf{k}\rangle$ to a final state $|j\mathbf{k}'\rangle$ in the QCL heterostructure, where $\mathbf{k} = (k_x, k_y)^\mathrm{T}$ and $\mathbf{k}' = (k_x', k_y')^\mathrm{T}$ are the corresponding in-plane wave vectors. The states are described by their wave functions Eq. (2) and eigenenergies Eq. (5). For the initial state, we have $\psi_{3\mathrm{D},i} = S^{-1/2}\psi_i(z)\exp(i\mathbf{k}\mathbf{r})$ with $\mathbf{r} = (x, y)^\mathrm{T}$ and $E_{i\mathbf{k}} = E_i + \hbar^2\mathbf{k}^2/(2m_i^\|)$ where $\psi_i(z)$ and E_i are obtained from Eq. (4) as described in Sec. II A, and analogously for the final state.

For elastic scattering processes, V is constant. The corresponding matrix element is defined as

$$V_{j\mathbf{k}',i\mathbf{k}} = \langle j\mathbf{k}'|V|i\mathbf{k}\rangle$$
$$= S^{-1}\int_S\int_{-\infty}^{\infty} V\psi_j^*\psi_i\exp[i(\mathbf{k}-\mathbf{k}')\mathbf{r}]\mathrm{d}^2r\mathrm{d}z. \quad (55)$$

For inelastic processes, we assume a harmonic potential of the form $V = V_0\exp(i\mathbf{Q}\mathbf{x} - i\omega_0 t) + V_0^*\exp(-i\mathbf{Q}\mathbf{x} + i\omega_0 t)$, where $\mathbf{Q}$ is, for example, the phonon wave vector. The matrix element can in analogy with Eq. (55) be obtained as

$$\begin{pmatrix} V_{j\mathbf{k}',i\mathbf{k}}^+(Q) \\ V_{j\mathbf{k}',i\mathbf{k}}^-(Q) \end{pmatrix} = \langle j\mathbf{k}'|\begin{pmatrix} V_0\exp(i\mathbf{Q}\mathbf{x}) \\ V_0^*\exp(-i\mathbf{Q}\mathbf{x}) \end{pmatrix}|i\mathbf{k}\rangle$$
$$= S^{-1}\int_S\int_{-\infty}^{\infty}\psi_j^*\psi_i\begin{pmatrix} V_0\exp(i\mathbf{Q}\mathbf{x}) \\ V_0^*\exp(-i\mathbf{Q}\mathbf{x}) \end{pmatrix}$$
$$\times \exp[i(\mathbf{k}-\mathbf{k}')\mathbf{r}]\mathrm{d}^2r\mathrm{d}z. \quad (56)$$

Assuming bound wave functions $\psi_{i,j}(z)$ in Eqs. (55) and (56), the integration over the z coordinate can be taken from $-\infty$ to ∞. This is also consistent with treating the gain medium as an infinitely extended periodic heterostructure. Furthermore, the in-plane cross section S is assumed to be macroscopic, and thus the integration can for scattering rate calculations be extended from $-\infty$ to ∞ also in x and y direction.

The transition rate from an initial state $|i\mathbf{k}\rangle$ to a final state $|j\mathbf{k}'\rangle$ is obtained from Fermi's golden rule,[61] given by

$$W_{i\mathbf{k},j\mathbf{k}'} = \frac{2\pi}{\hbar}|V_{j\mathbf{k}',i\mathbf{k}}|^2\delta(E_{j\mathbf{k}'} - E_{i\mathbf{k}}) \quad (57)$$

for elastic scattering processes and

$$W_{i\mathbf{k},j\mathbf{k}'}^\pm(Q) = \frac{2\pi}{\hbar}|V_{j\mathbf{k}',i\mathbf{k}}^\mp(Q)|^2\delta(E_{j\mathbf{k}'} - E_{i\mathbf{k}} \pm \hbar\omega_0) \quad (58)$$

for inelastic processes. Here, $W_{i\mathbf{k},j\mathbf{k}'}^+$ corresponds to the emission rate, caused by the component $\exp(i\omega_0 t)$, and $W_{i\mathbf{k},j\mathbf{k}'}^-$ refers to absorption due to the component $\exp(-i\omega_0 t)$. The Dirac δ function ensures energy conservation. For elastic processes, we obtain from $E_{j\mathbf{k}'} = E_{i\mathbf{k}}$

$$|\mathbf{k}'| = k' = \left(\frac{m_j^\|}{m_i^\|}k^2 + 2m_j^\|\frac{E_i - E_j}{\hbar^2}\right)^{1/2}. \quad (59)$$

Analogously, energy conservation yields for inelastic scattering

$$k' = \left(\frac{m_j^\|}{m_i^\|}k^2 + 2m_j^\|\frac{E_i - E_j \mp \hbar\omega_0}{\hbar^2}\right)^{1/2}. \quad (60)$$

The computation of the total transition rates from an initial state $|i\mathbf{k}\rangle$ to a final subband j involves the summation over wave vectors. These sums can be converted to integrals, introducing a factor of $L_d/(2\pi)$ per dimension where the device length L_d in the corresponding direction is assumed to be large enough that quantization effects can be neglected.[132] An example is the summation over the final in-plane wave vector $\mathbf{k}'$ which is two-dimensional. It can furthermore be advantageous to express $\mathbf{k}'$ in polar coordinates $|\mathbf{k}'|$ and ϕ, and introduce a kinetic energy variable $E_j^\mathrm{kin} = \hbar^2|\mathbf{k}'|^2/(2m_j^\|)$. Thus, we obtain

$$\sum_{\mathbf{k}'}\cdots \rightarrow \frac{S}{(2\pi)^2}\int\int_{-\infty}^{\infty}\cdots\mathrm{d}^2k'$$
$$= \frac{Sm_j^\|}{(2\pi\hbar)^2}\int_0^\infty\int_0^{2\pi}\cdots\mathrm{d}\phi\mathrm{d}E_j^\mathrm{kin}. \quad (61)$$

Spin degeneracy is not considered here, since for single-electron scattering processes the spin is conserved. For electron-electron scattering, the spin degeneracy must, however, be taken into account, as more closely discussed in the corresponding section. Analogously, summation over a three-dimensional wave vector, such as the phonon wave vector $\mathbf{Q}$, can in a crystal lattice of volume Ω_c be approximated by

$$\sum_{\mathbf{Q}} \ldots \rightarrow \frac{\Omega_c}{(2\pi)^3} \iiint_{-\infty}^{\infty} \ldots \mathrm{d}^3 Q. \tag{62}$$

B. Phonon scattering

A phonon is a quasiparticle associated with the lattice vibrations in a crystal, representing an excited quantum mechanical state in the quantization of the vibrational modes. Classically speaking, for an atom located at position $\mathbf{x}$, lattice vibrations are described by a displacement vector $\mathbf{u} = \mathbf{U}\sin(\mathbf{Q}\mathbf{x} - \omega_{\mathbf{Q}}t)$, with the amplitude $\mathbf{U}$, wave vector $\mathbf{Q}$, and angular frequency $\omega_{\mathbf{Q}}$. The normal modes are the solutions of $\mathbf{u}$ for which the lattice uniformly oscillates at a single frequency $\omega_{\mathbf{Q}}$, and a phonon corresponds to an elementary vibrational motion. The associated relation between wave vector $\mathbf{Q}$ and frequency $\omega_{\mathbf{Q}}$, the so-called dispersion relation, defines a phonon branch, as illustrated in Fig. 12. Acoustic modes are sound waves where two consecutive atoms move in the same direction, and we have $\omega_{\mathbf{Q}} = 0$ for $\mathbf{Q} = \mathbf{0}$. For optical modes, two consecutive atoms in the same unit cell move in opposite direction, and $\omega_{\mathbf{Q}}$ at $\mathbf{Q} = \mathbf{0}$ typically corresponds to infrared optical frequencies. Furthermore, the wave propagation can be predominantly longitudinal ($\mathbf{Q} \parallel \mathbf{U}$) or transverse ($\mathbf{Q} \perp \mathbf{U}$); the corresponding branches are then referred to as longitudinal or transverse branches. In three dimensions, there is a single longitudinal acoustic (LA) and 2 transverse acoustic (TA) branches as well as $N_u - 1$ longitudinal optical (LO) and $2(N_u - 1)$ transverse optical (TO) branches, where N_u denotes the number of atoms per unit cell. For example, GaAs has two atoms per unit cell, i.e., $N_u = 2$.

The lattice vibrations lead to a perturbation of the carriers and thus carrier scattering. Depending on the mechanisms, there are different types of phonon scattering. Non-polar phonon scattering occurs in all crystals and is due to acoustic and (for $N_u \geq 2$) TO phonons. Here, the lattice
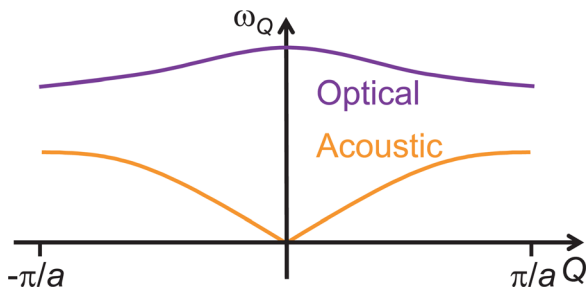

FIG. 12. Schematic illustration of the dispersion relation for an acoustic and optical phonon branch. This simplified example corresponds to a one-dimensional lattice with $N_u = 2$ and lattice constant a.

vibrations lead to a time dependent change of the conduction (and valence) band energy. Polar scattering only occurs in polar semiconductors, e.g., III–V semiconductors such as GaAs. It is due to LO phonons, where the out-of-phase movement of the neighbouring atoms of different types causes a local dipole moment, resulting in oscillating electric fields. Covalent semiconductors such as group IV materials do not exhibit polar scattering.

The dominant phonon scattering mechanism in QCLs is due to LO phonons. Since the optical phonon deformation potential approaches zero at the conduction band Γ point due to spherical symmetry, TO phonon scattering is negligible for the Γ valley; it can however play a significant role for intervalley transitions and intravalley scattering in the X and valence band Γ minimum.[133–135] Also acoustic phonons tend to play a secondary role in QCLs.[136]

1. Non-polar phonon scattering

Non-polar phonon scattering occurs in all crystals and is due to acoustic and (for $N_u \geq 2$) TO phonons. The energy of the vibrating atom is the sum of its potential and kinetic energy, which is equal to twice its average kinetic energy. Thus, the total energy of the vibrational mode becomes for $\mathbf{u} = \mathbf{U}\sin(\mathbf{Q}\mathbf{x} - \omega_{\mathbf{Q}}t)$

$$E = N_a m \langle |\partial_t \mathbf{u}|^2 \rangle_t = \frac{1}{2} N_a m U^2 \omega_{\mathbf{Q}}^2, \tag{63}$$

where the number of atoms with mass m in a crystal lattice of volume Ω_c and density ρ_c is $N_a = \Omega_c \rho_c / m$. Thus, we obtain the amplitude for a mode occupied by a single phonon of energy $E = \hbar\omega_{\mathbf{Q}}$

$$U = \sqrt{\frac{2\hbar}{\Omega_c \rho_c \omega_{\mathbf{Q}}}}. \tag{64}$$

a. Acoustic phonons. Since phonon confinement does not play an important role for the cascade structures (see discussion in Sec. VI B 2), scattering from bulk phonons obeying Bose distribution is considered. Lawaetz has shown in Ref. 137 that the deformation potential method of Bardeen and Shockley[138] may be applied to the scattering of electrons with acoustic phonons. This method is basically a Taylor expansion of the scattering potential in the phonon momentum $\mathbf{Q}$. In the case of vanishing screening, Kittel and Fong derive in Ref. 139 the change of the electronic energy in lowest order of the lattice deformation. Since for acoustic phonons, adjacent atoms move in the same direction, regions of compression or dilatation extend over several lattice sites, and the crystal can be described as an elastic continuum. The resulting strain gives rise to a time dependent change of the conduction band energy $V = \delta V_c$. The corresponding constant of proportionality is the deformation potential Ξ (for valleys without rotational symmetry, Ξ becomes a tensor), i.e., $V = \Xi\nabla\mathbf{u}$ for small displacements.[61] We obtain

$$V = \Xi\mathbf{Q}\mathbf{U}\cos(\mathbf{Q}\mathbf{x} - \omega_{\mathbf{Q}}t), \tag{65}$$

i.e., only LA phonons contribute since $\mathbf{QU} = 0$ for transverse phonons. For acoustic phonons, we have $\omega_\mathbf{Q} = 0$ for $Q = |\mathbf{Q}| = 0$, and the dispersion relation can be described by the linear approximation $\omega_\mathbf{Q} = v_s Q$ for small Q, where v_s is the longitudinal sound velocity (see Fig. 12). With Eq. (64), the perturbation potential is thus

$$V = M(Q)\{\exp[i(\mathbf{Qx} - \omega_\mathbf{Q}t)] + \exp[-i(\mathbf{Qx} - \omega_\mathbf{Q}t)]\}, \tag{66}$$

with the amplitude

$$M(Q) = \Xi\sqrt{\frac{\hbar Q^2}{2\Omega_c \rho_c \omega_\mathbf{Q}}} = \Xi\sqrt{\frac{\hbar Q}{2\Omega_c \rho_c v_s}}. \tag{67}$$

Equation (58) gives the transition rate for an electron in an initial state $|i\mathbf{k}\rangle$, which is scattered to a final state j with a wave vector $\mathbf{k}'$ by a phonon with wave vector $\mathbf{Q} = [\mathbf{q}, q_z]^\mathrm{T}$, where $\mathbf{q}$ corresponds to the in-plane component. Using the definition of the three-dimensional wave functions in Eq. (2), the corresponding matrix element of the perturbation potential Eq. (66) is with Eq. (56) given by

$$V^\pm_{j\mathbf{k}',i\mathbf{k}} = \frac{M(Q)}{S}\int_S \exp[i(\mathbf{k}\pm\mathbf{q}-\mathbf{k}')\mathbf{r}]d^2r F^\pm_{ji}(q_z), \tag{68}$$

with the form factor

$$F^\pm_{ji}(q_z) = \int_{-\infty}^\infty \psi_j^*(z)\exp(\pm iq_z z)\psi_i(z)dz. \tag{69}$$

For bound states, $\psi_{i,j}$ can be chosen to be real, and we have $|F^+_{ji}(q_z)|^2 = |F^-_{ji}(q_z)|^2 = |F_{ji}(q_z)|^2$. The scattering rate obtained from Eq. (58) is thus

$$W^\pm_{i\mathbf{k},j\mathbf{k}'}(\mathbf{Q}) = \frac{(2\pi)^3}{\hbar S}\left(N_\mathbf{Q} + \frac{1}{2}\pm\frac{1}{2}\right)|M(Q)|^2|F_{ji}(q_z)|^2$$
$$\times \delta(\mathbf{k}\mp\mathbf{q}-\mathbf{k}')\delta(E_{j\mathbf{k}'} - E_{i\mathbf{k}}\pm\hbar\omega_\mathbf{Q}), \tag{70}$$

where $W^+_{i\mathbf{k},j\mathbf{k}'}$ and $W^-_{i\mathbf{k},j\mathbf{k}'}$ refer to emission and absorption, respectively. Here, we have used that $|\int_S\exp[i(\mathbf{k}\pm\mathbf{q}-\mathbf{k}')\mathbf{r}]d^2r|^2$ can be approximated by $4\pi^2 S\delta(\mathbf{k}\mp\mathbf{q}-\mathbf{k}')$ for sufficiently large in-plane cross sections S. Since $M(Q)$ refers to a single phonon, the phonon occupation number of a mode in thermal equilibrium has been added in Eq. (70), given by the Bose-Einstein distribution

$$N_\mathbf{Q} = \left[\exp\left(\frac{\hbar\omega_\mathbf{Q}}{k_B T_L}\right) - 1\right]^{-1}. \tag{71}$$

For emission, the factor $N_\mathbf{Q} + 1$ is used to also include spontaneous emission processes.

The total transition rate from a given initial state $|i\mathbf{k}\rangle$ to a subband j is obtained by summing over all wave vectors $\mathbf{k}'$ and $\mathbf{Q}$. These sums can be converted to integrals using Eqs. (61) and (62). With Eqs. (67) and (70), we thus obtain the total transition rate from a given initial state $|i\mathbf{k}\rangle$ to a subband j[132]

$$W^\pm_{i\mathbf{k},j} = \frac{\Xi^2}{8\pi^2\rho_c v_s}\int Q\left(N_\mathbf{Q} + \frac{1}{2}\pm\frac{1}{2}\right)|F_{ji}(q_z)|^2$$
$$\times \delta[E_{j,\mathbf{k}\mp\mathbf{q}} - E_{i\mathbf{k}}\pm\hbar\omega_\mathbf{Q}]d^3Q. \tag{72}$$

Since acoustic phonons do not carry much energy, often the quasi-elastic approximation is applied, treating acoustic phonon scattering as elastic process. This is achieved by neglecting the phonon energy term $\pm\hbar\omega_\mathbf{Q}$ in Eqs. (70) and (72). Besides, we can approximate Eq. (71) as

$$N_\mathbf{Q} \approx N_\mathbf{Q} + 1 \approx \frac{k_B T_L}{\hbar\omega_\mathbf{Q}} \approx \frac{k_B T_L}{\hbar v_s Q}, \tag{73}$$

for all but the lowest temperatures, which is referred to as equipartition approximation. Thus, we obtain

$$W^\pm_{i\mathbf{k},j\mathbf{k}'}(\mathbf{Q}) = \Xi^2\frac{4\pi^3 k_B T_L}{\Omega_c \rho_c v_s^2 S\hbar}|F_{ji}(q_z)|^2$$
$$\times \delta(\mathbf{k}\mp\mathbf{q}-\mathbf{k}')\delta(E_{j\mathbf{k}'} - E_{i\mathbf{k}}). \tag{74}$$

With Eqs. (61) and (62), we furthermore obtain $W^\pm_{i\mathbf{k},j}(\mathbf{Q}) = 0$ for $E_j > E_{i\mathbf{k}}$, and otherwise

$$W^\pm_{i\mathbf{k},j} = \Xi^2\frac{k_B T_L m_j^\|}{4\pi\rho_c v_s^2\hbar^3}\int_{-\infty}^\infty |F_{ji}(q_z)|^2 dq_z. \tag{75}$$

The final wave vector magnitude k' is then given by Eq. (59). With Eq. (69), we obtain

$$\int_{-\infty}^\infty |F_{ji}(q_z)|^2 dq_z = 2\pi\int_{-\infty}^\infty |\psi_j(z)|^2|\psi_i(z)|^2 dz. \tag{76}$$

Thus, the total transition rate is given by

$$W_{i\mathbf{k},j} = \Xi^2\frac{k_B T_L m_j^\|}{\rho_c v_s^2\hbar^3}\int_{-\infty}^\infty |\psi_j(z)|^2|\psi_i(z)|^2 dz, \tag{77}$$

which includes both emission and absorption and thus contains an additional factor of 2.

b. Transverse optical phonons. Transverse optical phonons are treated in a similar manner as acoustic phonons. Here, adjacent atoms move in opposite directions. Thus, the crystal vibrations cannot be described as an elastic continuum as for acoustic phonons, but correspond rather to oscillations of the two sublattices with respect to each other. As a consequence, the potential change is directly a function of the displacement vector $\mathbf{u}$, $\delta V_c = \mathbf{D}_\mathrm{TO}\mathbf{u} = D_\mathrm{TO}u$. Here, $\mathbf{D}_\mathrm{TO}$ is the optical deformation potential field which assumes a different value for different kinds of intra- and intervalley transitions, and D_TO is the component parallel to $\mathbf{u}$. For optical phonons, $\omega_\mathbf{Q}$ reaches an extremum $\omega_\mathbf{Q} = \omega_\mathrm{TO}$ for $Q = 0$, and for small Q, the dispersion relation can thus be approximated by $\omega_\mathbf{Q} = \omega_\mathrm{TO}$ (see Fig. 12). The scattering rate for an electron from an initial state $|i\mathbf{k}\rangle$ to a final state $|j\mathbf{k}'\rangle$ due to phonons with wave vector $\mathbf{Q}$ is again given by Eq. (70), with

$$M(Q) = D_{TO}\sqrt{\frac{\hbar}{2\Omega_c\rho_c\omega_{TO}}}. \tag{78}$$

In analogy to Eq. (72), we obtain the total transition rate

$$W_{i\mathbf{k},j}^\pm = \frac{D_{TO}^2}{8\pi^2\rho_c\omega_{TO}}\int\left(N_\mathbf{Q}+\frac{1}{2}\pm\frac{1}{2}\right)|F_{ji}(q_z)|^2$$
$$\times \delta[E_{j,\mathbf{k}\mp\mathbf{q}}-E_{i\mathbf{k}}\pm\hbar\omega_{TO}]\mathrm{d}^3Q. \tag{79}$$

Approximating $\omega_\mathbf{Q}=\omega_{TO}$, the phonon number $N_\mathbf{Q}$ in Eq. (79) does not depend on $\mathbf{Q}$, and thus we can set $N_\mathbf{Q}=N_{Ph}$. With Eq. (76), the result then simplifies to

$$W_{i\mathbf{k},j}^\pm = \frac{D_{TO}^2 m_j^\parallel}{2\rho_c\omega_{TO}\hbar^2}\left(N_{Ph}+\frac{1}{2}\pm\frac{1}{2}\right)$$
$$\times \int_{-\infty}^\infty |\psi_j(z)|^2|\psi_i(z)|^2\mathrm{d}z, \tag{80}$$

for $E_j \leq E_{i\mathbf{k}}\mp\hbar\omega_{TO}$, and becomes 0 otherwise.

2. Longitudinal optical phonons

As for TO phonons, adjacent atoms move in opposite directions for LO phonons. Again the dispersion relation can be approximated with a constant value $\omega_\mathbf{Q}=\omega_{LO}$ for small Q (see Fig. 12), where $\hbar\omega_{LO}$ is the LO phonon energy at $Q=0$. Polar semiconductors contain different types of atoms carrying effective charges. For two different atoms with masses m_A and m_B in a unit cell, these effective charges are given by $\pm q_{eff}$ with $q_{eff}^2 = \epsilon_0\omega_{LO}^2\mu n_c^{-1}(\epsilon_{r,\infty}^{-1}-\epsilon_{r,0}^{-1})$.[61] Here, $\mu^{-1}=m_A^{-1}+m_B^{-1}$ is the inverse reduced mass, and n_c is the number of cells per unit volume. Furthermore, $\epsilon_{r,0}$ denotes the static dielectric constant and $\epsilon_{r,\infty}$ is the dielectric constant at very high frequencies where the ions cannot respond to a field anymore and thus only the electronic polarization remains. For LO waves, the out-of-phase movement of the atoms causes a local dipole moment $\mathbf{p}(t)=q_{eff}\mathbf{u}(t)$, where $\mathbf{u}=\mathbf{U}\sin(\mathbf{Qx}-\omega_\mathbf{Q}t)$ now indicates the relative displacement of the atoms in the unit cell; by contrast, transverse waves do not generate dipoles. The amplitude for excitation by a single phonon is given by Eq. (64) where ρ_c is replaced by μn_c. The resulting ionic polarization is $\mathbf{P}_{ion}=\mathbf{p}n_c$, and the electric field $\mathbf{E}$ can be obtained from the material equation $\mathbf{D}=\epsilon_0\mathbf{E}+\mathbf{P}_{ion}$ with $\mathbf{D}=\mathbf{0}$ due to the absence of free carriers. With the potential $V=e\int\mathbf{E}\mathrm{dx}$, the resulting scattering rate from a state $|i\mathbf{k}\rangle$ to $|j\mathbf{k}'\rangle$ is again given by Eq. (70), where

$$M(Q) = \frac{1}{2Q}e\sqrt{\epsilon_0^{-1}\left(\epsilon_{r,\infty}^{-1}-\epsilon_{r,0}^{-1}\right)}\sqrt{\frac{2\hbar\omega_{LO}}{\Omega_c}}. \tag{81}$$

The total transition rate from a given initial state $|i\mathbf{k}\rangle$ to a subband j is again obtained by summing over all wave vectors $\mathbf{k}'$ and $\mathbf{Q}$, using Eqs. (61) and (62). For $E_j \leq E_{i\mathbf{k}}\mp\hbar\omega_{LO}$, we obtain

$$W_{i\mathbf{k},j}^\pm = \frac{m_j^\parallel\omega_{LO}e^2}{8\pi^2\hbar^2\epsilon_0}\left(\epsilon_{r,\infty}^{-1}-\epsilon_{r,0}^{-1}\right)\left(N_{Ph}+\frac{1}{2}\pm\frac{1}{2}\right)\int_0^{2\pi}J(q)\mathrm{d}\theta, \tag{82}$$

and $W_{i\mathbf{k},j}^\pm = 0$ otherwise. Here,

$$J(q) = \frac{\pi}{q}\int\int_{-\infty}^\infty \psi_j^*(z)\psi_j(z')\psi_i(z)\psi_i^*(z')$$
$$\times \exp(-q|z-z'|)\mathrm{d}z\mathrm{d}z' \tag{83}$$

with the in-plane phonon wave vector magnitude

$$q = (k^2+k'^2-2kk'\cos\theta)^{1/2}, \tag{84}$$

where k' is given by Eq. (60) with $\omega_0=\omega_{LO}$.

a. Corrections due to screening, hot phonons, phonon modes, and lattice heating. In a heterostructure, the assumption of bulk phonon modes is justified for acoustic phonons, but not necessarily for LO phonons. The LO phonon modes arising from different dielectric constants in the well and barrier material can be described by microscopic and macroscopic approaches. The macroscopic dielectric continuum model, which is widely used in this context, yields slab modes which are confined in each layer, and interface modes corresponding to exponential solutions around the interfaces.[140] However, it has been shown that the bulk mode approximation is valid for most QCL designs.[48,101] Specifically, for not too different compositions of the barrier and well material and thus similar values of $\epsilon_{r,0}$, the use of bulk modes is a good approximation.[101]

Screening can be included by changing Eq. (81) to[141,142]

$$M(Q) = \frac{Q}{2(Q^2+q_s^2)}e\sqrt{\epsilon_0^{-1}\left(\epsilon_{r,\infty}^{-1}-\epsilon_{r,0}^{-1}\right)}\sqrt{\frac{2\hbar\omega_{LO}}{\Omega_c}}, \tag{85}$$

where q_s is the inverse screening length. For Eq. (83), we then obtain with $\tilde{q}=\sqrt{q^2+q_s^2}$

$$J(q) = \frac{\pi}{\tilde{q}}\int\int_{-\infty}^\infty \psi_j^*(z)\psi_j(z')\psi_i(z)\psi_i^*(z')\exp(-\tilde{q}|z-z'|)$$
$$\times \left(1-\frac{|z-z'|q_s^2}{2\tilde{q}}-\frac{q_s^2}{2\tilde{q}^2}\right)\mathrm{d}z\mathrm{d}z'. \tag{86}$$

The simplest model is to use Debye screening with the bulk inverse Debye screening length

$$q_s = \sqrt{\frac{n_e e^2}{\epsilon_0\epsilon_{r,0}k_B T_L}}, \tag{87}$$

where n_e is the averaged three-dimensional electron density. However, Debye screening is generally only valid for high temperatures,[143] since q_s diverges for $T_L \rightarrow 0$.

Deviations from the equilibrium phonon distribution Eq. (71) due to phonon emission and absorption in the heterostructure, also referred to as hot phonons, are frequently considered in simulations.[104,144] The hot phonon effect is commonly incorporated by using a relaxation time approach, describing the decay of LO phonons into other phonon modes by a corresponding lifetime.[145]

Typically, in simulations it is assumed that the lattice temperature of the gain medium is identical to a given heat sink temperature or to the room temperature. However, this approximation is only legitimate for pulsed operation at low duty cycle, and generally fails for continuous wave operation.[146] Lattice heating due to the dissipated electrical power can be self-consistently modeled by coupled carrier transport and thermal simulations,[147,148] yielding the actual temperature profile. The temperature distribution is computed by solving the heat diffusion equation, where the heat generation rate is obtained from the carrier transport simulation based on the dissipated electrical power or the LO phonon scattering rate.

C. Electron-electron scattering

Electron-electron scattering arises from the collision of electrons with other electrons. More precisely, this scattering mechanism can be divided into two main contributions, the short-range interaction between two electrons via the screened Coulomb potential and the long-range electron-plasmon coupling.[149] The latter is usually neglected in QCL simulations and is also not considered here, since it only becomes relevant at elevated doping levels.

Many-electron effects are to lowest order already considered in the Poisson equation, Eq. (20), corresponding to a mean field treatment of charges. In the semiclassical treatment, an inclusion of higher order effects is obtained by corresponding scattering rates. Electron-electron scattering is much more computationally demanding than the single-electron processes, hampering its inclusion in quantum mechanical simulations of QCLs beyond the mean-field approximation.[118]

In semiclassical simulations, electron-electron scattering is typically implemented as a two-electron process.[141,150] An electron in an initial state $|i\mathbf{k}\rangle$ scatters to a final state $|j\mathbf{k}'\rangle$, accompanied by a transition of a second electron from a state $|i_0\mathbf{k}_0\rangle$ to $|j_0\mathbf{k}_0'\rangle$. Nonparabolicity effects have for example been considered by Bonno and Thobel.[55] To facilitate the treatment, we assume here an identical mass m^* for all subbands. The total scattering rate from $|i\mathbf{k}\rangle$ to a subband j is then obtained as[141,150]

$$W_{i\mathbf{k},j} = \frac{m^*}{4\pi\hbar^3 S} \sum_{i_0,j_0,\mathbf{k}_0} f_{i_0}(\mathbf{k}_0) \int_0^{2\pi} d\theta |M_{ii_0jj_0}(q)|^2, \quad (88)$$

with the in-plane cross section area S and carrier distribution function $f_{i_0}(\mathbf{k}_0)$ for the state $|i_0\mathbf{k}_0\rangle$. θ is the angle between $\mathbf{g} = \mathbf{k}_0 - \mathbf{k}$ and $\mathbf{g}' = \mathbf{k}_0' - \mathbf{k}'$, and $\mathbf{q} = \mathbf{k} - \mathbf{k}'$ (with $q = |\mathbf{q}|$) denotes the exchanged wave vector. For the definition of the Coulomb matrix element, different conventions exist.[141,151] Here, we use the widely employed convention of Goodnick and Lugli,[110,141,150,152] where the bare Coulomb matrix element is with $\epsilon = \epsilon_0 \epsilon_{r,0}$ given by

$$V_{ii_0jj_0}^b(q) = \frac{e^2}{2\epsilon q} \int_{-\infty}^{\infty} dz \int_{-\infty}^{\infty} dz' [\psi_i(z)\psi_{i_0}(z')$$
$$\times \psi_j^*(z)\psi_{j_0}^*(z')\exp(-q|z-z'|)]. \quad (89)$$

Electron-electron scattering is sensitive with respect to the spin. Furthermore, screening has to be considered, caused by the response of the other electrons to changes in the electrostatic field associated with the scattering process. Often, only scattering of electron pairs with antiparallel spin is considered, since the contribution of carriers with parallel spin is smaller.[141,153] Furthermore, basic models consider screening by a constant screening wave number q_s. For the transition matrix element $M_{ii_0jj_0}$, we then obtain

$$|M_{ii_0jj_0}(q)|^2 = \frac{1}{2} \left| \frac{q}{q+q_s} V_{ii_0jj_0}^b(q) \right|^2, \quad (90)$$

where the factor $q/(q+q_s)$ corrects the bare Coulomb matrix element for screening effects, and the factor $1/2$ arises from neglecting parallel spin contributions. From energy conservation, we obtain for the exchanged wave vector with $g = |\mathbf{g}|$ and $g_0^2 = 4m^*(E_i + E_{i_0} - E_j - E_{j_0})/\hbar^2$

$$q = \frac{1}{2}[2g^2 + g_0^2 - 2g(g^2 + g_0^2)^{1/2}\cos\theta]^{1/2}. \quad (91)$$

A frequently used basic screening model is the single subband screening approximation, assuming that most electrons in each period reside in a single subband i, corresponding to the ground state. The obtained screening wave number is then[110,152]

$$q_s = \frac{e^2}{2\epsilon} \frac{m^*}{\pi\hbar^2} f_i(\mathbf{k} = \mathbf{0}). \quad (92)$$

1. Advanced spin and screening models

Advanced implementations of spin and screening effects have been developed for a more accurate description of electron-electron scattering.[68] Screening models with varying degrees of sophistication are employed to obtain the screened Coulomb matrix elements $V_{ii_0jj_0}^s$ from $V_{ii_0jj_0}^b$ in Eq. (89). In the random phase approximation (RPA), $V_{ii_0jj_0}^s(q)$ is found by solving the equation system[154]

$$V_{ii_0jj_0}^s = V_{ii_0jj_0}^b + \sum_{mn} V_{imjn}^b \Pi_{mn} V_{mi_0nj_0}^s. \quad (93)$$

Here, $\Pi_{mn}(q)$ is the polarizability tensor, given in the long wavelength limit $(q \to 0)$ by

$$\Pi_{mn} = \begin{cases} \dfrac{n_m^s - n_n^s}{E_m - E_n}, & m \neq n, \\ -\dfrac{m^*}{\pi\hbar^2} f_n(0), & m = n. \end{cases} \quad (94)$$

Commonly, simplified screening models using a constant screening wave number q_s are employed to avoid the numerical load associated with solving Eq. (93).[110,152] $V_{ii_0jj_0}^s$ is then obtained by replacing the prefactor $e^2/(2\epsilon q)$ in Eq. (89) by $e^2/[2\epsilon(q+q_s)]$. This implementation is also used in Eq. (90). The single subband screening model, Eq. (92), can be obtained from Eq. (93) by assuming that most of the electrons reside in a single subband, and screening is caused only

by this subband.[110,152] A somewhat improved approach which considers all subbands equally is the modified single subband model[152] with

$$q_{\mathrm{s}} = \frac{e^2}{2\epsilon} \frac{m^*}{\pi\hbar^2} \sum_i f_i(\mathbf{k} = \mathbf{0}), \qquad (95)$$

where i sums over the subbands in one period. An alternative approach consists in treating intersubband scattering as unscreened, thus applying Eq. (95) only to the intrasubband matrix elements.[46]

For collisions of electrons with parallel spin, interference occurs between $V^{\mathrm{s}}_{ii_0jj_0}$ and the "exchange" matrix element $V^{\mathrm{s}}_{ii_0j_0j}$.[153] Accounting for this exchange effect, we obtain[150,153]

$$|M_{ii_0jj_0}|^2 = \frac{p_{\mathrm{a}}}{2}[|V^{\mathrm{s}}_{ii_0jj_0}(q^-)|^2 + |V^{\mathrm{s}}_{ii_0j_0j}(q^+)|^2] + \frac{p_{\mathrm{p}}}{2}|V^{\mathrm{s}}_{ii_0jj_0}(q^-) - V^{\mathrm{s}}_{ii_0j_0j}(q^+)|^2, \qquad (96)$$

where

$$q^{\pm} = \frac{1}{2}[2g^2 + g_0^2 \pm 2g(g^2 + g_0^2)^{1/2}\cos\theta]^{1/2}, \qquad (97)$$

and $p_{\mathrm{a}} = p_{\mathrm{p}} = 1/2$ are the probabilities for antiparallel and parallel spin collisions, respectively. There are two common approaches to implement electron-electron scattering without explicitly considering the spin dependence. One method which has been used in Eq. (90) is to completely neglect the parallel spin collisions,[141,153] implying $p_{\mathrm{a}} = 1/2$, $p_{\mathrm{p}} = 0$ in Eq. (96). This approach tends to overestimate the exchange effect. Another common method is to ignore the exchange effect, i.e., parallel spin collisions are treated the same way as antiparallel spin contributions.[153] This corresponds to $p_{\mathrm{a}} = 1$, $p_{\mathrm{p}} = 0$ in Eq. (96).

Figure 13 contains EMC simulations of the gain spectra for different implementations of screening and spin effects.[68] In Fig. 13(a), the exchange effect is fully included, but different screening models are employed. The reference curve is based on the exact evaluation of the RPA (solid curve). Applying the simplified screening model Eq. (95) to all matrix elements overestimates the screening of the intersubband elements, thus leading to an underestimation of scattering. The resulting spectral gain profile (dashed curve) features an excessively narrow gain bandwidth and enhanced peak gain. On the other hand, completely ignoring the screening effect for the intersubband matrix elements overestimates the intersubband scattering, thus resulting in a lowered and broadened gain profile (dotted curve). In Fig. 13(b), screening is included in the RPA. Results are shown for $p_{\mathrm{a}} = p_{\mathrm{p}} = 1/2$ (solid curve), $p_{\mathrm{a}} = 1$, $p_{\mathrm{p}} = 0$ (dotted curve), $p_{\mathrm{a}} = 1/2$, $p_{\mathrm{p}} = 0$ (dashed curve), and $p_{\mathrm{a}} = p_{\mathrm{p}} = 0$ (dashed-dotted curve). The last case, which corresponds to completely neglecting electron-electron scattering in the simulation, yields two narrow gain spikes at around 2.8 and 3.6 THz, largely deviating from the experimental electroluminescence measurements.[69] Ignoring the exchange effect (dotted curve) leads to an overestimation of the scattering, and thus a reduced peak gain and increased gain bandwidth. On the other hand, completely
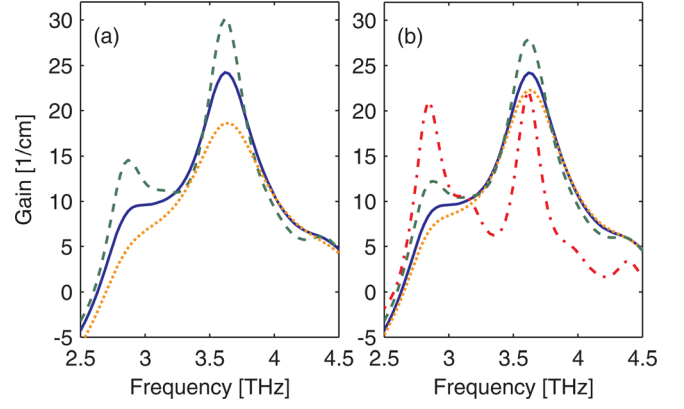

FIG. 13. EMC simulation results for the spectral gain vs. frequency of a bound-to-continuum terahertz QCL[69] at a lattice temperature $T_{\mathrm{L}} = 10\,\mathrm{K}$. (a) Comparison of different screening models: RPA (solid curve), modified single subband model for all matrix elements (dashed curve) or for the intrasubband elements only (dotted curve), thus treating intersubband elements as unscreened. (b) Comparison of different spin implementations. Shown are results for fully taking into account (solid curve) and ignoring the exchange effect (dotted curve), and ignoring parallel spin collisions (dashed curve). For comparison, the result obtained with no electron-electron scattering is also displayed (dashed-dotted curve). Reprinted with permission from J. Appl. Phys. **107**, 013104 (2010). Copyright 2010 AIP Publishing LLC.[68]

neglecting parallel spin collisions (dashed curve) leads to an underestimation of scattering and thus an enhanced peak gain.

D. Elastic scattering processes

Impurity scattering has been shown to play an important role in QCLs,[102,143,151] where it is often the dominant elastic scattering process. Also interface roughness can have a considerable impact on QCL operation.[107,155–157] Alloy scattering occurs in semiconductor alloys such as AlGaAs and other ternary materials, and, usually, only has to be considered for the well material since the wave functions are largely localized in the wells.[36] For elastic intrasubband scattering in the conduction band Γ valley, we have $k = k'$ due to energy conservation, i.e., elastic intrasubband scattering does not affect the carrier distribution due to in-plane isotropy.

1. Impurity scattering

Impurity scattering in QCLs arises from the doping, e.g., the ionized donors in the heterostructure. Charged impurities are to lowest order already considered in the Poisson equation, Eq. (20), corresponding to a mean field treatment of charges. An inclusion of higher order effects is in the semiclassical treatment obtained by corresponding scattering rates. The perturbation potential due to a charged impurity at position $(x', y', z')^{\mathrm{T}}$ is with $\mathbf{r}' = (x', y')^{\mathrm{T}}$ and $\epsilon = \epsilon_0\epsilon_{\mathrm{r},0}$ given by the Coulomb potential

$$V = -(4\pi\epsilon)^{-1}e^2(|\mathbf{r} - \mathbf{r}'|^2 + |z - z'|^2)^{-1/2}. \qquad (98)$$

For the definition of the corresponding matrix element, various conventions exist.[151,152] In analogy to Eq. (89) for electron-electron scattering, we define the bare matrix element as[152]

$$V_{ij}^{b}(z', q) = -\frac{e^2}{2\epsilon q} \int_{-\infty}^{\infty} \psi_i(z)\psi_j^*(z)\exp(-q|z - z'|)\mathrm{d}z. \quad (99)$$

The scattering rate is obtained from Eq. (57) as

$$W_{i\mathbf{k},j\mathbf{k}'} = \frac{2\pi}{S\hbar} \int |V_{ij}^{b}(z', q)|^2 n_{\mathrm{D}}(z')\mathrm{d}z'\delta(E_{j\mathbf{k}'} - E_{i\mathbf{k}})$$

$$= \frac{\pi e^4}{2S\epsilon^2\hbar q^2} F_{ij}(q)\delta(E_{j\mathbf{k}'} - E_{i\mathbf{k}}), \quad (100)$$

with the form factor

$$F_{ij}(q) = \int \mathrm{d}z' n_{\mathrm{D}}(z')$$

$$\times \left[\int_{-\infty}^{\infty} \psi_i(z)\psi_j^*(z)\exp(-q|z - z'|)\mathrm{d}z \right]^2. \quad (101)$$

Here, we have summed over the donors to include the effect of all ionized impurities, corresponding to an integration $S\int ... n_{\mathrm{D}}(z)\mathrm{d}z$, with the doping concentration $n_{\mathrm{D}}(z)$. Furthermore, an additional factor S^{-2} has been added since the prefactor S^{-1} is omitted in Eq. (99). Due to energy conservation $E_{i\mathbf{k}} = E_{j\mathbf{k}'}$, the final wave vector magnitude k' is given by Eq. (59).

The total transition rate from a given initial state $|i\mathbf{k}\rangle$ to a subband j is found by summing over all final wave vectors $\mathbf{k}'$ using Eq. (61). We obtain[54,152]

$$W_{i\mathbf{k},j} = \frac{m_j^{\|}e^4}{4\pi\epsilon^2\hbar^3} \int_0^{\pi} \frac{F_{ij}(q)}{q^2}\mathrm{d}\theta, \quad (102)$$

with $q(\theta) = |\mathbf{k} - \mathbf{k}'| = (k^2 + k'^2 - 2kk'\cos\theta)^{1/2}$. With Eq. (59), this yields

$$q(\theta) = \left[\left(1 + \frac{m_j^{\|}}{m_i^{\|}} \right) k^2 + q_0^2 - 2k \left(\frac{m_j^{\|}}{m_i^{\|}} k^2 + q_0^2 \right)^{1/2} \cos\theta \right]^{1/2}, \quad (103)$$

where $q_0^2 = 2m_j^{\|}(E_i - E_j)/\hbar^2$. Values of θ resulting in complex $q(\theta)$ should be excluded from the integration in Eq. (102) since scattering is then impossible.

a. Screening. As for electron-electron scattering discussed in Sec. VI C, the random phase approximation can also be used to obtain the screened matrix elements $V_{ij}^{s}(z', q)$ for impurity scattering from the bare elements, $V_{ij}^{b}(z', q)$ in Eq. (99). We obtain[151,152]

$$V_{ij}^{s} = V_{ij}^{b} + \sum_{mn} V_{imjn}^{b}\Pi_{mn}V_{mn}^{s}, \quad (104)$$

with V_{imjn}^{b} and Π_{mn} defined in Eqs. (89) and (94), respectively. In Eq. (100), $V_{ij}^{b}(z', q)$ has then to be replaced by $V_{ij}^{s}(z', q)$, and also Eq. (102) has to be changed correspondingly. As for electron-electron scattering, also here often simplified screening models using a constant screening wave number q_s are employed to avoid the numerical load

associated with solving Eq. (104).[54,151] Common models to handle screening are the model of Brooks and Herring[158] and the model of Conwell and Weisskopf.[159] While the model of Conwell and Weisskopf is based on a Rutherford-like scattering of electrons on bare Coulomb potentials with a cut-off radius, Brooks and Herring describe the impurity potential as being screened by free carriers. This latter approach is valid, if the constant screening length is much larger than the electronic wave length, which requires high temperatures and low carrier densities. For higher densities and low temperatures, the incorporation of more realistic screening has been shown to be essential,[160] whereas the limit of negligible free carrier screening is better described by the approach of Conwell and Weisskopf.[149] A detailed overview of the various refinements to the approach of Brooks and Herring has been given in Ref. 161. Electron scattering from screened impurities due to Brooks and Herring will be discussed in more detail in Sec. VIII D 3.

2. Interface roughness scattering

It has been shown that scattering from rough interfaces can change the QCL performance significantly.[107,155–157] There are two fundamentally different models for the interface roughness scattering in literature. In the first model, rough interfaces of quantum wells cause fluctuating well widths and fluctuating confinement energies of resonant states. In this model, the scattering potential is identified with the change of the well eigenenergies.[162,163] In the second model, fluctuations of the band offset yields the scattering potential.[164,165] The latter model may be favored for generality, since it does not require the existence of confined states, nor the distinguishability of level broadening by rough interfaces from other mechanisms. Nag has shown in Ref. 166 that both models agree well for quantum wells of various dimensions and materials. Thus, we consider interface roughness scattering due to the imperfections in the interface between the barrier and well material in the heterostructure, causing a local deviation of the interface $\Delta(x, y) = \Delta(\mathbf{r})$ from its average position z_0 as illustrated in Fig. 14. Since it is not feasible to measure $\Delta(\mathbf{r})$ directly, the interface roughness is characterized by its standard deviation Δ and correlation length Λ. Typically, a Gaussian autocorrelation function is assumed[162]

$$\langle \Delta(\mathbf{r})\Delta(\mathbf{r}') \rangle = \frac{1}{S} \int \Delta(\mathbf{r})\Delta(\mathbf{r} + \mathbf{d})\mathrm{d}^2r$$

$$= \Delta^2 \exp\left(-\frac{\mathbf{d}^2}{\Lambda^2} \right), \quad (105)$$

with $\mathbf{d} = \mathbf{r}' - \mathbf{r}$. A change $\Delta(\mathbf{r})$ of the interface position corresponds to a perturbing potential

$$V = \pm V_{\mathrm{o}}\left[H(z - z_0) - H(z - z_0 - \Delta(\mathbf{r})) \right], \quad (106)$$

where z_0 and V_{o} are the average interface position and the band offset, respectively. The "+" ("−") sign is obtained if the barrier (well) is located at $z < z_0$. Furthermore, H denotes the Heaviside step function. Since $\Delta(\mathbf{r})$ is small, i.e., on the order of a monolayer, we can approximate $\psi_{i,j}(z_0 + \Delta(\mathbf{r})) \approx \psi_{i,j}(z_0)$. Thus, we obtain with Eq. (55)
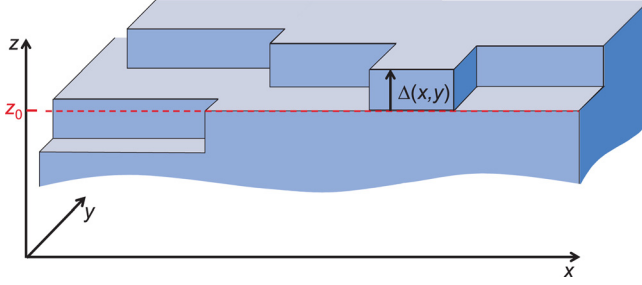
FIG. 14. Interface between two heterostructure layers consisting of different materials. The local deviation from a perfect interface due to roughness is expressed by $\Delta(x,y)$.

$$V_{j\mathbf{k}',i\mathbf{k}} = \pm \frac{V_\mathrm{o}}{S} \psi_i(z_0)\psi_j^*(z_0) \int \mathrm{d}^2 r \Delta(\mathbf{r})\exp(i\mathbf{q}\mathbf{r}), \quad (107)$$

and

$$|V_{j\mathbf{k}',i\mathbf{k}}|^2 = \frac{V_\mathrm{o}^2}{S^2} |\psi_i(z_0)\psi_j^*(z_0)|^2$$
$$\times \int \mathrm{d}^2 d \left[\exp(i\mathbf{q}\mathbf{d}) \int \mathrm{d}^2 r \Delta(\mathbf{r})\Delta(\mathbf{r}+\mathbf{d}) \right], \quad (108)$$

with $\mathbf{q} = \mathbf{k} - \mathbf{k}'$. Assuming a Gaussian autocorrelation Eq. (105), we obtain from Eq. (57) the scattering rate

$$W_{i\mathbf{k},j\mathbf{k}'} = \frac{2\pi^2}{\hbar S} V_\mathrm{o}^2 \Delta^2 \Lambda^2 \sum_n |\psi_i(z_n)\psi_j^*(z_n)|^2$$
$$\times \exp\left(-\frac{1}{4}\Lambda^2 q^2\right) \delta(E_{j\mathbf{k}'} - E_{i\mathbf{k}}), \quad (109)$$

where we additionally sum over all interfaces, located at positions z_n. Due to the energy conservation $E_{i\mathbf{k}} = E_{j\mathbf{k}'}$, the final wave vector magnitude k' is given by Eq. (59).

The total transition rate from a given initial state $|i\mathbf{k}\rangle$ to a subband j is found by summing over all final wave vectors $\mathbf{k}'$ using Eq. (61). We obtain

$$W_{i\mathbf{k},j} = \frac{m_j^\parallel}{\hbar^3} V_\mathrm{o}^2 \Delta^2 \Lambda^2 \sum_n |\psi_i(z_n)\psi_j^*(z_n)|^2$$
$$\times \int_0^\pi d\theta \exp\left(-\frac{1}{4}\Lambda^2 q^2\right) H(q^2), \quad (110)$$

where $q(\theta)$ is again given by Eq. (103), and values of θ with $q^2 < 0$ should be excluded from the integration since then scattering cannot occur.

The characterization of interface roughness by the parameters Λ and Δ is somewhat phenomenological, since it is difficult to directly extract them from experimental measurements of the interface profile. Rather, they are typically chosen to reproduce the measured optical transition linewidths.[167] For GaAs-based terahertz QCLs, frequently used values are $\Lambda \approx 10\,\mathrm{nm}$ and $\Delta \approx 0.1\,\mathrm{nm}$.[143,155] In Fig. 15, EMC simulation results for the gain profile are shown for two different types of terahertz QCLs.[107] In addition to scattering of electrons with phonons, impurities, and other electrons, interface roughness scattering with different values of
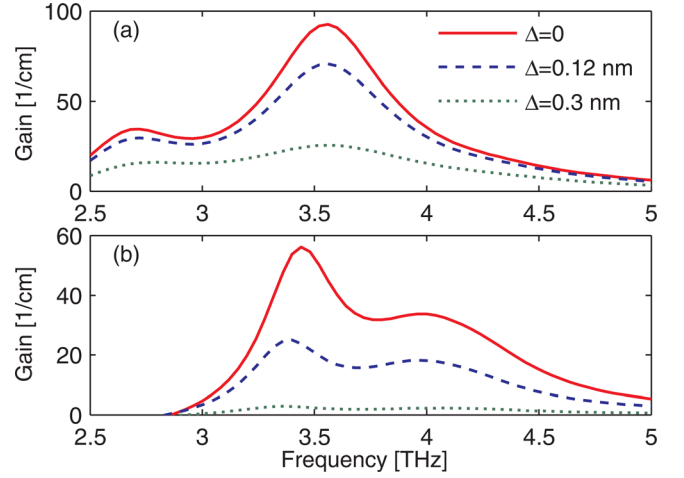

FIG. 15. Simulation results for the spectral gain vs. frequency, as obtained for a correlation length $\Lambda = 10\,\mathrm{nm}$ and different values of the interface roughness mean height Δ. (a) 3.4 THz resonant phonon depopulation structure; (b) 3.5 THz bound-to-continuum structure. Reprinted with permission from J. Appl. Phys. **105**, 123102 (2009). Copyright 2009 AIP Publishing LLC.[107]

the interface roughness mean height has been considered. Results are shown for $\Delta = 0$ (no interface roughness scattering), 0.12 nm and 0.3 nm (corresponding to one monolayer of GaAs). Depending on the assumed value of Δ, interface roughness can lead to considerable broadening of the gain profile and reduction in the peak gain. The phonon depopulation structure in Fig. 15(a) is less sensitive to interface roughness than the bound-to-continuum structure in Fig. 15(b), since the wave functions in the minibands of the latter structure are less localized and thus extend across more interfaces.

3. Alloy scattering

In a ternary semiconductor alloy $A_x B_{1-x} C$ such as $Al_x Ga_{1-x} As$, the A and B atoms are randomly distributed on the corresponding lattice sites. Thus, for a zinc-blende structure, each unit cell contains an atom of type A or B and an atom of type C. The potential in the semiconductor is then given by[36,168]

$$V_\mathrm{c}(\mathbf{x}) = \sum_n V_n(\mathbf{x} - \mathbf{x}_n), \quad (111)$$

where $\mathbf{x}_n$ denotes the position of unit cell n. Furthermore, $V_n = V_{AC}$ ($V_n = V_{BC}$) if the unit cell contains an atom of type A (B), where V_{AC} and V_{BC} are the potentials of the binary materials AC and BC. The conduction band profile of alloys is typically considered in virtual crystal approximation, obtained by averaging over the contributions of the two alloy materials AC and BC,

$$V_{\mathrm{c,VCA}} = \sum_n \left[x V_{AC}(\mathbf{x} - \mathbf{x}_n) + (1-x)V_{BC}(\mathbf{x} - \mathbf{x}_n) \right]. \quad (112)$$

The perturbing potential corresponds then to the difference between the real conduction band energy $V_\mathrm{c}(\mathbf{x})$ and $V_{\mathrm{c,VCA}}$ obtained by the virtual crystal approximation

$$V = V_c - V_{c,\text{VCA}} = \sum_n c_n \delta V(\mathbf{x} - \mathbf{x}_n), \qquad (113)$$

with $\delta V(\mathbf{x}) = V_{\text{AC}}(\mathbf{x}) - V_{\text{BC}}(\mathbf{x})$ and $c_n = (1-x)$ $(c_n = -x)$ if the unit cell contains an atom of type A (B).

The scattering potential δV is effective within the unit cell, and also screening effects are usually neglected due to the short-range nature of the scattering potential.[36] Thus, on the scale of the envelope functions $\psi(\mathbf{x})$ which are slowly varying on the atomic scale, δV can be modeled as a δ function potential[36]

$$\delta V(\mathbf{x}) = \delta V \Omega_0 \delta(\mathbf{x}), \qquad (114)$$

where Ω_0 denotes the unit cell volume, given by $\Omega_0 = a^3/4$ for a zinc-blende structure with lattice constant a. Furthermore, δV is the alloy scattering potential, which approximately corresponds to the conduction band offset of the binary materials involved, if they are not strongly lattice-mismatched.[36] Generally, however, δV deviates from the conduction band differences,[156] and should be estimated from empirical tightbinding parameters.[169–172] Using Eqs. (113) and (114), the matrix element Eq. (55) becomes with $\mathbf{q} = \mathbf{k} - \mathbf{k}'$

$$V_{j\mathbf{k}',i\mathbf{k}} = S^{-1} \delta V \Omega_0 \sum_n c_n \psi_j^*(z_n) \psi_i(z_n) \exp(i\mathbf{q}\mathbf{r}_n), \qquad (115)$$

and the square modulus is given by

$$|V_{j\mathbf{k}',i\mathbf{k}}|^2 = S^{-2}(\delta V \Omega_0)^2 \sum_n \sum_m \{ c_n c_m \exp[i\mathbf{q}(\mathbf{r}_n - \mathbf{r}_m)]$$
$$\times \psi_j^*(z_n) \psi_j(z_m) \psi_i(z_n) \psi_i^*(z_m) \}. \qquad (116)$$

For calculating the average of $|V_{j\mathbf{k}',i\mathbf{k}}|^2$, the expectation value $\langle c_n c_m \rangle$ must be determined. Each unit cell has the probability x of containing an atom of type A, and $1-x$ of containing an atom of type B. Thus, we obtain $\langle c_n c_m \rangle = 0$ for $m \neq n$ and $\langle c_n c_m \rangle = x(1-x)$ for $m = n$. Furthermore replacing the sum over the unit cell positions $\sum_{\mathbf{x}_n} f(z_n)$ by an integral $\Omega_0^{-1} \int f(z) d^3 x = \Omega_0^{-1} S \int f(z) dz$, we obtain

$$\langle |V_{j\mathbf{k}',i\mathbf{k}}|^2 \rangle = S^{-1} \Omega_0 (\delta V)^2 x(1-x)$$
$$\times \int_{-\infty}^{\infty} |\psi_j(z) \psi_i(z)|^2 dz. \qquad (117)$$

From Eq. (57), we obtain the scattering rate

$$W_{i\mathbf{k},j\mathbf{k}'} = \frac{2\pi}{\hbar S} \delta(E_{j\mathbf{k}'} - E_{i\mathbf{k}})$$
$$\times \int_{-\infty}^{\infty} \Omega_0 (\delta V)^2 x(1-x) |\psi_j \psi_i|^2 dz, \qquad (118)$$

where $\Omega_0(z)$, $\delta V(z)$, and $x(z)$ have been taken into the integral to account for varying alloy compositions along the growth direction z. Due to the energy conservation $E_{i\mathbf{k}} = E_{j\mathbf{k}'}$, the final wave vector magnitude k' is given by Eq. (59). The total transition rate from a given initial state $|i\mathbf{k}\rangle$ to a subband j is with Eq. (61)

$$W_{i\mathbf{k},j} = m_j^{\|} \hbar^{-3} \int_{-\infty}^{\infty} \Omega_0 (\delta V)^2 x(1-x) |\psi_j \psi_i|^2 dz \qquad (119)$$

for $E_j < E_{i\mathbf{k}}$, and 0 otherwise.

E. Self-consistent rate equation approach

In the self-consistent rate equation approach, the inter-subband scattering rates in the rate equation Eq. (36) or Eq. (37) are self-consistently determined based on the corresponding Hamiltonian. Thus, this approach only relies on well known material parameters such as the effective mass, and not on experimental or empirical lifetimes as in Sec. V A. Offering a compromise between accuracy and predictive power on the one hand and relative numerical efficiency on the other hand, the self-consistent rate equation approach is widely used for the simulation of QCLs.[67,77,91–93,173] Although extensions of the self-consistent rate equation approach to include the light field have been presented,[77] the optical cavity field is neglected in most cases. Such simulations yield the unsaturated population inversion or gain, indicating if lasing can start at all, but giving no information about the actual lasing operation.

1. Intersubband scattering rates

The number of electrons per unit energy and area in a thermalized subband i of a 2D system is given by $n_i^E f_i^{\text{FD}}$ with the Fermi-Dirac distribution

$$f_i^{\text{FD}}(E_{i\mathbf{k}}) = \left\{ \exp\left[(E_{i\mathbf{k}} - E_i^{\text{F}})/(k_B T_i) \right] + 1 \right\}^{-1}, \qquad (120)$$

where E_i^{F} is here a "quasi" Fermi energy describing the kinetic energy distribution of the electrons within the subband i,[132] and T_i is the associated electron temperature. Furthermore, n_i^E is the density of states per unit area and energy in a 2D system,[61] and $E_{i\mathbf{k}} = E_i + \hbar^2 k^2/(2m_i^{\|})$ is the energy of the electrons in subband i with an in-plane wave vector $\mathbf{k}$. The scattering rate from an initial subband i to a final subband j is then obtained by averaging over the carrier distribution. Assuming a Fermi-Dirac distribution in the initial and final state, we obtain[132]

$$(\tau_{ij}^{(m)})^{-1} = \frac{\int_{E_i}^{\infty} f_i^{\text{FD}}(E) W_{i\mathbf{k},j}^{(m)} dE}{\int_{E_i}^{\infty} f_i^{\text{FD}}(E) dE}. \qquad (121)$$

Here, n_i^E has been omitted in the enumerator and denominator because it is constant, $n_i^E = m_i^{\|}/(\pi \hbar^2)$ for $E \geq E_i$. The index m denotes the corresponding scattering mechanism. $W_{i\mathbf{k},j}^{(m)}$ is, for example, given by Eqs. (82) and (88) for LO phonon and electron-electron scattering, respectively. The total rate is obtained by summing over the individual contributions of the different scattering mechanisms, i.e., $\tau_{ij}^{-1} = \sum_m (\tau_{ij}^{(m)})^{-1}$.

The sheet density in a subband i, i.e., electron number per unit area, is given by

$$n_i^s = \int_{E_i}^{\infty} f_i^{FD}(E) n_i^E dE$$

$$= \frac{m_i^{\parallel} k_B T_i}{\pi \hbar^2} \left\{ \frac{E_i^F - E_i}{k_B T_i} + \ln\left[1 + \exp\left(-\frac{E_i^F - E_i}{k_B T_i} \right) \right] \right\}.$$

$$(122)$$

Furthermore expressing the carrier energy in terms of the in-plane wave vector, we obtain with $dE = \hbar^2 (m_i^{\parallel})^{-1} k dk$[132]

$$(\tau_{ij}^{(m)})^{-1} = \frac{\int_0^{\infty} f_i^{FD}\left[E_i + \hbar^2 k^2 / \left(2m_i^{\parallel} \right) \right] W_{ik,j}^{(m)} k dk}{\pi n_i^s}. \quad (123)$$

An additional factor $\{ 1 - f_j^{FD}[E_i + \hbar^2 k^2/(2m_i^{\parallel}) \pm E_0] \}$ can be included in the integral of Eq. (123), accounting for the final state blocking due to Pauli's exclusion principle.[132] This correction is only relevant for high doping levels. Here, E_0 is 0 for elastic scattering mechanisms and corresponds to the TO or LO phonon energy for optical phonon absorption ("+" sign) and emission ("–" sign), respectively. The Fermi energy can be calculated from Eq. (122)

$$E_i^F - E_i = k_B T_i \ln\left[\exp\left(\frac{n_i^s \pi \hbar^2}{m_i^{\parallel} k_B T_i} \right) - 1 \right]. \quad (124)$$

For lightly doped semiconductors, as is often the case in QCLs, we have $\left(E_{ik} - E_i^F \right) \gg k_B T_i$ in Eq. (120), which then approaches a classical Maxwell-Boltzmann distribution

$$f_i^{MB}(E_{ik}) = \exp\left[-\left(E_{ik} - E_i^F \right) / (k_B T_i) \right]. \quad (125)$$

Under this condition, Eq. (123) simplifies to

$$\left(\tau_{ij}^{(m)} \right)^{-1} = \frac{\hbar^2}{m_i^{\parallel} k_B T_i} \int_0^{\infty} \exp\left(-\frac{\hbar^2 k^2}{2m_i^{\parallel} k_B T_i} \right) W_{ik,j}^{(m)} k dk. \quad (126)$$

Numerically, the integration of Eq. (123) or Eq. (126) is performed from E_i up to a sufficiently large maximum value, e.g., the highest value of the simulated potential profile $V(z)$.[132]

2. Rate equations

In the self-consistent rate equation approach, typically the QCL is modeled as a biased periodic heterostructure, excluding effects such as domain formation.[127] Then the simulation can be restricted to a single representative period far away from the contacts, additionally applying periodic boundary conditions.[92] This corresponds to solving Eq. (36) with the self-consistently calculated scattering rates, Eq. (123). While Eq. (36) in principle includes the transitions to all equivalent levels in the different periods, in practice only scattering between the central period and adjacent periods has to be considered. Frequently the $1\frac{1}{2}$ period model is used, which applies to QCLs where a period consists of an injector region and an active region.[91,92] Here, for the active region states of the central period, only scattering transitions involving other states of the central period and the right-neighboring injector region are considered in Eq. (36). Analogously, for the states of the injector region, only scattering involving states of the central period and the left-neighboring active region are taken into account. Since the scattering rates Eq. (123) in general depend on the electron densities n_i^s, a direct solution of Eq. (36) is not possible, and an iterative scheme is commonly used. For simulations without lasing included, i.e., $W_{ij}^{opt} = 0$ in Eq. (36), setting $d_t n_i^s = 0$ yields the steady state solution[92]

$$n_i^s = \sum_{j \neq i} n_j^s \tau_{ji}^{-1} \bigg/ \sum_{j \neq i} \tau_{ij}^{-1}, \quad (127)$$

where $i = 1..N$ refers to the central period containing N subbands. The summation index j now only includes subbands in the central period and adjacent periods, as discussed above. For subbands outside the central period, the sheet density n_j^s of the equivalent level in the central subband is used. For the numerical computation of the subband populations, initially identical electron densities $n_i^s = n^s/N$ are assumed. Each iteration involves computing the scattering rates with Eq. (123) or Eq. (126), calculating the new values for n_i^s using Eq. (127), and renormalizing the sheet densities so that Eq. (38) is fulfilled. Convergence can be accelerated by combining the sheet densities of the previous two iterations, $\xi n_i^{s,new} + (1 - \xi) n_i^{s,old}$, as input for the next iteration,[91] where the relaxation parameter is typically chosen as $\xi = 0.5$. The simulation has converged when the obtained sheet densities do not significantly change anymore between iterations.

3. Kinetic energy balance method

Since T_i in Eq. (120) is unknown, the scattering rates [Eq. (123)] have to be evaluated assuming that the electron temperature in each subband is equivalent to the lattice temperature, $T_i = T_L$. However, the electron temperature can significantly exceed T_L in quantum cascade lasers.[174] Thus, the rate equation model has been extended to account for electron heating, typically assuming an identical electron temperature $T_i = T_e$ for all subbands.[173] The kinetic electron energy generation rate per period and unit device in-plane cross section is given by[173]

$$R^E = \sum_{i=1}^{N} \sum_j \sum_m n_i^s (\tau_{ij}^{(m)})^{-1} (E_i - E_j + E_0^{(m)}), \quad (128)$$

where i sums over the subbands of the central period and j also includes states from adjacent periods, as discussed above. Furthermore, m sums over the different scattering contributions, such as LO phonon emission ($E_0^{(m)} = -E_{LO}$) and absorption ($E_0^{(m)} = E_{LO}$), elastic scattering mechanisms ($E_0^{(m)} = 0$) and electron-electron scattering. For inelastic scattering mechanisms such as phonon-induced transitions, also intrasubband contributions $i = j$, have to be included.[173] Electron-electron scattering is often neglected in Eq. (128) since the net kinetic energy does not change for the important cases where both electrons stay within their respective subbands or swap the subbands. Also photon transitions do not change the kinetic energy for parabolic subbands due to

k conservation.[77] The numerical evaluation proceeds as follows. The average electron temperature corresponds to the value of T_e where $R^E = 0$ in Eq. (128). The rate equation is self-consistently solved as described above for an initial guess of T_e, e.g., $T_e = T_L$. Based on the obtained n_i^s and $\tau_{ij}^{(m)}$, R^E in Eq. (128) is calculated. This procedure is repeated and the guess for T_e is iteratively improved until $R_E \approx 0$ is obtained. Since the assumption of a single effective electron temperature is not always adequate in QCLs,[175] extended approaches have been developed allowing for different effective temperatures T_i in the individual subbands.[93]

F. Ensemble Monte Carlo method

Semiclassically, the carrier transport between states $|i\mathbf{k}\rangle$ in a quantum well system is given by the Boltzmann equation[176]

$$d_t f_{i\mathbf{k}} = \sum_j \sum_{\mathbf{k}'} \left(W_{j\mathbf{k}',i\mathbf{k}} f_{j\mathbf{k}'} - W_{i\mathbf{k},j\mathbf{k}'} f_{i\mathbf{k}} \right), \qquad (129)$$

with the scattering rates $W_{i\mathbf{k},j\mathbf{k}'} = \sum_m W_{i\mathbf{k},j\mathbf{k}'}^{(m)}$, where m sums over the individual contributions of the different scattering mechanisms. The distribution function $f_{i\mathbf{k}}(t)$ represents the probability of the state $|i\mathbf{k}\rangle$ being occupied at a given time t. Eq. (129) corresponds to an extended version of the rate equations Eq. (36), where i and j have been replaced by $i\mathbf{k}$ and $j\mathbf{k}'$, respectively. The physical quantities of interest, such as the sheet densities n_i^s and current density can be extracted from $f_{i\mathbf{k}}(t)$. For the numerical evaluation of Eq. (129), typically the EMC method is used, which is based on statistical sampling of the scattering events for a large ensemble of carriers,[149] here $N_e \approx 10^4..10^5$ electrons. The large number of carriers considered allows for an extraction of the physical quantities as function of t by statistical averaging over the carrier ensemble. Thus, this method is also applicable to time dependent processes where we cannot use temporal averaging as for ergodic systems.[149] Furthermore, electron-electron scattering can be implemented as a two-electron process where a second electron is randomly chosen as scattering partner.[141]

1. Simulation technique

Figure 16 contains a schematic diagram of the EMC algorithm. The system dynamics is evaluated up to a time t_{sim}, which must be chosen long enough to ensure convergence to the stationary solution. The simulation is divided into subintervals Δ_t, where the scattering dynamics is subsequently evaluated. Δ_t should be chosen so small that the average electron distribution does not change significantly over the time interval, but big enough to include several scattering events per electron on average. Assuming periodic boundary conditions, we can restrict our simulation to a few periods. Each electron is characterized by its subband i and in-plane wave vector **k**. All rates $W_{i\mathbf{k},j}^{(m)}$ for the various scattering mechanisms m are computed and tabulated at the beginning of the simulation to save computational resources. Here, we have to introduce a discrete grid for the wave vector. Since in the conduction band Γ valley, all states $|i\mathbf{k}\rangle$ in subband i
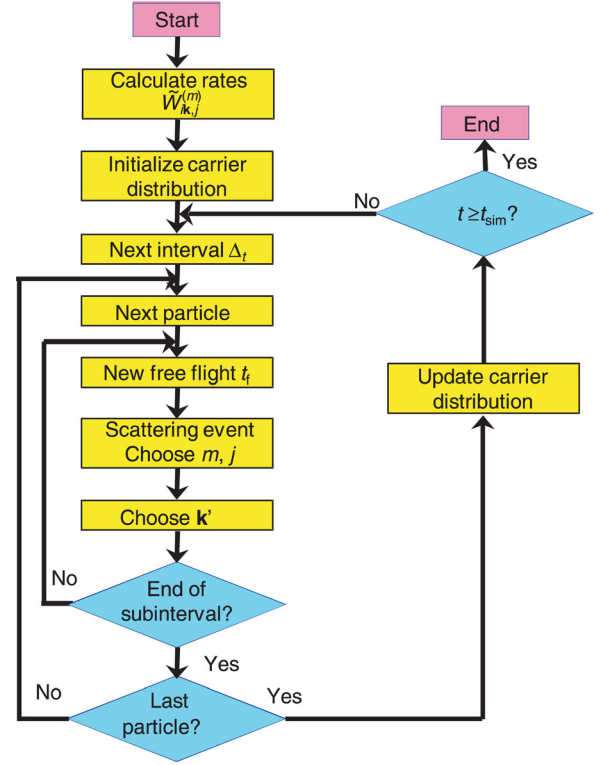

FIG. 16. Schematic diagram of the EMC algorithm

with the same value $k = |\mathbf{k}|$ are equivalent due to in-plane isotropy, it is practical to use the kinetic energy $E_{kin} = \hbar^2\mathbf{k}^2/(2m_i^\parallel)$ instead. The kinetic energy grid then divides the energy axis into segments n of widths $\Delta_E^{(n)}$ centered around discrete energies $E_{kin}^{(n)}$ with $k^{(n)} = (2m_i^\parallel E_{kin}^{(n)})^{1/2}/\hbar$. However, some rates $W_{i\mathbf{k},j}^{(m)}$, such as Eq. (88), depend on the initially unknown carrier distribution itself, i.e., $W_{i\mathbf{k},j}^{(m)} = W_{i\mathbf{k},j}^{(m)}(t)$. This problem can be overcome by tabulating an upper estimate $\tilde{W}_{i\mathbf{k},j}^{(m)}$ for time dependent scattering rates and compensating for the too high value by introducing artificial "self-scattering," as described further below.

An important quantity is the carrier distribution function, which can directly be obtained from $N_i^{(n)}$, denoting the number of simulated electrons in the nth energy cell of subband i at a time t. With the density of states per unit area and energy in a 2D system $n_i^E = m_i^\parallel/(\pi\hbar^2)$,[61] the number of available states in the nth energy cell is $n_i^E \Delta_E^{(n)} S$. The simulated device in-plane cross section is with the sheet doping density per period n^s given by $S = N_e/(n^s N_p)$, where N_p corresponds to the number of periods over which the N_e simulated electrons are distributed. The carrier distribution function is then approximately given as

$$f_i(E_{kin}, t) = \frac{n^s N_p N_i^{(n)}(t)}{n_i^E \Delta_E^{(n)} N_e}, \qquad (130)$$

where n indicates the energy cell containing the value E_{kin}. Equation (130) can also be expressed as a function of $k = (2m_i^\parallel E_{kin})^{1/2}/\hbar$. Additional temporal averaging of the carrier distribution function Eq. (130), for example over a simulation subinterval Δ_t, further reduces the fluctuations

resulting from the treatment of the carriers as discrete particles.

The simulation is initialized by assigning values i and $\mathbf{k}$ to each electron in the ensemble. Here, in principle arbitrarily chosen values can be used since Eq. (129) will converge to its stationary solution after a sufficiently long simulation time t_{sim} independently of the chosen initial conditions. However, the required t_{sim} to obtain convergence can be reduced by assuming a suitable initial electron distribution which is not too far from the converged solution. For example, the carriers can be distributed equally within the subbands, and a thermalized distribution Eq. (125) in each subband can be assumed. The Monte Carlo method is based on a stochastic evaluation of the scattering events, assuming that the electron undergoes scattering after a randomly selected free flight time t_{f}. For a time independent scattering rate τ_0^{-1}, the free flight time is given by $t_{\mathrm{f}} = -\tau_0 \ln(r)$, where r is a random number evenly distributed between 0 and 1.[149] Here, we choose $\tau_0^{-1} \geq \sum_m \sum_j \tilde{W}_{i\mathbf{k},j}^{(m)}$, corresponding to an upper bound for the total outscattering rate $W_{i\mathbf{k}}(t) = \sum_m \sum_j W_{i\mathbf{k},j}^{(m)}(t)$, where $j=i$ is included in the summations to account for intrasubband scattering. For each ensemble electron, the scattering dynamics is evaluated over a simulation subinterval Δ_t, where the last free flight is continued in the next subinterval. After each scattering event, the final subband j of the scattered electron and scattering mechanism m are randomly selected based on their associated probability $\tau_0 W_{i\mathbf{k},j}^{(m)}(t)$. The too high amount of scattering assumed is corrected by introducing artificial "self-scattering," which occurs with a probability $1 - \tau_0 \sum_m \sum_j W_{i\mathbf{k},j}^{(m)}(t)$ and does not change the carrier state at all.[149] Pauli's exclusion principle can be considered by subsequently rejecting the occurred scattering event with a probability $f_{j\mathbf{k}'}$. The final wave vector magnitude k' is obtained from Eq. (59) for elastic scattering, and from Eq. (60) for inelastic processes. Electron-electron scattering is in EMC typically implemented as a two-electron process with a randomly selected partner electron and scattering angle. To conserve energy for each scattering event individually, also the partner electron has to undergo scattering, i.e., the final wave vector has to be determined for both participating electrons.[141]

The periodic boundary conditions can, for example, be implemented by simulating three periods, where the N_e electrons are located in the central period, i.e., $N_p = 1$ in Eq. (130). Electrons scattered to the first or third period are automatically injected into the equivalent subband of the central period.[99] By counting the difference Δ_N of electrons scattered from the central period to the left- and right-neighbouring period, respectively, over a simulation subinterval Δ_t, the current density can be computed as

$$J = \frac{\Delta_N e}{\Delta_t S} = \frac{\Delta_N e n^{\mathrm{s}} N_{\mathrm{p}}}{\Delta_t N_e}. \qquad (131)$$

G. Inclusion of the optical cavity field

The optical gain and photon-induced transition rate are given by Eqs. (52) and (54) for a transition from a level i to a

level j, which here correspond to states $|i\mathbf{k}\rangle$ and $|j\mathbf{k}'\rangle$, respectively. The dipole matrix element then becomes

$$d_{i\mathbf{k},j\mathbf{k}'} = -e\langle i\mathbf{k}|z|j\mathbf{k}'\rangle = d_{ij}S^{-1} \int_S \exp[i(\mathbf{k}' - \mathbf{k})\mathbf{r}]\mathrm{d}^2r, \quad (132)$$

with

$$d_{ij} = -e\langle i|z|j\rangle = -e \int_{-\infty}^{\infty} \psi_i^*(z)z\psi_j(z)\mathrm{d}z, \qquad (133)$$

and for $|d_{i\mathbf{k},j\mathbf{k}'}|^2$, we obtain

$$|d_{i\mathbf{k},j\mathbf{k}'}|^2 = 4\pi^2 |d_{ij}|^2 S^{-1} \delta(\mathbf{k}' - \mathbf{k}). \qquad (134)$$

Here, we have used that $|\int_S \exp[i(\mathbf{k}' - \mathbf{k})\mathbf{r}]\mathrm{d}^2r|^2$ can be approximated by $4\pi^2 S\delta(\mathbf{k}' - \mathbf{k})$ for sufficiently large in-plane cross sections S. The photon-induced transition rate given in Eq. (54) now becomes

$$W_{i\mathbf{k},j\mathbf{k}'}^{\mathrm{opt}} = \frac{4\pi^3}{\epsilon_0 c n_0 \hbar^2 S} |d_{ij}|^2 \delta(\mathbf{k}' - \mathbf{k}) \sum_m I_m \mathcal{L}_{ij}(\omega_m, k), \quad (135)$$

where we sum over all relevant cavity modes with frequencies ω_m and intensities I_m to account for multimode lasing. Adapting Eq. (51) to the present case, the definition of the Lorentzian lineshape function becomes

$$\mathcal{L}_{ij}(\omega, k) = \frac{1}{\pi} \frac{\gamma_{ij}(k)}{\gamma_{ij}^2(k) + [\omega - |\omega_{ij}(k)|]^2}, \qquad (136)$$

where $\omega_{ij}(k) = (E_{i\mathbf{k}} - E_{j\mathbf{k}})/\hbar$ denotes the resonance frequency and $\gamma_{ij}(k)$ is the optical linewidth of the transition, which is given by

$$\gamma_{ij}(k) = \frac{1}{2}\left(\sum_{\ell \neq i} W_{i\mathbf{k},\ell} + \sum_{\ell \neq j} W_{j\mathbf{k},\ell}\right), \qquad (137)$$

when only the lifetime broadening contributions are considered.[107] The total photon-induced transition rate from a given initial state $|i\mathbf{k}\rangle$ to a subband j is found from Eq. (135) by summation over all final wave vectors $\mathbf{k}'$ using Eq. (61)

$$W_{i\mathbf{k},j}^{\mathrm{opt}} = \frac{\pi}{\epsilon_0 c n_0 \hbar^2} |d_{ij}|^2 \sum_m I_m \mathcal{L}_{ij}(\omega_m, k). \qquad (138)$$

The transition rate due to spontaneous photon emission can also be directly calculated from Eq. (138),[177] but is usually negligible compared to other scattering mechanisms in QCLs. The gain contribution at frequency ω of a single electron in state $|i\mathbf{k}\rangle$, i.e., $n_i^{\mathrm{s}} = 1/S$, is obtained from Eq. (52) by summing over the transitions to all available final states $|j\mathbf{k}'\rangle$. With Eqs. (61) and (134), we obtain

$$g(\omega) = \frac{\pi\omega}{\hbar\epsilon_0 c n_0 S L_{\mathrm{p}}} \sum_j \frac{\omega_{ij}(k)}{|\omega_{ij}(k)|} |d_{ij}|^2 \mathcal{L}_{ij}(\omega, k). \qquad (139)$$

For $E_{i\mathbf{k}} < E_{j\mathbf{k}}$, we have $\omega_{ij}(k)/|\omega_{ij}(k)| = -1$, indicating absorption and thus resulting in a negative contribution to

the optical gain.[76] For EMC simulations, the gain can be evaluated by summing Eq. (139) over all simulated electrons n in the corresponding states $|i_n\mathbf{k}\rangle$ of the central simulation period(s)[76]

$$g(\omega) = \frac{\pi\omega}{\hbar\epsilon_0 c n_0 S L_p} \sum_n \sum_j \frac{\omega_{i_n j}(k)}{|\omega_{i_n j}(k)|} |d_{i_n j}|^2 \mathcal{L}_{i_n j}(\omega, k), \quad (140)$$

where j also includes states in adjacent periods. S then corresponds to the simulated device in-plane cross section, which is for a sheet doping density per period n^s given by $S = N_e/(n^s N_p)$, where N_P corresponds to the number of periods over which the N_e simulated electrons are distributed.

For rate equations, the electron wave vector dependence is not considered, and thus a k independent averaged value γ_{ij} has to be taken in Eq. (136). Furthermore, ω_{ij} does not depend on the wave vector if nonparabolicity effects are neglected. Then the gain simplifies to Eq. (52), thus becoming

$$g = \frac{\pi\omega}{\hbar\epsilon_0 c n_0 L_p} \sum_{i,j} |d_{ij}|^2 (n_i^s - n_j^s) \mathcal{L}_{ij}(\omega). \quad (141)$$

The intensity evolution for a mode m is again given by Eq. (49)

$$n_0 c^{-1} \partial_t I_m = \Gamma_m g(\omega_m) I_m - a_m I_m. \quad (142)$$

The carrier transport and intensity evolution have to be simulated by a coupled approach, as illustrated in Fig. 17. The carrier transport simulation based on rate equations or EMC now includes the photon-induced transition rates Eq. (138), which depend on the intensities given by Eq. (142). On the other hand, Eq. (142) depends on the gain which is extracted from the carrier transport simulations by using Eq. (140) or Eq. (141).
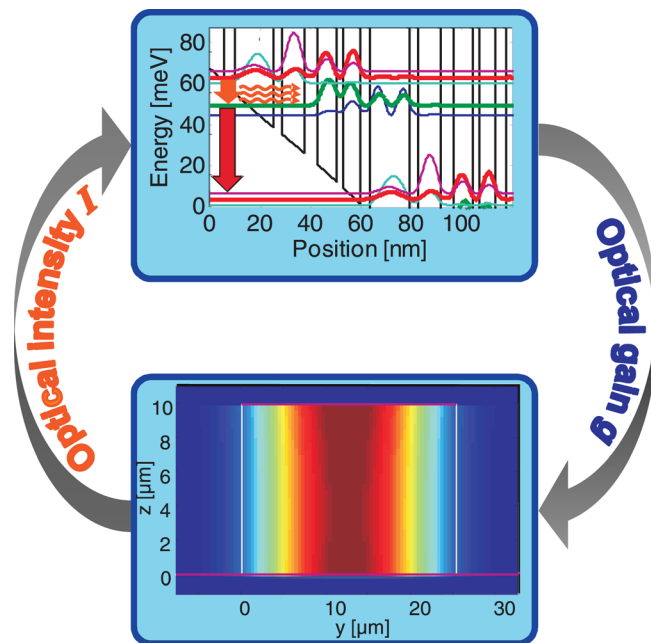

FIG. 17. Illustration of coupled simulations including both the carrier transport and optical cavity field.
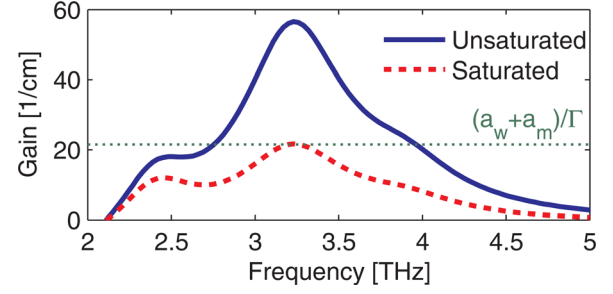

FIG. 18. Simulated unsaturated and saturated power gain coefficient vs. frequency. The dashed line indicates the threshold gain. Reprinted with permission from Appl. Phys. Lett. **96**, 011103 (2010). Copyright 2010 AIP Publishing LLC.[76]

H. Selected EMC simulation results

In the following, some EMC simulation results are presented for both mid-infrared and terahertz QCLs, illustrating the versatility of this approach. Examples without inclusion of the optical cavity field are shown in Secs. VI A–VI D, see Figs. 11, 13, and 15 where results for the unsaturated optical gain are displayed. Here, we focus on coupled simulations of the carrier transport and the optical cavity field.

The EMC method has been applied to study the optical power and gain saturation in terahertz QCLs, yielding good agreement with experimental data.[76] Here, examples are presented for a high temperature QCL lasing up to 164 K.[178] In Fig. 18, the simulated unsaturated spectral gain curve is compared to the result obtained for carrier-light coupling included, demonstrating gain saturation and clamping at the threshold gain value due to the lasing field. In Fig. 19, the obtained optical output power as a function of lattice temperature is compared to experimental data, demonstrating good agreement with experiment.

Due to the stimulated optical transitions, the lasing field significantly influences the carrier transport, affecting not only the subband populations and thus the optical gain but also the electric current. This is especially the case for high efficiency mid-infrared QCLs. In Fig. 20, the measured current density of a mid-infrared QCL with $\approx$50% wall-plug efficiency[9] is compared to EMC results with and without
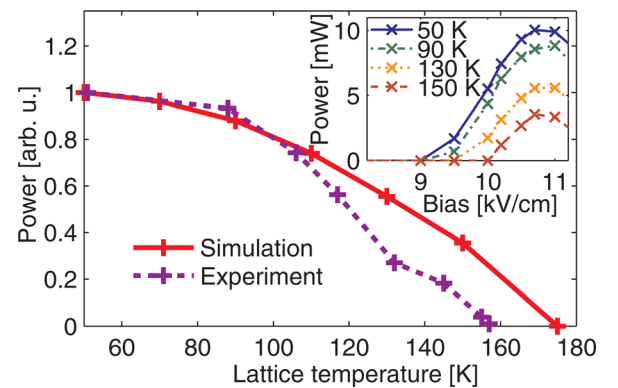

FIG. 19. Comparison of the measured optical power vs. lattice temperature[178] to EMC simulation results; the inset contains the simulated power vs. applied bias for various lattice temperatures. Reprinted with permission from Appl. Phys. Lett. **96**, 011103 (2010). Copyright 2010 AIP Publishing LLC.[76]
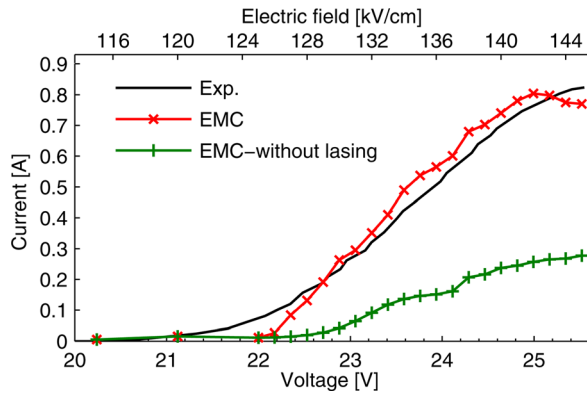
FIG. 20. The measured current-voltage characteristics for a high-efficiency mid-infrared QCL[9] is compared to results obtained from EMC simulations with and without lasing included. Reprinted with permission from J. Appl. Phys. **110**, 013108 (2011). Copyright 2011 AIP Publishing LLC.[106]

carrier-light coupling included.[106] Good agreement with experiment is only obtained in the former case, while the simulation neglecting photon-induced processes significantly underestimates the current density, demonstrating the significant influence of these processes on the carrier transport.

VII. DENSITY MATRIX APPROACHES

The one-dimensional density matrix approach is frequently used for the analysis and optimization of QCLs.[17,111–113] Also three-dimensional versions have been developed.[114,115] One-dimensional density matrix approaches can be seen as a quantum mechanical generalization of rate equations, Eq. (36), to include effects such as resonant tunneling and dephasing. In the semiclassical description, transport through a barrier occurs instantaneously due to electrons

being scattered into wave functions which are spatially extended across the barrier,[125] see Fig. 21(a). In the quantum mechanical picture, the electron transport across the barrier is described by a coherent superposition of the extended states with a narrow anticrossing energy gap Δ_E, resulting in a localized electron wavepacket. Due to the coherent time evolution of these states, the wavepacket oscillates between the left and right well with the so-called Rabi oscillation frequency $\Omega = \Delta_E/\hbar$,[125] and the corresponding tunneling time corresponds to half the duration of an oscillation cycle, $\tau_{\text{tun}} = \pi/\Omega$.[111] The semiclassical picture is adequate as long as the Rabi oscillations are not significantly dampened by dephasing, since then the wavepacket oscillates uniformly between the wells and the averaged population distribution corresponds to the semiclassical description. However, for strong dephasing with relaxation times $\tau \leq \tau_{\text{tun}}$, the wave packet no longer oscillates. Rather, the electrons accumulate in the left well, and the semiclassical picture of a uniform electron distribution across the extended states is no longer valid.[125] Now the current transport is not limited by the scattering-induced electron transport, but by tunneling through the barrier. This effect is commonly referred to as wave function localization due to dephasing, and occurs for thick barriers with narrow anticrossing gaps Δ_E and thus long tunneling times τ_{tun}.[125] Especially in terahertz QCLs, thick injection barriers are used with $\Delta_E \approx 1$ meV,[125] corresponding to τ_{tun} of around 2 ps, while the dephasing time can be estimated from the spontaneous emission linewidth of ~ 5 meV (Ref. 125) to be ≈ 0.3 ps.

Rather than describing localized wavepackets by a coherent superposition of eigenstates, frequently localized wave functions are used,[111–113] see Fig. 21(b). Here, the wave functions are computed for each QCL period separately, by
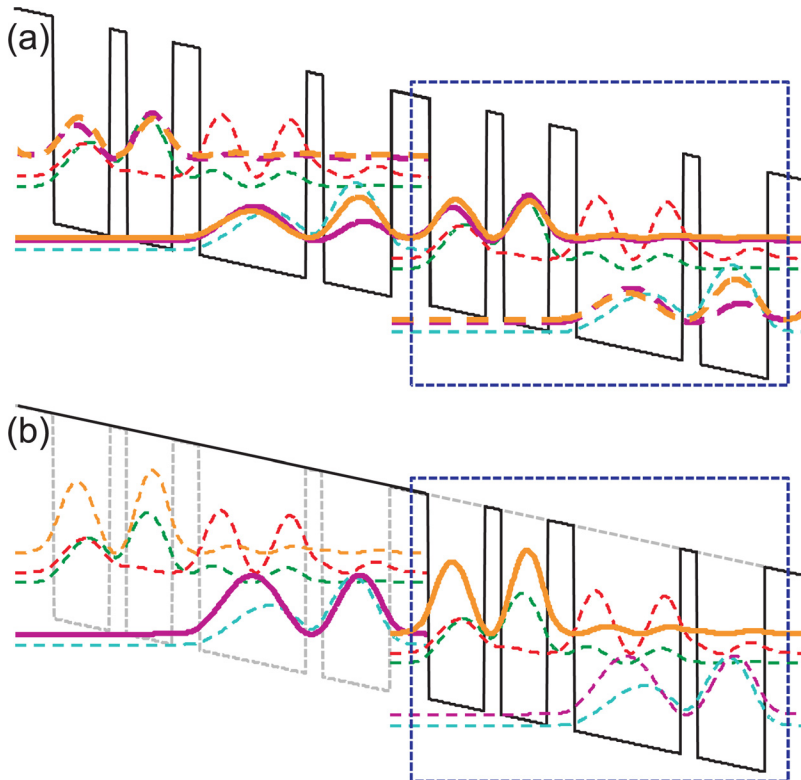

FIG. 21. Conduction band profile and probability densities for a terahertz QCL, computed based on (a) the actual potential V modeled as a periodic sequence of stages and (b) the tight-binding conduction band profile V_{tb} obtained by extending the barriers at the stage boundaries to confine the wavefunctions within the stage. The rectangles denote a single stage. The extended wavefunctions spanning the thick barrier and the corresponding localized wavefunctions are marked by bold lines.

assuming that the thick injection barriers at the left and right sides of the period are infinitely thick.[111] The Schrödinger equation, Eq. (4), is then solved for the corresponding tight-binding conduction band profile V_{tb} rather than the actual potential V. The corresponding Hamiltonian is obtained in the framework of tight-binding theory.[36] Specifically, the Rabi frequency for a doublet of states spanning the coupling barrier is given by $\Omega_{ij} = \hbar^{-1}\langle i|V - V_{tb}|j\rangle$.[111] Depending on the investigated QCL structure, resonant tunneling may have to be considered at more than a single coupling barrier in each period.[112,113] Then the period is subdivided into smaller regions, and the tight binding formalism has to be applied accordingly.

Wave functions ψ_i represent pure states, and thus are not adequate for describing decoherence effects due to the interaction with the environment. This can be accomplished by means of a density matrix ρ_{ij} (see also Sec. V B), which can represent both pure and mixed states and thus is suitable to model dephasing. The diagonal terms ρ_{ii} correspond to the occupation of level i, and the off-diagonal elements ρ_{ij} are the coherence or polarization terms for the doublet i, j.[125] The density matrix can be normalized so that the diagonal terms correspond to the sheet density, $\rho_{ii} = n_i^s$. The time evolution of the density matrix $\rho_{ij}(t)$ is described by the von Neumann equation

$$i\hbar d_t\rho_{ij} = \sum_\ell (H_{i\ell}\rho_{\ell j} - \rho_{i\ell}H_{\ell j}), \qquad (143)$$

where the Hamiltonian matrix elements are defined as $H_{ij} = \langle i|\hat{H}|j\rangle$. For describing the carrier transport in QCLs, the Hamiltonian is divided into two parts, $\hat{H} = \hat{H}_0 + \hat{H}'$. Here, $\hat{H}_0$ describes the coherent evolution of the quantum system due to the conduction band potential, while $\hat{H}'$ contains the perturbation potentials corresponding to the various scattering mechanisms and introduces dissipation to the system. $\hat{H}'$ is commonly implemented in a somewhat phenomenological manner by using transition and dephasing rates, similar as for the rate equation approach discussed in Secs. V A and VI E. Equation (143) can then be cast into the form[111,113]

$$d_t n_i^s = \sum_{j\neq i}\tau_{ji}^{-1}n_j^s - \tau_i^{-1}n_i^s + \sum_j i\Omega_{ij}(\rho_{ij} - \rho_{ji}),$$
$$d_t\rho_{ij} = i\Omega_{ij}(n_i^s - n_j^s) - i\rho_{ij}\omega_{ij} - \gamma_{ij}\rho_{ij}, \qquad (144)$$

where $\omega_{ij} = (E_i - E_j)/\hbar$, and $\tau_i^{-1} = \sum_{j\neq i}\tau_{ji}^{-1}$ indicates the total inverse lifetime of level i. The scattering rates τ_{ij}^{-1} have already been discussed in Secs. V A and VI E. The dephasing rate is given by[125]

$$\gamma_{ij} = (\tau_i^{-1} + \tau_j^{-1})/2 + \tau_{\text{pure},ij}^{-1}. \qquad (145)$$

Here, $(\tau_i^{-1} + \tau_j^{-1})/2$ is the lifetime broadening contribution [see also Eq. (137)], and $\tau_{\text{pure},ij}^{-1}$ contains the pure dephasing. The Rabi frequency Ω_{ij} is nonzero only for doublets spanning a coupling barrier; specifically, $\Omega_{ii} = 0$. In Eq. (144), resonant tunneling is assumed to be independent of the in-plane wave vector, and the rates $\tau_{ij}^{-1}, \gamma_{ij}$ are averaged over

the kinetic electron distribution within the subbands.[111] The stationary solution of Eq. (144) is obtained by setting $d_t = 0$, yielding[113]

$$\sum_{j\neq i}[\tau_{ji}^{-1}n_j^s + R_{ij}(n_j^s - n_i^s)] - \tau_i^{-1}n_i^s = 0, \qquad (146)$$

with

$$R_{ij} = \frac{2\Omega_{ij}^2\gamma_{ij}^{-1}}{1 + \omega_{ij}^2\gamma_{ij}^{-2}}. \qquad (147)$$

Furthermore, Eq. (38) has to be fulfilled, i.e., the total sheet density in each period is determined by the doping sheet density n^s.

Equation (144) can be solved using empirical rates, or implemented in a self-consistent manner by calculating the scattering rates τ_{ij}^{-1} as discussed in Sec. VI E. The pure dephasing rate $\tau_{\text{pure},ij}^{-1}$ in Eq. (145) can be computed based on intrasubband scattering transitions,[113,179] but is often treated by assuming an empirical value.[112,125] In a simplified model, the transport across the barrier can be restricted to the coupling between two states 1 and 2. For this case, Eq. (146) simplifies with $\tau_{12}^{-1} = \tau_{21}^{-1} = 0$ to $\tau_2^{-1}n_2^s + R_{12}(n_2^s - n_1^s) = 0$. The current density through the barrier is then given by $J = eR_{12}(n_1^s - n_2^s) = e\tau_2^{-1}n_2^s$. With Eq. (147) and $n_1^s + n_2^s = n^s$, we obtain[10]

$$J = \frac{2en^s\Omega_{12}^2\gamma_{12}^{-1}}{1 + \omega_{12}^2\gamma_{12}^{-2} + 4\Omega_{12}^2\gamma_{12}^{-1}\tau_2}. \qquad (148)$$

In the three-dimensional density matrix method, additionally the in-plane wave vector $\mathbf{k}$ is taken into account, i.e., the electrons are described by states $|i\mathbf{k}\rangle$ rather than $|i\rangle$. Consequently, the density matrix is given by $\rho_{i\mathbf{k},j\mathbf{k}'}$, and i, j, ℓ in Eq. (143) have to be replaced by $i\mathbf{k}$, $j\mathbf{k}'$, and $\ell\mathbf{k}''$, respectively. Three-dimensional density matrix approaches can be seen as a quantum mechanical generalization of the Boltzmann equation given in Eq. (129). Various approaches based on the three-dimensional density matrix have been developed for QCL simulation, where the scattering mechanisms are self-consistently implemented based on the corresponding Hamiltonians.[114,115] Specifically, also a hybrid density matrix-Monte Carlo approach has been introduced, where the tunneling transport through the coupling barrier is treated based on the density matrix formalism, while scattering inside each period is treated semiclassically using an EMC approach.[125]

VIII. NON-EQUILIBRIUM GREEN'S FUNCTION FORMALISM

A. General scope of the non-equilibrium Green's function method

Quantum cascade devices utilize charge transport in structures of the nanometer length scale. In this regime, quantum effects such as coherent tunneling, interference,

and confinement play a very important role for the transport physics. At the same time, however, the devices are run at finite temperatures, which support a significant amount of scattering with phonons. Since most devices are doped, alloyed and/or based on heterojunctions of different materials, impurity and alloy disorder as well as surface and interface roughness influence the transport, too. All these scattering effects share the fact that the full quantum information (i.e., the full phonon phase, the precise position of the impurities, etc.) is either unknown or lost in the statistics of large numbers of scattering events. Consequently, these effects contribute to incoherent scattering and dephasing.

It is well established that the NEGF theory is the most general scheme for the prediction of coherent and incoherent quantum transport. Since its introduction in the 1960s,[180–182] this formalism has been successfully applied on a great variety of transport problems. These problems include, but are not limited to spin,[183–185] phonon,[186–188] and electron transport,[189,190] covering different materials such as metals,[191,192] semiconductors,[193] topological insulators,[194] and even various dimensionalities such as layered structures,[195] nanotubes,[187,196] fullerenes,[197] and molecules.[198,199] For electronic transport, the NEGF method has been implemented in a large variety of representations, ranging from envelope function approximations such as the effective mass[200,201] and **k·p** method,[202] to atomistic representations such as tight binding[203,204] and even density functional theory models.[205] In most cases, the NEGF method is used for transport in open systems, i.e., devices connected to spin, charge or heat reservoirs via semi-infinite leads. Nevertheless, two different approaches have been successfully applied to mimic the field-periodic conditions cascade structures are facing.[118,122,206,207]

It is this high flexibility of NEGF that offers choosing specialized basis functions such as, e.g., Wannier-Stark states for cascade structures,[206] wire modes for homogeneous wire systems,[189] or generic basis functions for general cases within the low rank approximation.[208] It is true for all these cases that the closer the basis functions match the actual quasi particles of the system, the fewer basis functions are required to reliably predict the device performance and the more efficient the numerical implementation of NEGF will be. It is worth to mention here that the NEGF method natively contains more information than semiclassical methods such as the Boltzmann equation. In fact, it has already been shown in the 1960s how to approximate the NEGF equations to yield the Boltzmann equation.[181]

The NEGF method, however, faces one major drawback compared to most other methods: the numerical solution of the NEGF equations is expensive both in terms of memory and central processing unit (CPU) time. The numerical costs are particularly high if incoherent scattering in the self-consistent Born approximation is considered. For this reason, most numerical implementations of NEGF on real devices do not include electron-electron scattering beyond the first order (Hartree) approximation. Higher orders of electron-electron scattering (i.e., exchange terms) are required to model energy transfer during inelastic electron-electron

scattering, but their numerical load typically prohibits the implementation on concrete transport problems.

B. Overview of the non-equilibrium Green's function method

1. Fundamental equations and observables

As discussed in Sec. II A, electrons in quantum cascade lasers can be successfully described within the effective mass approximation. Hereby, the devices are typically considered as laterally homogeneous quantum well heterostructures. The electronic structure is then represented with the Hamiltonian

$$\hat{H}_0 = \frac{-\hbar^2}{2} \partial_z \frac{1}{m^*(z,E)} \partial_z + \frac{\hbar^2 k^2}{2m^{\parallel}(z,E)} + V(z),$$
$$V(z) = V_c(z) - e\Phi(z),$$
(149)

where k is the lateral electron momentum, $\Phi(z)$ the electrostatic potential, and $V_c(z)$ denotes the material and position dependent conduction band edge, including the band offsets. Note that the effective mass is energy-dependent to include the nonparabolicity as described in detail in Sec. II C. In the stationary limit, the NEGF method describes transport with the electronic retarded and lesser Green's function G^R and $G^<$, respectively. These functions solve four coupled partial differential equations that read in operator form for a given energy E and in-plane momentum $\mathbf{k}$[209]

$$G^R = \left(E\hat{1} - \hat{H}_0 - \Sigma^R\right)^{-1},$$
(150a)

$$G^< = G^R \Sigma^< G^{R^{\dagger}},$$
(150b)

$$\Sigma^< = G^< D^<,$$
(150c)

$$\Sigma^R = G^R D^R + G^R D^< + G^< D^R.$$
(150d)

Here, Σ^R and $\Sigma^<$ denote the retarded and lesser self-energies and D^R and $D^<$ are the sum of retarded and lesser Green's function of the environment. Equations (150a) and (150b) are also referred to as Dyson and Keldysh equation, respectively. The expressions for the self-energies, Eqs. (150c) and (150d), are discussed in Sec. VIII D for various scattering mechanisms. We note that in that section and in the rest of this paper all Green's functions and self-energies are given in real space representation which requires a transformation of Eqs. (150) from operator space into real space. Thus, the Green's functions and self-energies given in the following have the dimensions eV^{-1} m^{-1} and eV m^{-1}, respectively, unless stated otherwise. Green's functions and self-energies are solved in the self-consistent Born approximation, i.e., they are solved iteratively until convergence is achieved. Most of the known current conserving simplifications of the self-energies do not ease the convergence of the self-consistent Born iterations, but reduce the numerical burden of each individual iteration only. In contrast to methods that require the solution of the Schrödinger equation, the solutions of Eqs. (150) do not require to solve an eigenvalue

problem. Consequently, energy dependent effective masses in the Hamiltonian $\hat{H}_0$ do not increase the numerical complexity of NEGF. The Green's functions and self-energies are functions of two spatial coordinates z, z', the absolute value of the lateral momentum k and the energy E. The energy and spatially resolved spectral function $A(z, E) = i[G^R(z, z, 0, E) - G^{R^\dagger}(z, z, 0, E)]$ shows width and location of resonant states in the system. This indicates that G^R contains the information of resonant states and the density of states. The $G^<$ function contains in addition information of the occupancy of the electronic states. Consequently, occupancy related observables such as density, current, and optical gain are dependent on $G^<$: the spatially and energy-resolved density $n(z, E)$ and current density $J(z, E)$ are defined in relation to the density $n(z)$ and current density $J(z)$, respectively,

$$
\begin{aligned}
n(z) &= \int \mathrm{d}E n(z, E) \\
&= \frac{2}{(2\pi)^3} \Im \int \mathrm{d}E \int \mathrm{d}^2 k G^<(z, z, k, E),
\end{aligned} \tag{151}
$$

$$
\begin{aligned}
J(z) &= \int \mathrm{d}E J(z, E) \\
&= -\frac{\hbar e}{(2\pi)^3} \lim_{z' \to z} \int \mathrm{d}E \int \mathrm{d}^2 k \frac{1}{m^*(z, E)} \\
&\quad \times \Re(\partial_z - \partial_{z'}) G^<(z, z', k, E).
\end{aligned} \tag{152}
$$

The current density in Eq. (152) is only correct as long as the kinetic energy operator is the only term in the Hamiltonian that does not commute with the position operator. This can be seen from the fundamental probability current density operator in real space $\mathbf{x}$ and time t representation (see, e.g., Refs. 210 and 211)

$$
\langle \hat{J}(\mathbf{x}_1, t_1) \rangle = \lim_{t_2 \to t_1} \lim_{\mathbf{x}_2 \to \mathbf{x}_1} i\hbar \hat{v} G^<(\mathbf{x}_1, t_1; \mathbf{x}_2, t_2), \tag{153}
$$

with the velocity operator

$$
\hat{v} = \frac{i}{\hbar} [\hat{H}, \hat{x}]_-. \tag{154}
$$

Please note that the Green's function in Eq. (153) is given in real space $\mathbf{x}$ and time t representation and has the dimension $\mathrm{eV}^{-1} \, \mathrm{m}^{-3} \, \mathrm{s}^{-1}$. The optical field amplitude absorption coefficient $\alpha(z, \omega)$ for a photon of frequency ω is a function of the permittivity $\epsilon(z, \omega)$[212]

$$
\alpha(z, \omega) = \Im\left[\frac{\epsilon(z, \omega)}{\epsilon_0}\right] \frac{\sqrt{2}\omega}{c} \left\{ \Re\left[\frac{\epsilon(z, \omega)}{\epsilon_0}\right] + \left|\frac{\epsilon(z, \omega)}{\epsilon_0}\right| \right\}^{-1/2}. \tag{155}
$$

Here, ϵ_0 is the vaccum permittivity and c denotes the speed of light. The optical absorption coefficient is usually defined with respect to the field intensity (a) rather than with respect to the field amplitude,[88] see Sec. III. In particular, in the gain regime one often refers to the power gain $-a$. These quantities are related by $a = 2\alpha$. The permittivity depends on the complex conductance $\sigma(z, \omega)$ and the materials dielectric constant $\epsilon_r(z)$

$$
\epsilon(z, \omega) = \epsilon_0 \epsilon_r(z) + i\sigma(z, \omega)/\omega. \tag{156}
$$

Before lasing starts and the perturbation $\delta\hat{V}(\omega)$ due to the optical field is still small, the optical absorption can be extracted from the linear response of the Green's functions. To first order, the change of the lesser Green's function is given by (for a given $\mathbf{k}$)[116]

$$
\begin{aligned}
\delta G^<(\omega, E) &= G^R(E + \hbar\omega)\delta\hat{V}(\omega)G^<(E) \\
&\quad + G^<(E + \hbar\omega)\delta\hat{V}(\omega)G^{R^\dagger}(E) \\
&\quad + G^R(E + \hbar\omega)\delta\Sigma^R(E + \hbar\omega, -E)G^<(E) \\
&\quad + G^<(E + \hbar\omega)\delta\Sigma^{R^\dagger}(E + \hbar\omega, -E)G^{R^\dagger}(E) \\
&\quad + G^R(E + \hbar\omega)\delta\Sigma^<(E + \hbar\omega, -E)G^{R^\dagger}(E).
\end{aligned} \tag{157}
$$

Changes to the scattering self-energies $\delta\Sigma$ due to the photonic perturbation $\delta\hat{V}$ are first order vertex corrections.

The wavelengths of terahertz lasers are much larger than the typical dimensions of the QCL periods. Thus, for the electron transport calculation, the optical electric field can be considered constant in the active device. Since the QCL laser light is usually linearly polarized in transport direction z, the perturbing potential reads in Coulomb gauge

$$
\delta\hat{V}(\omega) = \frac{\hbar e}{m^*(z)\omega} E_z(\omega)\partial_z, \tag{158}
$$

with the photon electric field component in z direction E_z. This results in the change of the current density $\delta J(z, \omega)$ in linear order of $\delta\hat{V}(\omega)$

$$
\begin{aligned}
\delta J(z_1, \omega) &= -\lim_{z_2 \to z_1} \frac{\hbar^2 e}{m^*(z_1)} (\partial_{z_1} - \partial_{z_2}) \\
&\quad \times \int \frac{\mathrm{d}E}{2\pi\hbar} \int \frac{\mathrm{d}^2 k}{(2\pi)^2} \delta G^<(z_1, z_2, k, \omega, E) \\
&\quad - \frac{2\hbar e^2 E_z(\omega)}{m^*(z_1)\omega} \int \frac{\mathrm{d}E}{2\pi\hbar} \int \frac{\mathrm{d}^2 k}{(2\pi)^2} G^<(z_1, z_1, k, E).
\end{aligned} \tag{159}
$$

The first term of the last equation results from the current operator in Eq. (152) applied on the change of the lesser Green's functions $\delta G^<$. The second term results from the change of the velocity and of the current operator due to the fact that $\delta\hat{V}(\omega)$ does not commute with the position operator [see Eq. (154)]. The quotient of the perturbation of the current density $\delta J(z, \omega)$ and the electric field $E_z(\omega)$ of the photon gives us the optical conductance

$$
\sigma(z, \omega) = \frac{\delta J(z, \omega)}{E_z(\omega)}. \tag{160}
$$

If vertex corrections $\delta\Sigma$ are ignored for simplicity, Eqs. (158), (157), and (159) can be combined into an equation for the optical conductance

$$\sigma(z_1, \omega) = \lim_{z_2 \to z_1} \frac{\hbar^2 e^2}{m^*(z_1)^2 (2\pi)^3 \omega} (\partial_{z_1} - \partial_{z_2}) \int dE d^2k dz_3$$
$$\times [G^R(z_1, z_3, k, E + \hbar\omega) \partial_{z'} G^<(z', z_2, k, E)|_{z'=z_3}$$
$$+ G^<(z_1, z_3, k, E + \hbar\omega) \partial_{z'} G^A(z', z_2, k, E)|_{z'=z_3}]$$
$$- 2e^2 \int \frac{dE d^2k}{(2\pi)^3 m^*(z_1)\omega} G^<(z_1, z_1, k, E). \quad (161)$$

When the self-consistently solved Green's functions are used for the optical conductance, Eq. (161) fully accounts for the self-consistently calculated electron states and their non-equilibrium state occupations. The vertex corrections $\delta\Sigma$ in Eq. (157) increase the numerical load significantly, since they require self-consistent iterations of δG with $\delta\Sigma$. Both δG and $\delta\Sigma$ depend on the optical frequency ω in addition to the "standard" dependence of the Green's function [e.g., $\delta\Sigma = \delta\Sigma(z, z', k, E, \omega)$]. Thus, such self-consistency is expensive in terms of memory and time. It has been shown that vertex corrections narrow the optical linewidths and increase the peak height of the absorption coefficient in terahertz QCLs.[117] Calculations without $\delta\Sigma$ may not predict quantitative values of $\alpha(z, \omega)$, but can serve for qualitative predictions only. If the interaction with the photonic field should be considered beyond linear response—such as in the case of electron transport during lasing, time dependent NEGF with a periodically oscillating electric field is required. Details of this approach can be found in Ref. 124.

The balance between the benefits of the NEGF method and its numerical load is strongly device physics dependent. If the electron transport is clearly dominated by incoherent scattering and tunneling across multiple barriers is negligible, semiclassical models are clearly more efficient than NEGF. Otherwise, if scattering is negligible, the NEGF method has to compete with numerically more efficient methods such as the Schrödinger equation or the quantum transmitting boundary method.[213] If scattering in low order captures the physics and energy resolved information is not desired, the density matrix method is more efficient than NEGF as well. The strength of NEGF is that it allows to predict transport in all these regimes and gives deep insight to the ongoing processes in any of the before mentioned situations. It will be discussed in Sec. VIII E in detail that a single terahertz QCL can move from the ballistic to the scattering dominated regime with the applied electric field. Therefore, it is appropriate to use NEGF on these devices.

2. Different basis representations and low rank approximations

The solution of the NEGF equations, Eqs. (150), requires many matrix-matrix products and matrix inversions. Both scale cubically with N, the number of degrees of freedom the equations are discretized in. As a result, highly resolved NEGF calculations can easily require modern supercomputers.[44] It is obvious that any reduction of N will reduce the number of floating point operations significantly.[208] Real space discretization of Eqs. (150) faces the challenge that any device feature such as differing widths of

semiconductor layers has to match to the chosen resolution. Particularly small device features require an inhomogeneous grid in real space, increasing the numerical complexity.

The most common technique to reduce N is to transform the NEGF equations into a system of basis functions that are equal or close to the actual propagating eigenstates of the system's Hamiltonian. The more the basis functions agree with the eigenfunctions of the system, the less the basis functions couple with each other and the fewer functions need to be included in the actual transport calculation. If the target basis has less states than the rank N of the original (real space) representation, the transformation matrices are rectangular. In the ideal case, when the basis functions are isolated from each other, their individual contribution to transport can be solved and summed. Prominent examples of such rank reductions are the mode space approaches for transport in nanowire structures. If the basis functions are coupled, it is common to still limit the number of considered functions. In this case, however, this approach gives only approximate results. In general cases, transformations with rectangular matrices that reduce the rank of the equations are called low rank approximations.[208]

In the framework of cascade devices, a small number N of Wannier or Wannier-Stark basis functions have proven to represent quantum transport well enough for reliable predictions. For instance, Lee et al. were able to solve the IV characteristics and optical output performance of terahertz QCLs with only 5 states per period.[214] Although it is a common and numerically very efficient technique, caution is required when results of the low rank space are transformed back into the original real space representation if the rank of the two differs a lot. Numerically, the back transformation produces results with the high resolution of the original space, but the reliability of the data in the real space resolution strongly depends on how well the basis functions of the low rank space match the physics of the device. Unphysical oscillations of the current density with real space can indicate such reliability issues. Increasing the number of considered basis functions in the low rank space can improve the reliability of the approximate results.[208,214]

C. Boundary conditions for cascade systems

Boundary conditions in stationary NEGF always fall into two categories: boundary conditions for G^R that are related to the density of states and boundary conditions for $G^<$ that are related to state occupancy and particle distributions. Since quantum cascade lasers consist of many repetitions of the same building block—a sequence of quantum wells and barriers—NEGF is often applied on cascade structures with field periodic boundary conditions.[121,206,215] Field periodic boundary conditions for the retarded Green's function can be mimicked by solving the NEGF equations in a system of 3 or more periods: the innermost period is considered as the actual active device while the boundary periods are considered as charge suppliers only. The Poisson equation is solved with periodic boundary condition and an applied homogeneous electric field is added to the resulting potential. Since the source sided periods are on average on a

higher potential (and contrary for the drain sided periods), the system effectively faces field periodic charge boundary conditions. The initial guess for the lesser Green's function $G^<$ is also field periodic. Since the system is closed, self-consistent Born iterations of the higher order scattering terms may change the number of electrons. To maintain global charge neutrality, the $G^<$ function needs rescaling between iterations.[215] Although this type of boundary treatment is appealing for cascade systems, care on the numerical details has to be taken: In the first iteration of the self-consistent Born approximation the system is purely coherent. Then, the equation for G^R is van-Hove singular whenever the energy E agrees with a resonance of the system. Otherwise, the resulting retarded Green's function is real and the spectral function A vanishes (i.e., no electronic states are found). To solve this issue, a small artificial retarded scattering self-energy is typically included in Eqs. (150). This self-energy maintains a minimal broadening of resonances. This artificial broadening has to be negligible compared to the actual scattering self-energies. Note that purely coherent and field periodic calculations would clearly yield divergent electron energies in nonequilibrium.

Most applications of NEGF tackle transport in open systems. In open systems, the active device is connected via semi-infinite leads with charge reservoirs that supply or collect electrons with a constant, typically equilibrium distribution. Technically, the boundary conditions of G^R and $G^<$ are then introduced by including contact self-energies Σ^R_{con} and $\Sigma^<_{con}$ in the equations for G^R and $G^<$ in Eqs. (150), respectively. In this case, the large number of cascade periods can be mimicked by considering one or a few periods as the actual active device and including the potential landscape and material sequences of the adjacent periods in the calculation of Σ^R.[118,122] In this way, the presence of the leads broadens resonant states in the active device and allows for the solution of purely ballistic transport without adding artificial retarded self-energies. It also allows for the assessment of individual scattering mechanisms, since scattering and its broadening of the states is not essential for numerical convergence and can be turned off selectively. However, the boundary condition for $G^<$ requires a model for the electron distribution in the leads. It is common in NEGF for open systems to assume equilibrium Fermi distributions. In the case of cascade devices, however, such equilibrium distributions do not include possible heating effects within previous cascade periods and therefore allow results of nonperiodic electron distributions: When heating effects exist, the electrons do not dissipate all energy they gain while traversing the potential drop of a single period. Then, the electron distributions at source and drain side of each period differ.[118]

Alternatively, an electron distribution can be read out of $G^<$ at device positions that are equivalent to the lead/device interfaces. In this case, the nonequilibrium distribution is defined as

$$f(z,k,E) \equiv -iG^<(z,z,k,E)/A(z,z,k,E), \quad (162)$$

with the spectral function

$$A(z,z',k,E) = i\left[G^R(z,z',k,E) - G^{R\dagger}(z,z',k,E)\right]. \quad (163)$$

In equilibrium, the function $f(z,k,E)$ equals the Fermi distribution $f(E,\mu)$.[209] When the distribution function of source ($z=0$) and drain ($z=L$) are assumed to be equivalent to the distribution functions at positions in the device that are a single QCL period apart

$$f(0,k,E) \overset{!}{=} f(nL_p,k,E+ne\phi),$$
$$f(L,k,E) \overset{!}{=} f(L-nL_p,k,E-ne\phi),$$

field periodic boundary conditions for $G^<$ are mimicked.[122] Here, L_p is the length of a QCL period, ϕ is the potential drop per period, and n is the number of explicitly considered QCL periods. To ensure global charge neutrality, the Poisson equation is solved with the applied electric field at both boundaries.

D. Scattering self-energies

In the following, the resulting expressions for the self-energies, given in Eqs. (150c) and (150d), are summarized for the most important scattering mechanisms in QCLs. If multiple scattering mechanisms are considered, the individual self-energies of the respective scattering mechanisms can be summed up and the summed self-energy is then used in Eqs. (150). Some of the scattering self-energies are numerically so expensive to implement that approximations are inevitable. Some approximations are shown that ease the numerical burden significantly but still faithfully reproduce the scattering rates expected from Fermi's golden rule. It is worth mentioning that all presented approximations have been carefully assessed. Approximations that violate Pauli blocking, underestimate effects of nonlocality of quantum mechanics or miss important characteristics of the scattering potentials have been avoided. Coherent effects or the balance between incoherent and coherent QCL physics are to the best of the authors' knowledge not affected by the presented approximations. A detailed discussion of the validity of approximations in NEGF can be found in Ref. 120. This section ends with a brief discussion on general approximations of scattering self-energies and some numerical details for efficient implementations.

1. Scattering from longitudinal acoustic phonons

To avoid many-particle Green's functions, the phonon gas is assumed to remain unchanged by the electron propagation. This requires applying the perturbation potential of Eq. (66) in second order and to group the phonon creation with annihilation operators.[216] The resulting products of phonon and electron operators are then translated into the electron and phonon Green's functions in Eq. (150) for the lesser and retarded self-energies[195,209,211]

$$\Sigma^<(z_3,z_4,k,E) = \frac{1}{(2\pi)^3}\int d^2q\,dq_z|C_\mathbf{Q}|^2 e^{iq_z(z_3-z_4)}$$
$$\times [N_\mathbf{Q}G^<(z_3,z_4,|\mathbf{k}-\mathbf{q}|,E-\hbar\omega_\mathbf{Q})$$
$$+(1+N_\mathbf{Q})G^<(z_3,z_4,|\mathbf{k}-\mathbf{q}|,E+\hbar\omega_\mathbf{Q})],$$
$$(164)$$

$$\Sigma^R(z_3, z_4, k, E) = \frac{1}{(2\pi)^3} \int d^2q dq_z |C_{\mathbf{Q}}|^2 e^{iq_z(z_3-z_4)}$$
$$\times [N_{\mathbf{Q}} G^R(z_3, z_4, |\mathbf{k}-\mathbf{q}|, E+\hbar\omega_{\mathbf{Q}})$$
$$+(1+N_{\mathbf{Q}}) G^R(z_3, z_4, |\mathbf{k}-\mathbf{q}|, E-\hbar\omega_{\mathbf{Q}})]$$
$$+\frac{1}{(2\pi)^3} \int d^2q dq_z |C_{\mathbf{Q}}|^2 e^{iq_z(z_3-z_4)}$$
$$\times \left[\frac{1}{2} G^<(z_3, z_4, |\mathbf{k}-\mathbf{q}|, E-\hbar\omega_{\mathbf{Q}}) \right.$$
$$\left. -\frac{1}{2} G^<(z_3, z_4, |\mathbf{k}-\mathbf{q}|, E+\hbar\omega_{\mathbf{Q}}) \right]$$
$$-\frac{i}{(2\pi)^4} \int dE' \int d^2q dq_z |C_{\mathbf{Q}}|^2 e^{iq_z(z_3-z_4)}$$
$$\times G^<(z_3, z_4, |\mathbf{k}-\mathbf{q}|, E-E')$$
$$\times \mathrm{Pr}\left(\frac{1}{E'+\hbar\omega_{\mathbf{Q}}} - \frac{1}{E'-\hbar\omega_{\mathbf{Q}}} \right). \tag{165}$$

Here, $\mathbf{Q} = [\mathbf{q}, q_z]^T$ is the phonon wave vector with the in-plane component $\mathbf{q}$, and $Q = |\mathbf{Q}| = \sqrt{q^2 + q_z^2}$. The coupling constant for a phonon of wave vector $\mathbf{Q}$ and phonon frequency $\omega_{\mathbf{Q}}$ is denoted with $C_{\mathbf{Q}}$, and the analytical phonon Green's functions are already inserted.[216] It is worth noting that Eqs. (164) and (165) are valid for electrons scattering from any type of bulk equilibrium phonons. In the case of the deformation potential perturbation and a linearized phonon dispersion relation,[149] the coupling constant reads [similar to Eq. (67)]

$$|C_{\mathbf{Q}}|^2 = \frac{\hbar\Xi^2}{2\rho_c \omega_{\mathbf{Q}}} Q^2. \tag{166}$$

Even in this approximate shape, Eqs. (164) and (165) require numerical solution of the three-dimensional $\mathbf{Q}$ integrals. Since that is typically beyond numerical feasibility, further approximations are commonly applied.

a. Elastic acoustic phonon scattering. A rather common approach[165,217] is to neglect energy changes of the electrons in the scattering process with the phonons

$$E \pm \hbar\omega_{\mathbf{Q}} \approx E. \tag{167}$$

In addition, high temperatures are assumed[218]

$$k_B T \gg \hbar\omega_{\mathbf{Q}}, \tag{168}$$

which allows to apply the equipartition approximation to the Bose distribution (see also Sec. VI B 1)

$$N_{\mathbf{Q}} + 1 \approx N_{\mathbf{Q}} \approx \frac{k_B T}{\hbar\omega_{\mathbf{Q}}} = \frac{k_B T}{\hbar v_s Q}. \tag{169}$$

The lower the phonon energy, the better justified the equipartition approximation. In the case of GaAs the acoustic phonon energy extends up to approximately 19 meV. At this phonon energy, the equipartition approximation is well justified for phonon bath temperatures above approximately 30 K, as illustrated in Fig. 22. This figure shows the exact
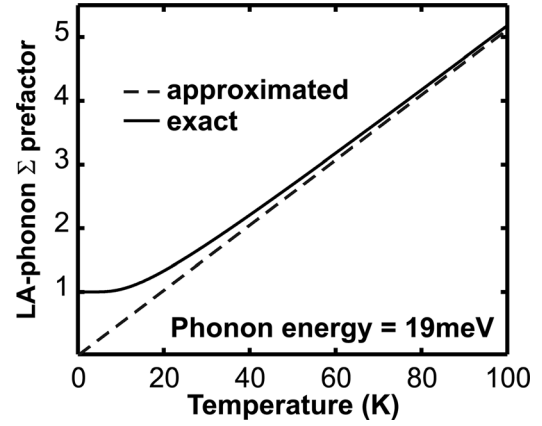

FIG. 22. Phonon-number-dependent prefactor in the integral of Eqs. (164) and (165) at 19 meV phonon energy as a function of the bath temperature. The high temperature approximation described in the text (dashed) converges to the exact solution (solid) at approximately 30 K. The lower the phonon energy, the faster both curves converge.

(solid) as well as the approximated [Eq. (169), dashed] result of $(1 + 2N_{\mathbf{Q}})$ at 19 meV.

With all these approximations the formula for the lesser self-energy finally reads

$$\Sigma^<(z_3, z_4, k, E) = \frac{1}{(2\pi)^3} \frac{k_B T \Xi^2}{\rho_c v_s^2}$$
$$\times \int d^2q dq_z e^{iq_z(z_3-z_4)} G^<(z_3, z_4, |\mathbf{k}-\mathbf{q}|, E). \tag{170}$$

When the integral over q_z is approximated[141] to run from $-\infty$ to ∞ it can be solved analytically which results in a local scattering self-energy

$$\Sigma^<(z_3, z_4, k, E) = \frac{1}{(2\pi)^2} \frac{k_B T \Xi^2}{\rho_c v_s^2} \delta(z_3 - z_4) \int d^2l G^<(z_3, z_4, l, E). \tag{171}$$

This scattering self-energy describes elastic scattering processes. Therefore, terms involving the lesser Green's function in the formula of the retarded self-energy Eq. (165) vanish exactly and the retarded self-energy can be obtained from Eq. (171) by replacing the "$<$" with "R".

b. Inelastic acoustic phonon scattering. An elastic approximation of the scattering with acoustic phonons misses an important physical effect: Inelastic scattering with acoustic phonons allows dissipation of arbitrarily small amounts of energy. In particular, when the electronic energy is insufficient to emit polar optical phonons, such an elastic treatment of acoustic phonons may underestimate electron thermalization. Setting the energy of acoustic phonons to a constant, but finite value still limits the minimum amount of dissipated energy.[206] Such small amounts of dissipated energy become important when the electrons do not carry enough energy to dissipate polar optical phonons. A full thermalization of such electrons requires the possibility to dissipate arbitrarily small amounts of energy.

A simple approximate inclusion of inelasticity is to replace the Green's function in Eq. (171) with an approximate Green's function $\tilde{G}$, averaged over the energy range of ± 1 acoustic phonon

$$\tilde{G}^{<,R}\left(z_3, z_4, l, \tilde{E}\right) = \frac{1}{2\hbar\omega_D} \int_{E-\hbar\omega_D}^{E+\hbar\omega_D} dE' G^{<,R}(z_3, z_4, l, E'),$$

$$\forall \tilde{E} \in [E - \hbar\omega_D, E + \hbar\omega_D]. \tag{172}$$

Hereby, the Debye frequency

$$\omega_D = \left(\frac{3\rho_N}{4\pi}\right)^{1/3} v_s, \tag{173}$$

with the number density ρ_N limits the width of that average. In connection with the equipartition approximation of Subsection VIII D 1 a, the numerical benefit from this approximation is threefold. First, the q_z integral in Eq. (164) can be solved analytically and yields a local scattering self-energy similar to Eq. (171)

$$\Sigma^<(z_3, z_4, k, E) = \frac{1}{(2\pi)^2} \frac{\Xi^2 k_B T}{\rho_c v_s^2 2\hbar\omega_D} \delta(z_3 - z_4)$$

$$\times \int_{E-\hbar\omega_D}^{E+\hbar\omega_D} dE' \int d^2 l G^<(z_3, z_4, l, E'). \tag{174}$$

Second, scattering with acoustic phonons is inelastically implemented and allows dissipation of arbitrarily small energies. Third, all terms containing $G^<$ in Eq. (165) vanish exactly and the retarded self-energy is independent of the lesser Green's functions

$$\Sigma^R(z_3, z_4, k, E) = \frac{1}{(2\pi)^2} \frac{\Xi^2 k_B T}{\rho_c v_s^2 2\hbar\omega_D} \delta(z_3 - z_4)$$

$$\times \int_{E-\hbar\omega_D}^{E+\hbar\omega_D} dE' \int d^2 l G^R(z_3, z_4, l, E'). \tag{175}$$

If all Σ^R are independent of $G^<$, the retarded functions can be solved in advance of the lesser Green's functions and self-energies. Once the retarded self-energy of a homogeneous system in equilibrium is known, the bulk, on-shell scattering rate can be extracted: In homogeneous systems, the self-energies depend only on the difference of the propagation coordinates ($r = z_3 - z_4$). A Fourier transform of the imaginary part of the retarded self-energy with respect to r agrees then with the scattering rate[195]

$$\tilde{\Gamma}(\mathbf{k}, k_z, E) = -\frac{2}{\hbar} \Im \int dr \exp(ik_z r) \Sigma^R(r, k, E). \tag{176}$$

If $\tilde{\Gamma}$ is evaluated at

$$k_z = \sqrt{2m^*(E)E/\hbar^2 - k^2 m^*(E)/m^\|(E)}, \tag{177}$$

it agrees with the on-shell scattering rate of bulk electrons with energy

$$E = \frac{\hbar^2}{2}\left[\frac{k^2}{m^\|(E)} + \frac{k_z^2}{m^*(E)}\right]. \tag{178}$$

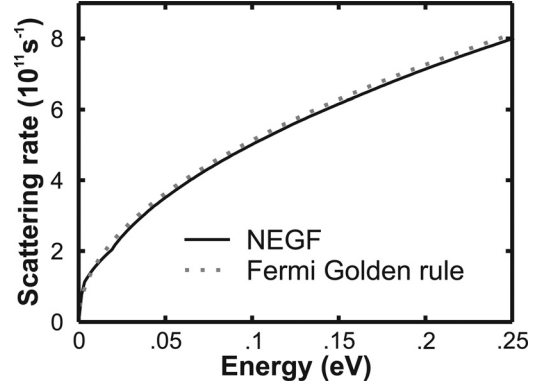

FIG. 23. On-shell scattering rate of electrons by longitudinal acoustic phonons in bulk n-doped GaAs at 300 K and $n = 2 \times 10^{18}$ cm^{-3}. The Fermi golden rule (dotted) and the NEGF calculation (full line) match perfectly.

Figure 23 shows that scattering rate for the self-energy Eq. (175). It nicely reproduces the on-shell scattering rate of Fermi's golden rule.

In spite of the scattering rate, the presented approximations of inelastic longitudinal acoustic phonons lead to an incorrect electron distribution: It can be shown that any phonon distribution that deviates from the Bose distribution [such as the approximate one of Eq. (169)] will cause the equilibrium electron distribution to deviate from the Fermi distribution.[219] This is illustrated in Fig. 24 as it shows the electronic occupation of the first state of an unbiased 10 nm wide In$_{.0165}$Ga$_{.9835}$As quantum well that is surrounded by 10 nm thick GaAs layers at various temperatures. In this calculation, only scattering on acoustic phonons given by Eqs. (174) and (175) is included. The zero of energy is set to the chemical potential of the device and energies below the conduction band edge are neglected. The higher the device temperature, the better is the agreement between the high-energy tail and the Fermi distribution. However, the occupations at
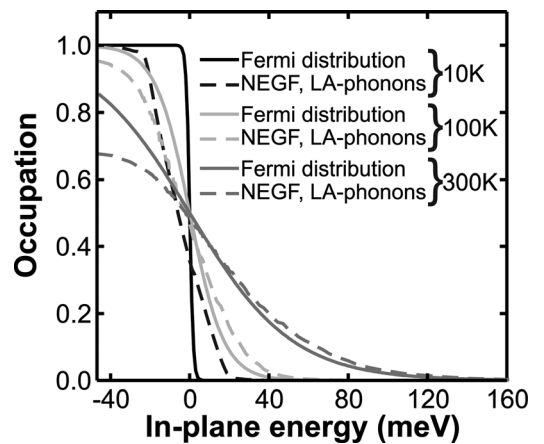

FIG. 24. In-plane electron distribution in the first state of a 10 nm wide In$_{.0165}$Ga$_{.9835}$As quantum well at equilibrium and various temperatures. The dashed lines show occupations resulting from NEGF calculations when the approximate inelastic scattering with LA-phonons of Eqs. (174) and (175) is the only included incoherent scattering mechanism. The Fermi distributions are depicted with solid lines. The zero in energy marks the chemical potential.

energies below the chemical potential are underestimated at any temperature.

Apart from these deviations from the Fermi distribution, Eqs. (174) and (175) are numerically efficient and describe the probability for inelastic scattering with acoustic phonons well enough. As soon as another inelastic scattering mechanism is included that is significantly stronger than the acoustic phonon scattering, such deviations from the Fermi distribution become negligible.

2. Scattering from polar longitudinal optical phonons

The scattering of electrons from polar optical phonons is discussed in detail in Sec. VI B 2. With the screened polar optical scattering potential of Eq. (85) the lesser self-energy reads similar to Eq. (86)

$$
\begin{aligned}
&\Sigma^<(z_3, z_4, k, E) \\
&= \frac{\gamma\pi}{(2\pi)^3} \int d^2 l \frac{e^{-\sqrt{|\mathbf{k}-\mathbf{l}|^2 + q_s^2}|z_3 - z_4|}}{\sqrt{|\mathbf{k}-\mathbf{l}|^2 + q_s^2}} \\
&\times \left[1 - \frac{q_s^2 |z_3 - z_4|}{2\sqrt{|\mathbf{k}-\mathbf{l}|^2 + q_s^2}} - \frac{q_s^2}{2\left(|\mathbf{k}-\mathbf{l}|^2 + q_s^2\right)} \right] \\
&\times [N_{\mathrm{Ph}} G^<(z_3, z_4, l, E - E_{\mathrm{LO}}) \\
&+ (1 + N_{\mathrm{Ph}}) G^<(z_3, z_4, l, E + E_{\mathrm{LO}})],
\end{aligned}
\tag{179}
$$

with the LO phonon energy $E_{\mathrm{LO}} = \hbar\omega_{\mathrm{LO}}$, phonon occupation number N_{Ph} given in Eq. (71), and

$$
\gamma = e^2 \frac{E_{\mathrm{LO}}}{2\epsilon_0} \left(\frac{1}{\epsilon_{r,\infty}} - \frac{1}{\epsilon_{r,0}} \right).
\tag{180}
$$

Analogously, the formula for the retarded self-energy can be derived to

$$
\begin{aligned}
&\Sigma^R(z_3, z_4, k, E) \\
&= \frac{\gamma\pi}{(2\pi)^3} \int d^2 l \frac{e^{-\sqrt{|\mathbf{k}-\mathbf{l}|^2 + q_s^2}|z_3 - z_4|}}{\sqrt{|\mathbf{k}-\mathbf{l}|^2 + q_s^2}} \\
&\times \left[1 - \frac{q_s^2 |z_3 - z_4|}{2\sqrt{|\mathbf{k}-\mathbf{l}|^2 + q_s^2}} - \frac{q_s^2}{2\left(|\mathbf{k}-\mathbf{l}|^2 + q_s^2\right)} \right] \\
&\times \left[(1 + N_{\mathrm{Ph}}) G^R(z_3, z_4, l, E - E_{\mathrm{LO}}) \right. \\
&+ N_{\mathrm{Ph}} G^R(z_3, z_4, l, E + E_{\mathrm{LO}}) + \frac{1}{2} G^<(z_3, z_4, l, E - E_{\mathrm{LO}}) \\
&- \frac{1}{2} G^<(z_3, z_4, l, E + E_{\mathrm{LO}}) + i \int \frac{d\tilde{E}}{2\pi} G^<(z_3, z_4, l, \tilde{E}) \\
&\times \left. \left(\Pr\frac{1}{E - \tilde{E} - E_{\mathrm{LO}}} - \Pr\frac{1}{E - \tilde{E} + E_{\mathrm{LO}}} \right) \right].
\end{aligned}
\tag{181}
$$

Svizhenko and Anantram have shown in one-dimensional systems that the principal value integrals of the last line in Eq. (181) shift the energies of resonant states.[220] When that
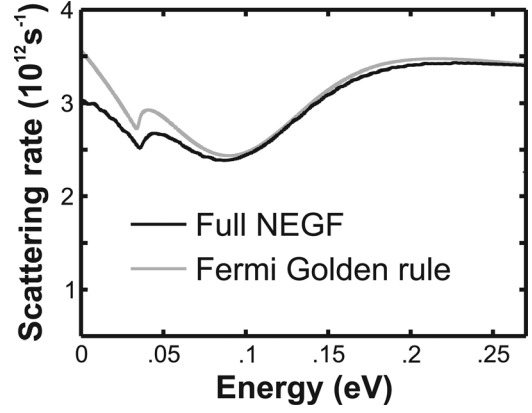

FIG. 25. On-shell scattering rate of electrons with screened polar optical phonons in homogeneous n-doped GaAs at 300 K. The doping concentration is 2×10^{18} cm^{-3} and the screening length is set to 5 nm for comparison. Results of NEGF calculations (black) agree nicely with Fermi's golden rule (grey). The scattering rate below the chemical potential at approximately 90 meV originates mainly from the scattering of conduction band holes.

shift is not differing significantly between different confined states, it leads only to a rigorous shift of the current-voltage characteristics. Therefore and since the principal value integrals are numerically expensive to solve, these principal value integrals are often neglected.[165,206,215] The on-shell scattering rate that corresponds to Eq. (181) is illustrated in Fig. 25. It shows results for a homogeneously n-doped GaAs device with $n = 2 \times 10^{18}$ cm^{-3}. It is worth emphasizing that the self-energies of the polar optical phonon scattering are finite when the two propagation coordinates z_3 and z_4 differ. This nonlocality of the scattering originates from the long-range nature of the Coulomb potential. Screening can efficiently limit this effect as can be seen from the exponents in Eqs. (179) and (181).

In Fig. 25, NEGF is solved in equilibrium at room temperature while only scattering with polar optical phonons is included. The black line in Fig. 25 shows the scattering rate Γ resulting from Eq. (176) while the gray line shows results of Fermi's golden rule (see, e.g., Ref. 218). Due to the high electron density, the chemical potential lies at approximately 90 meV above the conduction band edge (at 0 meV). In this case, holes in the conduction band at energies lower than the chemical potential contribute to the total scattering rate.[120] Therefore, the Fermi golden rule result is a sum of scattering probabilities of four kinds of scattering events: emission and absorption of LO-phonons by electrons and holes in the conduction band. When present, all these scattering mechanisms are automatically included in the self-energy of Eq. (181).

3. Brooks-Herring scattering on charged impurities

It has been discussed already in Sec. VI D 1 that for electrons scattering from charged impurities, the perturbing potential $V_{\mathrm{imp}}(\mathbf{x}; \mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_N)$ at position $\mathbf{x}$ is created by the N impurities at the position $\mathbf{s}_1, ..., \mathbf{s}_N$. In this section, more details on the perturbation of Green's functions due to impurity scattering are presented. It is particularly relevant since the NEGF formulation of impurity scattering differs from the semiclassical picture in Sec. VI D 1: the scattering on charged impurities may influence the electronic correlation function

between distinct positions. This is also true for scattering on polar optical phonons in Eqs. (179) and (181), but the following derivation illustrates nicely this effect. If the impurities are randomly distributed, the probability density to find an impurity at position $\mathbf{s}_i$ is given by

$$P(\mathbf{s}_i) = \frac{n_D(\mathbf{s}_i)}{N}. \tag{182}$$

Here, the total number of impurities N is given by the integral of the impurity density over the total volume V

$$\int_V n_D(\mathbf{s}) d^3 s = N. \tag{183}$$

As discussed in Sec. VI D 1, the first order of the impurity scattering is covered by the Poisson equation. For the second order, the squared scattering potential is required. The actual distribution of the impurities is unknown, thus the product of the perturbing potential has to be averaged over the impurity positions which introduces the impurity potential autocorrelation function

$$\langle V_{imp}(\mathbf{x}_3; \mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N) V_{imp}(\mathbf{x}_4; \mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N) \rangle_{imp}$$
$$= \frac{1}{N^N} \int \prod_{i=1}^{N} d^3 s_i n_D(\mathbf{s}_i) \frac{1}{(2\pi)^6}$$
$$\times \int d^3 Q d^3 Q' V_{imp}(Q) V_{imp}(Q') \sum_{j=1}^{N} e^{i\mathbf{Q}\cdot(\mathbf{x}_3 - \mathbf{s}_j)} e^{i\mathbf{Q}'\cdot(\mathbf{x}_4 - \mathbf{s}_j)}. \tag{184}$$

With some algebra, this simplifies to

$$\langle V_{imp}(\mathbf{x}_3; \mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N) V_{imp}(\mathbf{x}_4; \mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N) \rangle_{imp}$$
$$= \frac{1}{(2\pi)^6} \int d^3 s d^3 Q d^3 Q' n_D(\mathbf{s}) V_{imp}(Q)$$
$$\times V_{imp}(Q') e^{-i(\mathbf{Q}+\mathbf{Q}')\cdot\mathbf{s}} e^{i\mathbf{Q}\cdot\mathbf{x}_3} e^{i\mathbf{Q}'\cdot\mathbf{x}_4}. \tag{185}$$

Typical cascade lasers are set in the regime of the approach of Brooks and Herring (see discussion in Sec. VI D 1)[158]

$$V_{imp}(Q) = \frac{e^2}{\epsilon} \frac{1}{Q^2 + q_s^2}, \tag{186}$$

with the inverse Debye screening length q_s. In these systems, the doping profile is independent of in-plane positions $[n_D(\mathbf{s}) = n_D(z)]$ and the impurity potential correlation Eq. (185) can be simplified further

$$\langle V_{imp}(\mathbf{x}_3; \mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N) V_{imp}(\mathbf{x}_4; \mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N) \rangle_{imp}$$
$$= \frac{1}{(2\pi)^2} \int dz d^3 Q n_D(z) e^{i\mathbf{Q}\cdot(\mathbf{x}_3 - \mathbf{x}_4)}$$
$$\times V_{imp}(Q, z_3 - z) V_{imp}(-Q, z_4 - z). \tag{187}$$

The two-dimensional Fourier transform of the Debye-Hückel potential with respect to the in-plane coordinates reads[165]

$$V_{imp}(q, r_z) = \frac{e^2}{2\epsilon} \frac{\exp(-\sqrt{q_s^2 + q^2}|r_z|)}{\sqrt{q_s^2 + q^2}}. \tag{188}$$

In 3D real space representation, the elastic charged impurity scattering self-energy is a product of the electronic Green's function with the scattering potential correlation. After the Fourier transform with respect to the in-plane momentum the scattering self-energies read[165]

$$\Sigma_{imp}^{<,R}(z_3, z_4, k, E) = \frac{e^4}{16\pi^2\epsilon^2}$$
$$\times \int d^2 q F(z_3, z_4, |\mathbf{k} - \mathbf{q}|) G^{<,R}(z_3, z_4, q, E),$$
$$F\left(z_3, z_4, \sqrt{q_s^2 + p^2}\right) = \int dz n_D(z) \frac{e^{-\sqrt{q_s^2 + p^2}(|z_3 - z| + |z_4 - z|)}}{q_s^2 + p^2}. \tag{189}$$

Unfortunately, this result requires a three-dimensional integral for each value of (z_3, z_4, k, E). It turns out that a numerical implementation of such a self-energy is very time consuming and an approximation of the self-energy is necessary.

a. Averaged remote scattering. The scattering self-energy of the last paragraph is only in so far "exact," that the correlation function of impurities is not further approximated. However, the assumption of a constant inverse screening length q_s is already a significant approximation and in reality, screening in inhomogeneous devices is neither homogeneous, nor constant with respect to momentum and frequency. Instead, it would be a more realistic ansatz to describe the screening of charges with a polarization that depends on both propagation coordinates and is capable to describe the many particle effects correctly (see, e.g., Ref. 211). However, such a dielectric function is numerically too demanding.

Given this fact, it appears questionable to put great efforts in an "exact" numerical implementation of Eq. (189). Instead, the position dependent number of charged impurities n_D in Eq. (189) can be approximated with its average $\langle n_D \rangle_{\mathbf{x}_3, \mathbf{x}_4}$ along the shortest propagation path between both propagation coordinates $\mathbf{x}_3$ and $\mathbf{x}_4$. In this way, the influence of scattering at inhomogeneously distributed charged impurities is approximated with an effective scattering at homogeneously distributed impurities. This approximation effectively affects the function F of Eq. (189)

$$F\left(z_3, z_4, \sqrt{q_s^2 + p^2}\right) \approx n_D(z_3, z_4) \int dz \frac{e^{-\sqrt{q_s^2 + p^2}(|z_3 - z| + |z_4 - z|)}}{q_s^2 + p^2},$$

$$\langle n_D \rangle_{z_3, z_4} = \begin{cases} n_D(z_3), & \text{for } z_3 = z_4 \\ \dfrac{1}{z_4 - z_3} \displaystyle\int_{z_3}^{z_4} d\zeta \, n_D(\zeta), & \text{elsewhere.} \end{cases} \tag{190}$$

With this approximation, the scattering self-energy simplifies to

$$\Sigma^{<,R}(z_3, z_4, k, E) = \frac{\langle n_D \rangle_{z_3, z_4} e^4}{4\epsilon^2 (2\pi)^2}$$
$$\times \int d^2q \left[\frac{|z_3 - z_4| + 1/\sqrt{q_s^2 + q^2}}{q_s^2 + q^2} e^{-\sqrt{q_s^2+q^2}|z_3-z_4|} G^{<,R}(z_3, z_4, |\mathbf{k} - \mathbf{q}|, E) \right]. \quad (191)$$

To assess the validity of Eq. (190), we compare the exact solution of F of Eq. (189) with the approximate F of Eq. (190) in the case of a step like impurity density

$$n_D(z) \propto \begin{cases} 1, \forall |z| \le 7 \text{ nm} \\ 0, \qquad \text{else.} \end{cases} \quad (192)$$

Figure 26 shows the approximated [Eq. (190)] as well as the exact integral [Eq. (189)], integrated over z_3 and z_4 for various values of q. The smaller the transferred momentum in devices with large screening lengths is, the more important is the actual shape of the impurity density $n_D(z)$ and the more is the function F affected by the approximation in Eq. (190). Figure 26 shows that the approximation of Eq. (190) effectively overestimates scattering only for rather small transferred in-plane momenta. If the screening length is shorter than 20 nm (which corresponds to $q_s > 0.05 \text{ nm}^{-1}$), the discrepancy of the effective scattering strength is negligible.

Figure 27 shows F in the case of the impurity density in Eq. (192) with z_4 centered in the compact support of $n_D(z)$. The direct path between z_3 and z_4 may cross an area with nonzero $n_D(z)$. Then, the approximation of F (dashed) overestimates the exact solution (solid) for values of z_3 close to and within the compact support of $n_D(z)$. However, when the direct path between z_3 and z_4 does not touch the compact support of n_D, i.e.,

$$z_3, z_4 > 7 \text{ nm} \vee z_3, z_4 < -7 \text{ nm}, \quad (193)$$

the approximation of Eq. (190) underestimates the scattering (see Fig. 28). Consequently, the approximation in Eq. (190)

partly neglects remote scattering at charged impurities. Nevertheless, nonlocality of quantum mechanics as well as the self-consistent Born approximation ensure that a propagation between points z_3 and z_4 that fulfill Eq. (193) is still affected by scattering at the charged impurities. Figure 29 illustrates that Eq. (191) reproduces the on-shell scattering rate of Fermi's golden rule in a homogeneous GaAs system.

4. Scattering from rough interfaces

The scattering potential has been discussed already in Sec. VI D 2. Since a detailed knowledge of the (x, y)-dependence of the realistic interface is futile, the interface roughness potential $V_{IR}(\mathbf{x})$ has to be averaged in the (x, y)-direction. When the mid-point between the two materials is assumed to be randomly distributed within a roughness interval 2Δ the two possible values of $V_{IR}(x, y, z)$, i.e., $\pm V_o/2$ (where V_o is the conduction band offset) are equally likely and the (x, y)-average of the roughness potential $\langle V_{IR}(z) \rangle_{(x,y)}$ vanishes exactly. Consequently, the lowest order scattering at rough interfaces is of second order in V_o. Its derivation is analog to the scattering at charged impurities ($\forall z_3, z_4$ within the roughness interval)

$$\Sigma^{<,R}(z_3, z_4, \mathbf{r}_3, \mathbf{r}_4, E) = G^{<,R}(z_3, z_4, \mathbf{r}_3, \mathbf{r}_4, E)$$
$$\times \langle V_{IR}(z_3, \mathbf{r}_3) V_{IR}(z_4, \mathbf{r}_4) \rangle_{(x,y)}. \quad (194)$$

In general, the product of the perturbing potentials averaged with respect to the actual interface configuration $\langle V_{IR}(z_3, \mathbf{r}_3) V_{IR}(z_4, \mathbf{r}_4) \rangle_{(x,y)}$ depends on the growth coordinates z_3 and z_4. However, modern molecular beam epitaxy techniques allow for (depending on the growth direction)
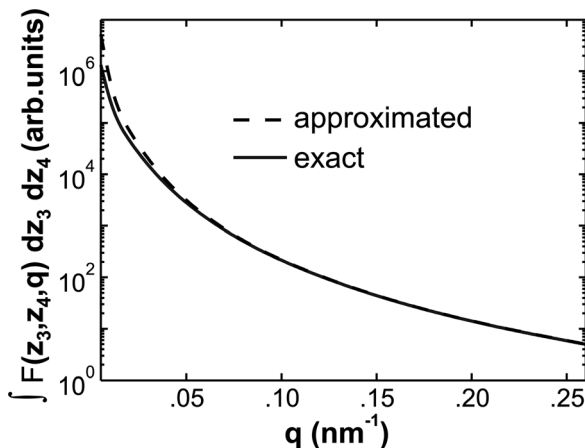

FIG. 26. The function F of Eqs. (189) and (190) integrated over both propagation coordinates. The approximation (dashed) of F deviates from the exact one (solid) only for small q.
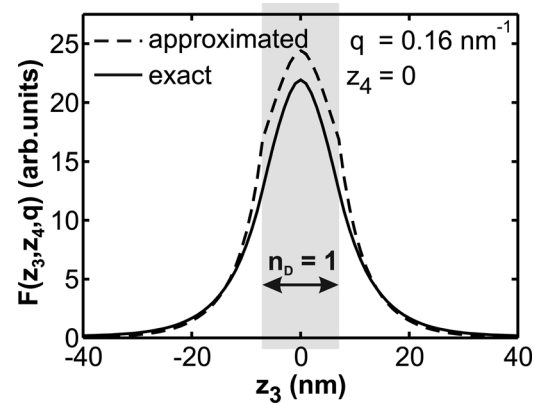

FIG. 27. The function $F(z_3, z_4 = 0, q)$ of Eqs. (189) and (190) at $q = 0.16 \text{ nm}^{-1}$. The impurity density $n_D(z)$ is given in Eq. (192) and is nonzero only in the gray shaded region. Note that the propagation coordinate z_4 is centered in the compact support of $n_D(z)$.
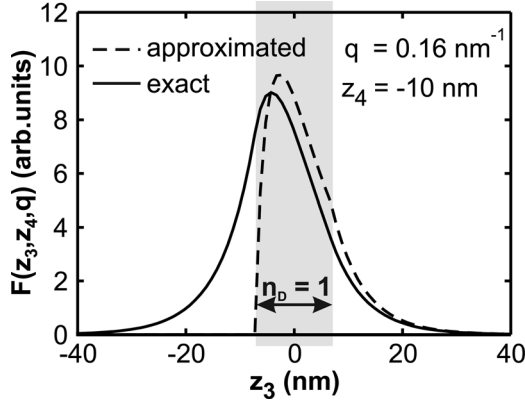
FIG. 28. Same as in Fig. 27, but with $z_4 = -10$ nm, i.e., outside of the compact support of $n_D(z)$ (grey shaded region).

almost monoatomic or monomolecular resolved growth of layers, so that the typical roughness interval extends only over a few Angstrom (see Sec. VI D 2). If this is much smaller than the typical numerical resolution in envelope function approximations (such as the effective mass representations), Δ is usually represented by a single discretization interval ($\forall z_3, z_4$ within the roughness interval)

$$\langle V_{IR}(x_3, y_3, z_3) V_{IR}(x_4, y_4, z_4) \rangle_{(x,y)}$$
$$= \langle V_{IR}(x_3, y_3, C) V_{IR}(x_4, y_4, C) \rangle_{(x,y)}. \quad (195)$$

If Δ is smaller than the numerical grid spacing s, the self-energies are finite only when its both propagation coordinates lie on the ideal midpoint C. Deviations of the step height from the grid spacing can be linearly interpolated

$$\Sigma^{<,R}(C,C,\mathbf{r}_3,\mathbf{r}_4,E) = \frac{2\Delta}{s} G^{<,R}(C,C,\mathbf{r}_3,\mathbf{r}_4,E)$$
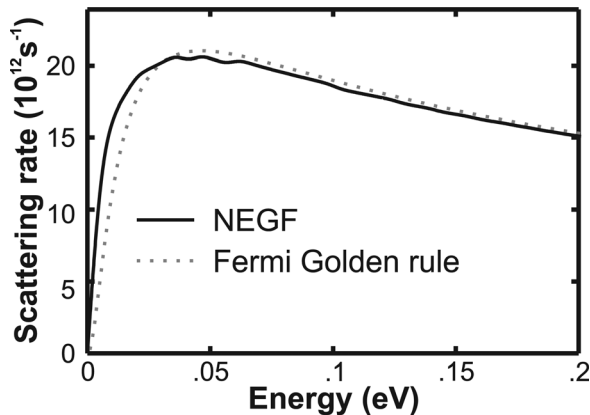$$\times \langle V_{IR}(\mathbf{r}_3,C) V_{IR}(\mathbf{r}_4,C) \rangle_{(x,y)}. \quad (196)$$

FIG. 29. On-shell scattering rate of electrons by charged impurities at 300 K in homogeneous n-doped GaAs. The doping concentration is 2×10^{18} cm^{-3}. The screening length was set constant to 5 nm for comparison (instead of the realistic 3 nm). The dotted line (Fermi golden rule) and the full line (NEGF calculation) agree nicely.

If no discretization point lies directly on the steep interface (as in most finite differences discretizations), the interface is at the center between two adjacent grid points. Then, one half of this self-energy can be distributed on each point adjacent to the interface. In this way, the Green's function on the interface is approximated with the Green's function of the respective adjacent grid point. It is worth noting that the prefactor of Δ^2 in formulas of Sec. VI D 2 originates from the transformation of z_3 and z_4 to delocalized basis states (see, e.g., Ref. 206).

a. Gaussian roughness correlation. As pointed out in Sec. VI D 2, it is a common approximation to assume a Gaussian-shaped in-plane interface roughness autocorrelation $\langle V_{IR}(\mathbf{r}_3, C) V_{IR}(\mathbf{r}_4, C) \rangle_{(x,y)}$ (see, e.g., Refs. 166, 179, and 221)

$$\langle V_{IR}(\mathbf{r}_3, C) V_{IR}(\mathbf{r}_4, C) \rangle_{(x,y)} = \frac{V_o^2}{4} \exp\left(-\frac{|\mathbf{r}_3 - \mathbf{r}_4|^2}{\Lambda^2}\right), \quad (197)$$

with the correlation length Λ of the roughness in (x,y)-direction. When this function is inserted into Eq. (196), Fourier transformed into the in-plane momentum space and integrated over the in-plane scattering angle, the scattering self-energy reads[165]

$$\Sigma^{<,R}(C,C,k,E) = \frac{\Lambda^2}{4\pi} \frac{V_o^2}{4} \exp(-k^2\Lambda^2/4) \times \int d^2q I_0\left(\frac{kq\Lambda^2}{2}\right)$$
$$\times \exp(-q^2\Lambda^2/4) G^{<,R}(C,C,q,E), \quad (198)$$

with the modified Bessel function

$$I_0(x) = \frac{1}{\pi} \int_0^\pi d\phi \exp(x\cos\phi). \quad (199)$$

b. Exponential shaped roughness correlation. Studies of different material systems have shown,[221,222] that an exponential shape of the roughness autocorrelation may better reproduce experimental data. Then, the autocorrelation of the scattering potential reads

$$\langle V_{IR}(\mathbf{r}_3, C) V_{IR}(\mathbf{r}_4, C) \rangle_{(x,y)} = \frac{V_o^2}{4} \exp\left(-\frac{|\mathbf{r}_3 - \mathbf{r}_4|}{\lambda}\right), \quad (200)$$

where λ is the exponential correlation length. When the Fourier transform of this function is inserted in Eq. (196) and the angle integral of the convolution is performed, the self-energy reads

$$\Sigma^{<,R}(C,C,k,E) = \frac{\lambda^2}{\pi} \frac{V_o^2}{2} \int_0^\infty dq E\left(2\sqrt{\frac{\lambda^2 kq}{1 + \lambda^2(k+q)^2}}\right)$$
$$\times \frac{q G^{<,R}(C,C,q,E)}{\sqrt{1 + \lambda^2(k+q)^2}\left[1 + \lambda^2(k-q)^2\right]}. \quad (201)$$

Here, $E(x)$ denotes the complete elliptical integral of the second kind,

$$E(x) = \int_0^x \frac{\sqrt{1 - k^2 t^2}}{1 - t^2} dt. \qquad (202)$$

5. Scattering from alloy disorder

As discussed in Sec. VI D 3, the perturbing potential of electrons scattering on alloy disorder is given by the difference of the real conduction band offset of the material at position z to the idealized one [see Eq. (113)]. The derivation of the self-energies for the scattering on alloy disorder is very analog to the scattering on charged impurities: In a random alloy, the first order contribution of δV vanishes and the second order in δV gives the first nonvanishing scattering self-energy. Band edge fluctuations,

$$\langle \delta V(x_3, y_3, z_3) \delta V(x_4, y_4, z_4) \rangle_{(x,y)}$$
$$= \delta V^2 \langle \delta x(x_3, y_3, z_3) \delta x(x_4, y_4, z_4) \rangle_{(x,y)}, \qquad (203)$$

are caused by concentration fluctuations δx. With the assumption of local randomness within the volume Ω_0 [see also Eq. (114)]

$$\langle \delta x(\mathbf{x}) \delta x(\mathbf{x}') \rangle = \Omega_0 x(1 - x) \delta(\mathbf{x} - \mathbf{x}'), \qquad (204)$$

the alloy scattering self-energy reads

$$\Sigma^{<,R}(z_3, z_4, k, E) = \Omega_0 x(1 - x) \delta V^2 \delta(z_3 - z_4)$$
$$\times \int d^2 q G^{<,R}(z_3, z_4, q, E). \qquad (205)$$

6. Inelastic electron-electron scattering

In most of the NEGF implementations, the electron-electron interaction is only included in the Hartree approximation. This mean field approach requires solving the Poisson equation self-consistently with the electron distribution within the active device. The inclusion of inelastic scattering, however, requires solving the NEGF equations beyond the Hartree-Fock approximation, as the Fock term gives only elastic scattering contributions. Higher order correlation terms increase the numerical effort dramatically: Correlation terms are proportional to convolutions of at least three single electron Green's functions. Approximations of these convolution integrals typically combine most of the electron Green's functions into the interaction term W and assume a product ansatz for the self-energy, i.e., GW approximations.[223] Both terms of this approximation are subject to further approximations. Assessments of typical approximations have shown that lower order approximations may even yield more realistic results than inclusions of higher order terms.[224,225] Examples for NEGF implementations of inelastic electron-electron scattering on cascade devices so far only include the plasmon pole approximation for W and a low-order GW_0 approximation.[118,119] The challenge of a numerically efficient implementation of electron-electron scattering in NEGF is not conclusively solved in literature. The assessment of possible approximations of electron-electron scattering in NEGF is beyond the scope of this article. We refer the reader to the literature for more thorough discussions.[118,119,205]

7. General remarks on scattering self-energies

It is worth making a few remarks on the implementation of scattering in NEGF in general. When NEGF is solved self-consistently with the Poisson equation, convergence essentially requires meshes in energy and momentum that resolve resonant states and van Hove singularities well. Most often that is only possible with inhomogeneous and adaptive meshes in energy and momentum.[165,219,226] All scattering self-energies of Secs. VIII D 1–VIII D 5 are given in their analytical form. When they are discretized in energy and momentum, they represent effective scattering between energy and momentum intervals of different sizes. For numerical current conservation, the discretization of the energy and momentum integrals in the scattering self-energies has to carefully account for these different sizes.[219] In most NEGF implementations, current conservation is ensured by solving the scattering self-energies in the self-consistent Born approximation. Alternative approaches that solve the self-energies in current conserving non-self-consistent approximations are the Büttiker probe approach[227] and the lowest order approximation of Ref. 228. These approaches are either limited to close-to equilibrium situations or to weak scattering only.[219,228] In other words, the Green's functions that appear in the scattering self-energies of the previous paragraphs are the full scattered Green's functions. Solving the Green's functions and self-energies in the full self-consistent Born approximation requires iterative solutions of the involved NEGF equations. Since this is numerically challenging, it is tempting to either truncate the iterations before convergence or to apply lower order approximations. While some of the low order approximations may still conserve the current, they can easily face artifacts. For instance, the neglect of $G^<$ contributions to the inelastic Σ^R self-energy has been shown to violate the Pauli blocking.[118,165]

The full energy and momentum integrals of self-energies such as the polar optical phonon scattering represent another high numerical burden. This has motivated several authors to simplify these integrals with representative transferred momenta. While this approximation eases the numerical load a lot, it has to be treated with great care, since the integrand functions vary significantly with the transferred momentum.[120,121]

Nonlocal scattering mechanisms such as the charged impurity scattering cause a much higher numerical load than local scattering mechanisms: Nonlocal scattering increases the number of nonvanishing elements in the inverse of the retarded Green's function in Eqs. (150) and thereby increases the number of floating point operations to solve G^R. Even the solutions of the scattering self-energies obviously require more operations and more memory, too. When nonlocal scattering is approximated with local scattering, at least an

appropriate compensation factor has to be introduced to avoid underestimation of the scattering rate.[120,229,230]

E. Selected results of NEGF on terahertz QCLs

In the following, some NEGF results for terahertz QCLs are presented that illustrate typical and important features of NEGF. All results represent the electron propagation in the QCLs in terms of stationary vertical transport in laterally homogeneous quantum well heterostructures. The QCLs are considered to be in contact with two charge reservoirs at $z = R$ and $z = L$, respectively. Thereby, the charge transport is treated as a scattering problem from source to drain, with the open device forming the scattering center. Cascade periods that surround the active device are included within the contact self-energies that supply electrons in equilibrium Fermi distributions. Incoherent scattering of electrons on optical and acoustic phonons, charged impurities, and rough interfaces is included. The electron-electron interaction in the Hartree approximation is taken into account. Material parameters are taken from Ref. 231. All results shown in this section are for terahertz QCLs consisting of periodically repeated GaAs and $Al_{.15}Ga_{.85}As$ layers. Each period consists of layers of the widths (30) 92 (55) 80 (27) 66 (41) 155 Å, where the values in parentheses indicate the $Al_{.15}Ga_{.85}As$ barriers.[232] Only the widest well is doped with a sheet doping density of 1.9×10^{10} cm^{-2}. Only a single period is explicitly calculated, whereas the remaining periods are covered within the lead model.

1. QCL work principle—Energy resolved spectral function

One important difference between NEGF and the density matrix method is that the resulting Green's function and self-energies are energy resolved. This additional information increases the numerical load, but it can unveil important insight into the device physics. One of the energy resolved quantities is the spectral function of Eq. (163) as it can illustrate the mechanisms that are responsible for gain in the presently studied QCL structures.

Figure 30 depicts a contour plot of the energy and spatially resolved spectral function of the terahertz QCL for vanishing lateral momentum $k = 0$ at a bias voltage of 33 mV per period which is below but close to the lasing threshold (of about 50 mV per period). The maxima of the spectral function represent resonant states. All states show a finite width and a fine structure that results from the coherent and incoherent coupling of all well states with one another. In other words, the width of the broadened levels corresponds to the total lifetime of the electrons in the respective device state. The upper laser level (labeled by #4) which is predominantly an antibonding state is aligned with the confined state #5 in the leftmost source-sided quantum well and therefore gets filled by resonant tunneling. The lower laser level #3 gets efficiently emptied by two mechanisms. First, the bonding state #3 is aligned with the states #2 and #2' of the rightmost well which allows its coherent depletion by tunneling. Second, the energy difference between this state and the lowest resonance state (#1)

matches approximately the energy of an LO phonon (36 meV) which leads to an additional depletion by the resonant emission of LO phonons. At the shown voltage the alignment is visible, but not fully established. The detuning of the alignment leads to a strong coherent leakage which is in more detail discussed below.

2. Effect of incoherent scattering

Transport calculations of mid-infrared QCLs suggest that the contribution of coherent propagation to the charge transport in QCLs is insignificant compared to the efficient incoherent scattering.[114] Furthermore, Monte Carlo solutions of the semiclassical Boltzmann equation, which neglect (coherent) correlation effects between laser states, have successfully predicted charge transport in terahertz QCLs near threshold.[107] Nevertheless, it can be shown that a general answer to the question whether the transport in terahertz QCLs is mainly coherent or incoherent cannot be given, since the balance between both is sensitive to details of the device structure. It actually turns out that the four well resonant phonon terahertz QCLs of Fig. 30 is a very instructive example for the interplay of coherent and incoherent transport.

A rather large portion of the current in the resonant phonon terahertz QCLs of Ref. 232 stems from coherent transport. This can be deduced from Fig. 31, comparing experimental and various theoretical results for the current density of this terahertz QCL.

The solid curve shows the calculation where phonon, impurity, interface roughness, and electron-electron scattering in the Hartree approximation have been fully included. The dash-dotted curve in Fig. 31 shows the limiting case where all scattering self-energies have been artificially turned off. Obviously, incoherent scattering enhances the current density in this QCL. In contrast, incoherent scattering
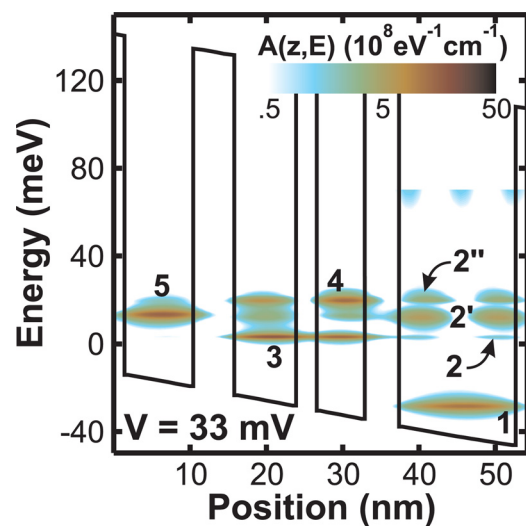

FIG. 30. Contour plot of the spectral function $A(z, E)$ of the QCL, in units of 10^8 eV^{-1} cm^{-1}, as a function of position z in nm and energy E in meV. The zero in energy marks the chemical potential of the source. The applied bias voltage is 33 mV per period. The solid line indicates the self-consistent potential profile. The spectral function is only shown within the energy interval from −50 to 70 meV. The labels number the relevant resonances that are discussed in the main text.

reduces the current density when no confining barriers are present (such as in transistor devices). In fact, it has been shown that incoherent scattering decreases the current in low resistive devices, whereas the current in high resistive devices is increased by incoherent scattering.[227]

3. Coherent regime

The two maxima of the ballistic current in Fig. 31 near 33 and 48 mV correspond to aligned laser states. At these voltages, electronic states are generated that extend across the entire QCL period. Such a delocalized state can be found in Fig. 30. At a bias voltage of 33 mV per period, the lowest state in the injector well (labeled with #5) is aligned with the second state ($\#2'$) of the rightmost quantum well, i.e., the collector well. This alignment generates a finite spectral function in the gap between the upper (#4) and the lower (#3) laser level. In this way, the spectral function at the energy $E = 12$ meV remains finite in every quantum well of the active period. Thus, electrons of this energy can coherently tunnel throughout the QCL period and maintain a maximum in the ballistic I-V characteristic.

Resonances in the current density caused by states that extend across the total QCL period are already known in literature. In the area of Monte Carlo simulations of QCLs, anticrossing of laser states also lead to such highly delocalized states which eases the coherent multibarrier tunneling. In the Monte Carlo formalism, however, the laser states are determined with a Hermitian Schrödinger equation which yields infinite state lifetimes. Thus, when the alignment conditions of the above delocalized states are met, the corresponding resonance in the I–V characteristics predicted in the Monte Carlo formalism is very large. It has been shown in Ref. 125 that a finite lifetime of the laser states reduces the height of the artificial current peaks significantly. The finite lifetime of the electrons in the ballistic calculations of Fig. 31 originates from the finite probability for electrons to leave the device and thereby to "decay" into lead states. In fact, the Schrödinger equation that corresponds to the solution of the Dyson equation [Eq. (150)] is non-Hermitian,
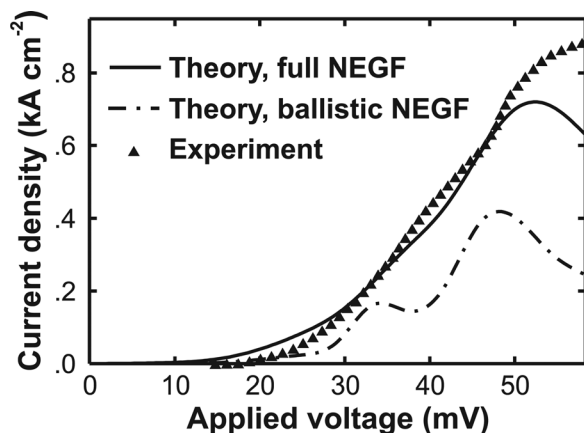
irrespective whether incoherent scattering is implemented or not. Thus, artificial spikes of the I–V characteristics cannot be seen in NEGF for open devices.

4. Incoherent regime

When the applied voltage of the QCL in Fig. 30 exceeds 33 mV the completely delocalized state breaks apart and the ballistic I–V characteristic shows a negative differential resistivity. Such a situation is depicted by the contour lines of Fig. 32 as they show the spectral function of the same QCL as in Fig. 30, but at an applied bias voltage of 52 mV per period. Here, the most prominent maxima of the spectral function separate into two groups of partly delocalized states: One group consists of the aligned injector (#5 in Fig. 30) and upper laser level (#4) and allows for the coherent propagation from the source sided device boundary to the center of each QCL period. The second group of delocalized states is generated by the alignment of the lower laser level (#3) and the second state of the collector well (#2) and eases the electronic propagation from the center of each QCL period to its drain sided limit. Since the states of both groups are energetically separated, electronic transitions between them require dissipation of energy. Thus, at this bias voltage, the coherent propagation throughout the total QCL period is suppressed, and the ballistic current is significantly smaller than the corresponding result for incoherent scattering included (see Fig. 31).

5. Incommensurate transport periodicity—Energy resolved current density

Typically, electrons in quantum cascade devices are expected to follow the periodicity of the structure. In theory, this is not necessarily the case, as can be seen from the contour plot of the spectral function at vanishing in-plane
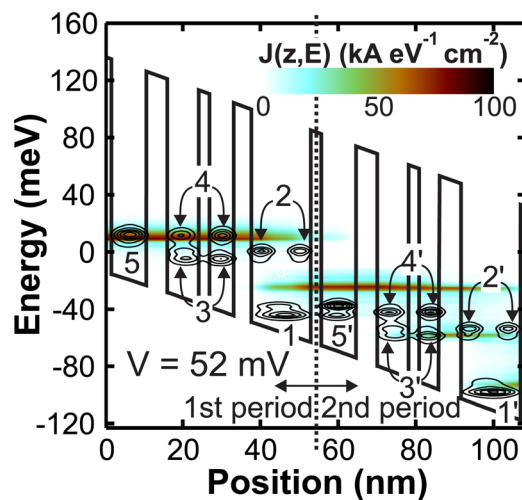

FIG. 32. Calculated conduction band profile (thick line), contour lines of the energy and spatially resolved spectral function $A(z, E)$ at vanishing in-plane momentum in two adjacent periods of the QCL at a bias voltage of 52 mV per period in the relevant energy range between -121 meV and 78 meV. The contour plot shows the spatially and energy resolved current density $J(z, E)$ in the same energy range. The zero in energy marks the chemical potential of the source. The dotted line marks the boundary between first and second period.

FIG. 31. Experimental (triangles) and theoretically predicted (solid lines) I–V characteristics of the QCL of Fig. 30. Ballistic calculations (dashed-dotted), i.e., NEGF calculations ignoring any incoherent scattering mechanism, underestimate the current density. At the two ballistic resonances, however, a large portion of the realistic current is coherent.

momentum shown in Fig. 32 for two adjacent QCL periods. The states associated with the first and second QCL period are labeled by numbers and primed numbers, respectively. In both periods, the alignment of the states follows the scheme described in detail for Fig. 30. In so far, the spectral function is periodic. In particular, the energy difference between the lower laser levels (3 and 3′) and the lowest collector states (1 and 1′) matches the energy of an LO phonon in both periods. Nevertheless, the carrier distribution deviates from the geometric QCL periodicity. This can be seen by a contour graph of the local energy resolved current density $J(z,E)$ [as defined in Eq. (152)] in Fig. 32. The function $J(z,E)$ shows spatially constant (i.e., horizontal) stripes in regions where the electrons propagate without dissipating energy. Disruptions of these horizontal stripes mark positions where LO-phonons get emitted. The figure shows that the number of emitted LO phonons is not equal for adjacent QCL periods. When the electrons have passed the first period and traversed a potential drop of 52 meV, they have emitted only one LO-phonon of energy 36 meV. The included elastic and inelastic scattering mechanisms are not able to dissipate the remaining 16 meV within this QCL period. This is a consequence of the good state alignment that supports efficient coherent multi-barrier tunneling. Consequently, the electrons enter the second period with an in-plane kinetic energy of 16 meV. As can be seen in Fig. 32, the energy of the leftmost current stripe coincides with states 5 and 4, whereas the

following current stripe lies above the corresponding states 5′ and 4′. Thus, these propagating electrons are now able to emit an LO-phonon, ending up in and occupying the lower laser level 3′. This occupancy leads to the build-up of the current stripe near $z = 80$ nm in Fig. 32 that is absent in the first period. The electrons can now tunnel resonantly from states 3′ into 2′ and scatter into the lowest collector state 1′ by the emission of an additional LO-phonon. Thus, the electrons have emitted a total of 3 LO phonons (3×36 meV $= 108$ meV) across 2 QCL periods (voltage drop of 2×52 meV $= 104$ meV) and are finally fully thermalized. The remaining small energy discrepancy can be gained from absorbing or emitting acoustic phonons. This process is repeated in the subsequent QCL periods such that a commensurable charge distribution with period two is established. Since the detailed energy balance depends on the applied bias voltage, the carrier density and current distribution may even become incommensurable with the geometric periods. A consequence of this incomplete carrier thermalization is a significant reduction in the occupation inversion and the optical gain in every other period. The calculated gain shows a drop of approximately 65% in the second period in Fig. 32. It has been estimated that electron-electron scattering cannot relax the electrons and restore the periodicity of the carrier distribution to a single QCL period.[118] This is mainly due to the efficient coherent tunneling of the electrons which supports resonant LO phonon emission instead.[156] Recent experimental findings indicate the heating of the electron gas described here.[233]

6. Temperature degradation—Energy resolved density

With increasing temperature, the number of electrons with high in-plane kinetic energies becomes significant. This can be seen in Fig. 33(a) that shows the energy resolved electron density $n(z,E)$ in the injector well (labeled with I), the two active quantum wells (II and III) and the collector well (IV) of the four well terahertz QCL of Fig. 30 at a temperature of 200 K. The subbands are depicted in Fig. 33(b) which shows the spatially integrated spectral function at vanishing in-plane momentum. The dashed lines mark the subband energies of the lowest state in the QCL period ($E = 17.4$ meV) as well as the lower ($E = 55.5$ meV and $E = 60.9$ meV) and the upper laser levels ($E = 71.1$ meV and $E = 73.4$ meV). Electrons with more than 20 meV of kinetic energy above the subband minima can be found in the upper laser level. This kinetic energy suffices for electrons in the upper laser level to scatter into the lower laser level by emitting an optical phonon (thermally activated phonon emission). In addition, some electrons in the collector well (IV) can reach sufficiently high kinetic energies to fill the lower laser level. These mechanisms reduce the occupation inversion and the optical gain.

For comparison, the energy resolved density and the spectral function of this QCL at 40 K lattice temperature are shown in Figs. 34(a) and 34(b), respectively. It is easy to see that both mechanisms, the thermally activated phonon emission and the thermal backfilling, are absent at this lower temperature.[234]
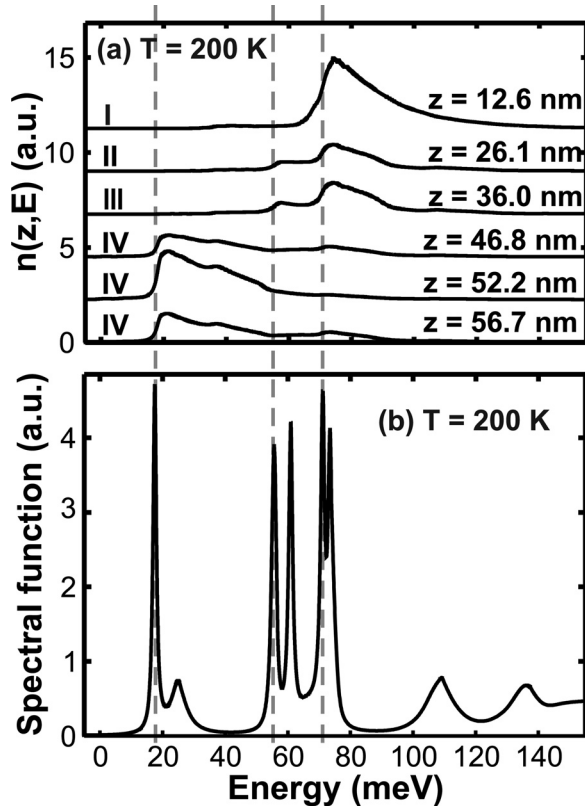

FIG. 33. (a) Energy resolved density at a bias of 50 mV per period at a lattice temperature of 200 K for various spatial positions in the QCL of Fig. 30. The Roman numbers denote the injector (I), the two active (II and III), and the collector well (IV). (b) Spatially integrated spectral function at 200 K; the peaks mark the energy of resonant states.
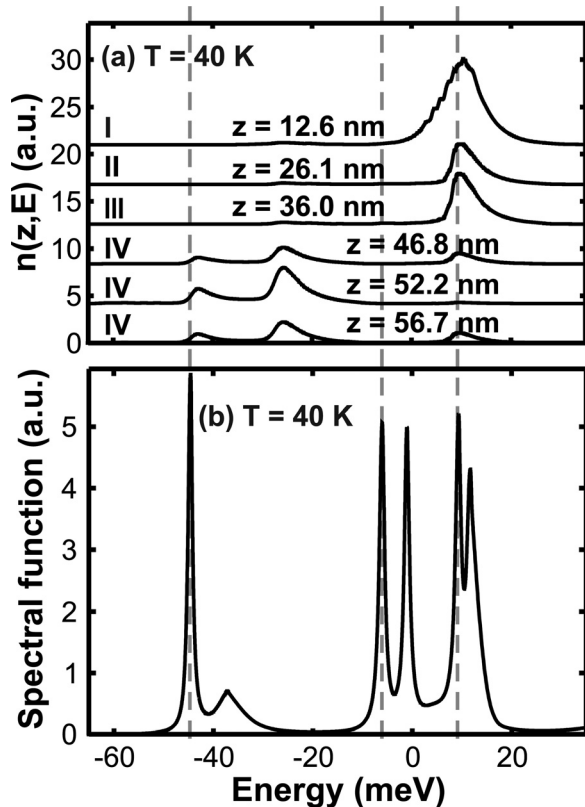
FIG. 34. (a) Energy resolved density and (b) spatially integrated spectral function for the same QCL as in Fig. 33, but at a lattice temperature of 40 K.

IX. CONCLUSION AND OUTLOOK

The advancement of QCL simulation approaches is driven by an intrinsic motivation to further improve the description and numerical implementation of the underlying physical processes, and by experimental progress requiring improved and extended simulations. A current research topic is the consideration of the relevant quantum effects at an acceptable numerical efficiency. Furthermore, the self-consistent inclusion of the laser field in the simulation is of great interest, especially in the context of frequency conversion structures for the terahertz and near infrared regime. Besides, alternative material systems extending the spectral range covered by QCLs or enabling high temperature terahertz operation might necessitate a theoretical treatment beyond the conduction band Γ valley.

A. Development of hybrid quantum-semiclassical and approximate quantum simulation approaches

Quantum coherence effects can play a pronounced role especially in terahertz QCLs where the energetic spacing between the quantized levels is relatively small.[115,125] On the other hand, quantum transport approaches such as the density matrix formalism or NEGF are numerically much more demanding than their semiclassical counterparts. Especially for full 3D simulations, this impedes their application to QCL design and optimization. Thus, there have been various efforts to implement certain aspects of the quantum transport theory into the semiclassical description to increase its accuracy and range of validity, but preserving its relative

numerical efficiency. Examples include phenomenologically extending the standard rate-equation models to include coherence,[235] incorporating energetic broadening of the quantized states into EMC simulations,[236,237] and combining the semiclassical Monte Carlo method with a density matrix approach.[125,238] Hybrid density matrix-Monte Carlo methods exploit the fact that resonant tunneling dominates the transport only in some regions, e.g., through a thick barrier such as an injection barrier, while a semiclassical transport description is adequate in the rest of the QCL structure.[125] A special challenge in hybrid quantum-semiclassical simulations will be a self-consistent description of dephasing going beyond phenomenological dephasing time models.[125] Furthermore, free carrier absorption[239,240] and electron leakage into the continuum of states[241,242] can have an effect and thus should be adequately implemented into the simulations.

B. Modeling of innovative QCL designs based on alternative material systems

Innovative QCL designs based on alternative material systems hold the potential of extending the spectral range covered by QCLs or enabling high temperature terahertz operation. The exploration of alternative QCL designs and subsequent systematic optimization is greatly facilitated by careful modeling. Up to now, lasing has only been obtained for n-type QCLs, using InGaAs/InAlAs on InP substrate or GaAs/AlGaAs on GaAs,[2,3] and to a smaller extent antimonides such as InAs/AlSb.[4,5] Since all these materials (apart from AlSb) have direct bandgaps, QCL simulations have up to now focused on the conduction band Γ valley. Thus, for alternative material systems with an indirect bandgap or using valence band transitions, the simulation methods will have to be correspondingly extended and adapted. Furthermore, additional effects such as strong polarization fields might have to be included.

One example for an alternative material system which could extend the application range of QCLs is GaN/AlGaN. Due to the large conduction band discontinuity, nitride-based QCL structures are promising candidates for short wavelength applications.[243,244] Furthermore, the GaN/AlGaN material system is also interesting for the development of high temperature terahertz QCLs because of the large optical phonon energies ($E_{LO} \approx 90 \, \text{meV}$).[243,245–247] However, no working nitride-based QCL has been demonstrated to date, only absorption and electroluminescence has been observed.[243,244] First NEGF results on a GaN-based QCL indicated a too large level broadening to maintain lasing.[248] The feasibility of terahertz GaN/AlGaN QCLs has also been studied using self-consistent rate equation models[245,247] and Monte Carlo simulations.[246] For simulating such structures, the band bending effects due to the strong intrinsic polarization fields have to be considered. These are clearly visible in Fig. 35, showing the conduction band profile and probability densities for an experimental GaN/AlGaN structure.[243]

An alternative to the conventional n-type QCLs are designs employing hole-based intersubband transitions in the valence band. Si/SiGe-based p-type structures are interesting since this material system offers a high integration potential
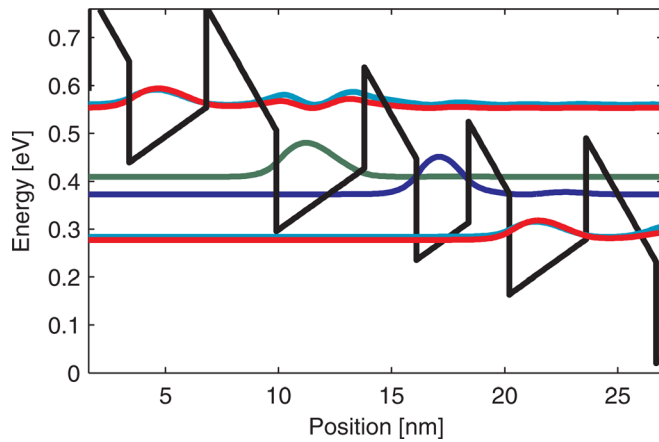
FIG. 35. Conduction band profile and probability densities for an experimental GaN/AlGaN structure.

and low process costs. For terahertz applications, a major advantage versus GaAs-based terahertz emitters is that potentially higher operating temperatures can be obtained due to the absence of LO phonon scattering in type IV semiconductors.[249] A working Si/SiGe QCL has not yet been demonstrated; however, electroluminescence could be achieved in the mid-infrared regime due to transitions between quantized heavy hole states,[250] and in the terahertz regime due to light hole-heavy hole transitions.[249] The simulation of hole-based devices is far more elaborate than for comparable n-type structures, since mixing effects of the heavy hole, light hole and split-off band have to be considered in the carrier transport simulations and the Schrödinger solver. Thus, a more complex description than the effective mass approximation typically used for n-type devices has to be employed, such as **k·p** theory.[251,252] Due to the absence of working p-type QCLs, only few theoretical studies based on rate equations[251,253] and the Monte Carlo method[252] have been published. If p-type QCLs based on Si/SiGe or another material system become technically feasible, the adaption and efficient numerical implementation of the different simulation approaches discussed in this review paper for the valence band will become a demanding task.

Alternatively, n-type terahertz QCLs based on Ge/SiGe have been considered.[254] Here, the lasing transition takes place in the conduction band L valleys, which are in contrast to the Γ valley both anisotropic and degenerate. Rate equation simulations have been performed to investigate the feasibility of such designs.[255] Intersubband absorption has been experimentally observed in the conduction band of Ge/SiGe quantum well structures,[256] but up to now no operating QCL has been demonstrated.

C. Inclusion of the optical cavity field

Simulations of the coupled electron and optical dynamics have mainly been performed using one-dimensional approaches such as rate[77,126] and Maxwell-Bloch[96–98] equations or the density matrix method.[113] On the other hand, with very few exceptions[76,123,238] advanced three-dimensional carrier transport simulations have focused on the electron dynamics, completely ignoring the light field.

However, the inclusion of the lasing field is not only required to study the actual lasing operation but also to model terahertz and infrared frequency conversion QCL sources. Here, the nonlinear optical properties of the QCL heterostructure must be adequately implemented, and the modeling of the optical cavity field is crucial to evaluate the nonlinear frequency conversion process. An example is terahertz difference frequency generation in QCL structures, enabling room temperature terahertz generation[20,22,23] with broadband frequency tunability.[23,24] A major goal is here to push the available room temperature output power from currently 120 μW (Ref. 23) to a few mW, as required for most technical applications. Recently, such a QCL device has been modeled using an EMC approach.[25] Artificial optical nonlinearities are also attractive for extending QCL operation towards shorter wavelengths, e.g., by using frequency doubling.[27]

The self-consistent inclusion of the optical cavity field requires adequate electromagnetic modeling of the resonator to determine the mode solutions and the corresponding overlap factor Γ and losses a_w, a_m. Increasingly, special resonator designs based on plasmonic effects or exhibiting subwavelength structuring are used for beam shaping or to enhance the efficiency and spectral purity. For example, the recent performance improvement of terahertz difference frequency sources has largely benefited from special cavity designs employing distributed feedback structures[31] or the Cherenkov effect.[23,28] In addition to general electromagnetic modeling approaches such as the finite element method, the development of adapted methods for specific cavity types, which are numerically efficient and provide more intuitive insight, is helpful for device simulation and optimization.

[1] J. Faist, F. Capasso, D. L. Sivco, C. Sirtori, A. L. Hutchinson, and A. Y. Cho, Science **264**, 553 (1994).
[2] B. S. Williams, Nat. Photonics **1**, 517 (2007).
[3] Y. Yao, A. J. Hoffman, and C. F. Gmachl, Nat. Photonics **6**, 432 (2012).
[4] K. Ohtani and H. Ohno, Appl. Phys. Lett. **82**, 1003 (2003).
[5] O. Cathabard, R. Teissier, J. Devenson, J. C. Moreno, and A. N. Baranov, Appl. Phys. Lett. **96**, 141110 (2010).
[6] M. Nobile, P. Klang, E. Mujagic, H. Detz, A. Andrews, W. Schrenk, and G. Strasser, Electron. Lett. **45**, 1031 (2009).
[7] C. Deutsch, A. Benz, H. Detz, P. Klang, M. Nobile, A. M. Andrews, W. Schrenk, T. Kubis, P. Vogl, G. Strasser, and K. Unterrainer, Appl. Phys. Lett. **97**, 261110 (2010).
[8] R. F. Curl, F. Capasso, C. Gmachl, A. A. Kosterev, B. McManus, R. Lewicki, M. Pusharsky, G. Wysocki, and F. K. Tittel, Chem. Phys. Lett. **487**, 1 (2010).
[9] Y. Bai, S. Slivken, S. Kuboya, S. R. Darvish, and M. Razeghi, Nat. Photonics **4**, 99 (2010).

[10]R. F. Kazarinov and R. A. Suris, Sov. Phys. Semicond. **5**, 707 (1971).

[11]J. Faist, F. Capasso, C. Sirtori, D. L. Sivco, A. L. Hutchinson, and A. Y. Cho, Electron. Lett. **32**, 560 (1996).

[12]Y. Bai, N. Bandyopadhyay, S. Tsao, S. Slivken, and M. Razeghi, Appl. Phys. Lett. **98**, 181102 (2011).

[13]G. Scalari, D. Turčinková, J. Lloyd-Hughes, M. I. Amanti, M. Fischer, M. Beck, and J. Faist, Appl. Phys. Lett. **97**, 081110 (2010).

[14]A. Hugi, R. Terazzi, Y. Bonetti, A. Wittmann, M. Fischer, M. Beck, J. Faist, and E. Gini, Appl. Phys. Lett. **95**, 061103 (2009).

[15]P. Q. Liu, A. J. Hoffman, M. D. Escarra, K. J. Franz, J. B. Khurgin, Y. Dikmelik, X. Wang, J. Fan, and C. F. Gmachl, Nat. Photonics **4**, 95 (2010).

[16]R. Köhler, A. Tredicucci, F. Beltram, H. E. Beere, E. H. Linfield, A. G. Davies, D. A. Ritchie, R. C. Iotti, and F. Rossi, Nature **417**, 156 (2002).

[17]S. Fathololoumi, E. Dupont, C. Chan, Z. Wasilewski, S. Laframboise, D. Ban, A. Mátyás, C. Jirauschek, Q. Hu, and H. Liu, Opt. Express **20**, 3866 (2012).

[18]A. Wade, G. Fedorov, D. Smirnov, S. Kumar, B. S. Williams, Q. Hu, and J. L. Reno, Nat. Photonics **3**, 41 (2009).

[19]M. A. Belkin, J. A. Fan, S. Hormoz, F. Capasso, S. P. Khanna, M. Lachab, A. G. Davies, and E. H. Linfield, Opt. Express **16**, 3242 (2008).

[20]M. Belkin, F. Capasso, F. Xie, A. Belyanin, M. Fischer, A. Wittmann, and J. Faist, Appl. Phys. Lett. **92**, 201101 (2008).

[21]R. W. Adams, A. Vizbaras, M. Jang, C. Grasse, S. Katz, G. Boehm, M. C. Amann, and M. A. Belkin, Appl. Phys. Lett. **98**, 151114 (2011).

[22]Q. Y. Lu, N. Bandyopadhyay, S. Slivken, Y. Bai, and M. Razeghi, Opt. Express **21**, 968 (2013).

[23]K. Vijayraghavan, Y. Jiang, M. Jang, A. Jiang, K. Choutagunta, A. Vizbaras, F. Demmerle, G. Boehm, M. C. Amann, and M. A. Belkin, Nat. Commun. **4**, 2021 (2013).

[24]Q. Y. Lu, N. Bandyopadhyay, S. Slivken, Y. Bai, and M. Razeghi, Appl. Phys. Lett. **101**, 251121 (2012).

[25]C. Jirauschek, A. Matyas, P. Lugli, and M.-C. Amann, Opt. Express **21**, 6180 (2013).

[26]N. Owschimikow, C. Gmachl, A. Belyanin, V. Kocharovsky, D. L. Sivco, R. Colombelli, F. Capasso, and A. Y. Cho, Phys. Rev. Lett. **90**, 043902 (2003).

[27]A. Vizbaras, M. Anders, S. Katz, C. Grasse, G. Boehm, R. Meyer, M. A. Belkin, and M.-C. Amann, IEEE J. Quantum Electron. **47**, 691 (2011).

[28]K. Vijayraghavan, R. W. Adams, A. Vizbaras, M. Jang, C. Grasse, G. Boehm, M. C. Amann, and M. A. Belkin, Appl. Phys. Lett. **100**, 251104 (2012).

[29]O. Demichel, L. Mahler, T. Losco, C. Mauro, R. Green, A. Tredicucci, J. Xu, F. Beltram, H. E. Beere, D. A. Ritchie, and V. Tamošinuas, Opt. Express **14**, 5335 (2006).

[30]A. Bousseksou, Y. Chassagneux, J. R. Coudevylle, R. Colombelli, C. Sirtori, G. Patriarche, G. Beaudoin, and I. Sagnes, Appl. Phys. Lett. **95**, 091105 (2009).

[31]Q. Y. Lu, N. Bandyopadhyay, S. Slivken, Y. Bai, and M. Razeghi, Appl. Phys. Lett. **99**, 131106 (2011).

[32]Y. Chassagneux, R. Colombelli, W. Maineult, S. Barbieri, H. E. Beere, D. A. Ritchie, S. P. Khanna, E. H. Linfield, and A. G. Davies, Nature **457**, 174 (2009).

[33]R. Colombelli, K. Srinivasan, M. Troccoli, O. Painter, C. F. Gmachl, D. M. Tennant, A. M. Sergent, D. L. Sivco, A. Y. Cho, and F. Capasso, Science **302**, 1374 (2003).

[34]M. Schubert and F. Rana, IEEE J. Quantum Electron. **42**, 257 (2006).

[35]A. Taflove and S. C. Hagness, *The Finite-Difference Time-Domain Method* (Artech House, Boston, 2000).

[36]G. Bastard, *Wave Mechanics Applied to Semiconductor Heterostructures* (Les Editions de Physique, Paris, 1988).

[37]I. Vurgaftman, J. R. Meyer, and L. R. Ram-Mohan, J. Appl. Phys. **89**, 5815 (2001).

[38]M. Cardona, Phys. Rev. **121**, 752 (1961).

[39]M. Sugawara, N. Okazaki, T. Fujii, and S. Yamazaki, Phys. Rev. B **48**, 8102 (1993).

[40]L. D. Landau and E. M. Lifshits, *Quantum Mechanics: Nonrelativistic Theory* (Pergamon Press, Oxford, 1977).

[41]B. Jonsson and S. T. Eng, IEEE J. Quantum Electron. **26**, 2025 (1990).

[42]E. Anemogiannis, E. N. Glytsis, and T. K. Gaylord, IEEE J. Quantum Electron. **29**, 2731 (1993).

[43]C. Jirauschek, IEEE J. Quantum Electron. **45**, 1059 (2009).

[44]S. Steiger, M. Povolotskyi, H.-H. Park, T. Kubis, and G. Klimeck, IEEE Trans. Nanotechnol. **10**, 1464 (2011).

[45]X. Gao, D. Botez, and I. Knezevic, Appl. Phys. Lett. **89**, 191119 (2006).

[46]X. Gao, D. Botez, and I. Knezevic, J. Appl. Phys. **101**, 063101 (2007).

[47]X. Gao, M. D'Souza, D. Botez, and I. Knezevic, J. Appl. Phys. **102**, 113107 (2007).

[48]X. Gao, D. Botez, and I. Knezevic, J. Appl. Phys. **103**, 073101 (2008).

[49]D. F. Nelson, R. C. Miller, and D. A. Kleinman, Phys. Rev. B **35**, 7770 (1987).

[50]G. Liu and S.-L. Chuang, Phys. Rev. B **65**, 165220 (2002).

[51]U. Ekenberg, Phys. Rev. B **40**, 7714 (1989).

[52]M. Braun and U. Rossler, J. Phys. C: Solid State Phys. **18**, 3365 (1985).

[53]C. Sirtori, F. Capasso, J. Faist, and S. Scandolo, Phys. Rev. B **50**, 8663 (1994).

[54]H. Tanimoto, N. Yasuda, K. Taniguchi, and C. Hamaguchi, Jpn. J. Appl. Phys., Part 1 **27**, 563 (1988).

[55]O. Bonno and J.-L. Thobel, J. Appl. Phys. **104**, 053719 (2008).

[56]Y. Ando and T. Itoh, J. Appl. Phys. **61**, 1497 (1987).

[57]E. Cassan, J. Appl. Phys. **87**, 7931 (2000).

[58]W. R. Frensley, *Heterostructures and Quantum Devices, VLSI Electronics: Microstructure Science*, edited by W. R. Frensley and N. G. Einspruch (Academic Press, 1994).

[59]C. Juang, K. J. Kuhn, and R. B. Darling, Phys. Rev. B **41**, 12047 (1990).

[60]D. Y. Ko and J. C. Inkson, Phys. Rev. B **38**, 9945 (1988).

[61]J. H. Davies, *The Physics of Low-Dimensional Semiconductors* (Cambridge University Press, Cambridge, 1997).

[62]S. Vatannia and G. Gildenblat, IEEE J. Quantum Electron. **32**, 1093 (1996).

[63]J.-G. S. Demers and R. Maciejko, J. Appl. Phys. **90**, 6120 (2001).

[64]W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing* (Cambridge University Press, Cambridge, 1992).

[65]J. D. Cooper, A. Valavanis, Z. Ikonić, P. Harrison, and J. E. Cunningham, J. Appl. Phys. **108**, 113109 (2010).

[66]H. Li, J. C. Cao, and H. C. Liu, Semicond. Sci. Technol. **23**, 125040 (2008).

[67]V. D. Jovanović, D. Indjin, N. Vukmirović, Z. Ikonić, P. Harrison, E. H. Linfield, H. Page, X. Marcadet, C. Sirtori, C. Worrall, H. E. Beere, and D. A. Ritchie, Appl. Phys. Lett. **86**, 211117 (2005).

[68]C. Jirauschek, A. Matyas, and P. Lugli, J. Appl. Phys. **107**, 013104 (2010).

[69]G. Scalari, L. Ajili, J. Faist, H. Beere, E. Linfield, D. Ritchie, and G. Davies, Appl. Phys. Lett. **82**, 3165 (2003).

[70]A. Trellakis, A. Galick, A. Pacelli, and U. Ravaioli, J. Appl. Phys. **81**, 7880 (1997).

[71]S. Kohen, B. S. Williams, and Q. Hu, J. Appl. Phys. **97**, 053106 (2005).

[72]J. A. Fan, M. A. Belkin, F. Capasso, S. Khanna, M. Lachab, A. G. Davies, and E. H. Linfield, Opt. Express **14**, 11672 (2006).

[73]P. Monk, *Finite Element Methods for Maxwell's Equations* (Oxford University Press, Oxford, 2003).

[74]A. Benz, C. Deutsch, G. Fasching, K. Unterrainer, A. M. Andrews, P. Klang, W. Schrenk, and G. Strasser, Opt. Express **17**, 941 (2009).

[75]N. Schulz, K. Bierwirth, F. Arndt, and U. Koster, IEEE Trans. Microwave Theory Tech. **38**, 722 (1990).

[76]C. Jirauschek, Appl. Phys. Lett. **96**, 011103 (2010).

[77]C. Peng, G. Chen, T. Yang, S.-W. Park, and R. Martini, Semicond. Sci. Technol. **28**, 105008 (2013).

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

[79]J. Butler and J. Zoroofchi, IEEE J. Quantum Electron. **10**, 809 (1974).

[80]T. Ikegami, IEEE J. Quantum Electron. **8**, 470 (1972).

[81]P. C. Kendall, D. A. Roberts, P. N. Robson, M. J. Adams, and M. J. Robertson, IEEE Photonics Technol. Lett. **5**, 148 (1993).

[82]F. K. Reinhart, I. Hayashi, and M. B. Panish, J. Appl. Phys. **42**, 4466 (1971).

[83]D. Dietze, J. Darmo, and K. Unterrainer, IEEE J. Quantum Electron. **46**, 618 (2010).

[84]M. Born and E. Wolf, *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light* (Cambridge University Press, Cambridge, 1999).

[85]R. Bräuer and O. Bryngdahl, Appl. Opt. **33**, 7875 (1994).

[86]K. Chiang, Opt. Lett. **16**, 714 (1991).

[87]A. S. Sudbo, Pure Appl. Opt. **3**, 381 (1994).

[88]A. E. Siegman, *Lasers* (University Science Books, Mill Valley, 1986).

[89]K. Faist, D. Hofstetter, M. Beck, T. Aellen, M. Rochat, and S. Blaser, IEEE J. Quantum Electron. **38**, 533 (2002).

[90]M. Yamanishi, T. Edamura, K. Fujita, N. Akikusa, and H. Kan, IEEE J. Quantum Electron. **44**, 12 (2008).

[91]D. Indjin, P. Harrison, R. W. Kelsall, and Z. Ikonić, J. Appl. Phys. **91**, 9019 (2002).

[92]K. Donovan, P. Harrison, and R. W. Kelsall, J. Appl. Phys. **89**, 3084 (2001).

[93]P. Slingerland, C. Baird, and R. H. Giles, Semicond. Sci. Technol. **27**, 065009 (2012).

[94]A. Mirčetić, D. Indjin, Z. Ikonić, P. Harrison, V. Milanović, and R. W. Kelsall, J. Appl. Phys. **97**, 084506 (2005).

[95]R. W. Boyd, *Nonlinear Optics* (Academic Press, San Diego, 2003).

[96]C. Y. Wang, L. Diehl, A. Gordon, C. Jirauschek, F. X. Kärtner, A. Belyanin, D. Bour, S. Corzine, G. Höfler, M. Troccoli, J. Faist, and F. Capasso, Phys. Rev. A **75**, 031802 (2007).

[97]V.-M. Gkortsas, C. Wang, L. Kuznetsova, L. Diehl, A. Gordon, C. Jirauschek, M. A. Belkin, A. Belyanin, F. Capasso, and F. X. Kärtner, Opt. Express **18**, 13616 (2010).

[98]C. R. Menyuk and M. A. Talukder, Phys. Rev. Lett. **102**, 023903 (2009).

[99]R. C. Iotti and F. Rossi, Appl. Phys. Lett. **76**, 2265 (2000).

[100]R. Köhler, R. C. Iotti, A. Tredicucci, and F. Rossi, Appl. Phys. Lett. **79**, 3920 (2001).

[101]H. Callebaut, S. Kumar, B. S. Williams, Q. Hu, and J. L. Reno, Appl. Phys. Lett. **83**, 207 (2003).

[102]H. Callebaut, S. Kumar, B. S. Williams, Q. Hu, and J. L. Reno, Appl. Phys. Lett. **84**, 645 (2004).

[103]F. Compagnone, A. di Carlo, and P. Lugli, Appl. Phys. Lett. **80**, 920 (2002).

[104]J. T. Lü and J. C. Cao, Appl. Phys. Lett. **88**, 061119 (2006).

[105]A. Mátyás, M. Belkin, P. Lugli, and C. Jirauschek, Appl. Phys. Lett. **96**, 201110 (2010).

[106]A. Mátyás, P. Lugli, and C. Jirauschek, J. Appl. Phys. **110**, 013108 (2011).

[107]C. Jirauschek and P. Lugli, J. Appl. Phys. **105**, 123102 (2009).

[108]C. Jirauschek and P. Lugli, Phys. Status Solidi C **5**, 221 (2008).

[109]H. Li, J. C. Cao, Z. Y. Tan, Y. J. Han, X. G. Guo, S. L. Feng, H. Luo, S. R. Laframboise, and H. C. Liu, J. Phys. D: Appl. Phys. **42**, 025101 (2009).

[110]O. Bonno, J.-L. Thobel, and F. Dessenne, J. Appl. Phys. **97**, 043702 (2005).

[111]S. Kumar and Q. Hu, Phys. Rev. B **80**, 245316 (2009).

[112]E. Dupont, S. Fathololoumi, and H. C. Liu, Phys. Rev. B **81**, 205311 (2010).

[113]R. Terazzi and J. Faist, New J. Phys. **12**, 033045 (2010).

[114]R. C. Iotti and F. Rossi, Phys. Rev. Lett. **87**, 146603 (2001).

[115]C. Weber, A. Wacker, and A. Knorr, Phys. Rev. B **79**, 165322 (2009).

[116]A. Wacker, Phys. Rev. B **66**, 085326 (2002).

[117]F. Banit, S.-C. Lee, A. Knorr, and A. Wacker, Appl. Phys. Lett. **86**, 041108 (2005).

[118]T. Kubis, C. Yeh, P. Vogl, A. Benz, G. Fasching, and C. Deutsch, Phys. Rev. B **79**, 195323 (2009).

[119]T. Schmielau and M. F. Pereira, Appl. Phys. Lett. **95**, 231111 (2009).

[120]T. Kubis and P. Vogl, Phys. Rev. B **83**, 195304 (2011).

[121]T. Schmielau and M. F. Pereira, Phys. Status Solidi B **246**, 329 (2009).

[122]A. Kolek, G. Haldas, and M. Bugajski, Appl. Phys. Lett. **101**, 061110 (2012).

[123]D. O. Winge, M. Lindskog, and A. Wacker, Appl. Phys. Lett. **101**, 211113 (2012).

[124]A. Wacker, M. Lindskog, and D. O. Winge, IEEE J. Sel. Top. Quantum Electron. **19**, 1 (2013).

[125]H. Callebaut and Q. Hu, J. Appl. Phys. **98**, 104505 (2005).

[126]L. Schrottke, M. Giehler, M. Wienold, R. Hey, and H. T. Grahn, Semicond. Sci. Technol. **25**, 045025 (2010).

[127]S. L. Lu, L. Schrottke, S. W. Teitsworth, R. Hey, and H. T. Grahn, Phys. Rev. B **73**, 033311 (2006).

[128]A. Hugi, G. Villares, S. Blaser, H. C. Liu, and J. Faist, Nature **492**, 229 (2012).

[129]M. A. Talukder and C. R. Menyuk, Appl. Phys. Lett. **95**, 071109 (2009).

[130]M. A. Talukder and C. R. Menyuk, Opt. Express **18**, 5639 (2010).

[131]R. C. Iotti, E. Ciancio, and F. Rossi, Phys. Rev. B **72**, 125347 (2005).

[132]P. Harrison, *Quantum Wells, Wires and Dots: Theoretical and Computational Physics* (Wiley, Chichester, 1999).

[133]W. A. Harrison, Phys. Rev. **104**, 1281 (1956).

[134]K. Hess, *Advanced Theory of Semiconductor Devices* (Wiley-IEEE Press, New York, 1999).

[135]C. Y. Chang and F. Kai, *GaAs High-Speed Devices* (John Wiley & Sons, New York, 1994).

[136]R. Nelander and A. Wacker, Phys. Status Solidi B **6**, 579 (2009).

[137]P. Lawaetz, Phys. Rev. **183**, 730 (1969).

[138]J. Bardeen and W. Shockley, Phys. Rev. **80**, 72 (1950).

[139]C. Kittel and C. Y. Fong, *Quantentheorie der Festkörper* (Oldenbourg, München, 1988).

[140]B. S. Williams and Q. Hu, J. Appl. Phys. **90**, 5504 (2001).

[141]S. M. Goodnick and P. Lugli, Phys. Rev. B **37**, 2578 (1988).

[142]S.-H. Park, D. Ahn, and Y.-T. Lee, Jpn. J. Appl. Phys., Part 1 **39**, 6601 (2000).

[143]R. Nelander and A. Wacker, Appl. Phys. Lett. **92**, 081102 (2008).

[144]F. Compagnone, M. Manenti, A. Di Carlo, and P. Lugli, Physica B **314**, 336 (2002).

[145]P. Lugli, C. Jacoboni, L. Reggiani, and P. Kocevar, Appl. Phys. Lett. **50**, 1251 (1987).

[146]V. Spagnolo, M. Troccoli, G. Scamarcio, C. Gmachl, F. Capasso, A. Tredicucci, A. M. Sergent, A. L. Hutchinson, D. L. Sivco, and A. Y. Cho, Appl. Phys. Lett. **78**, 2095 (2001).

[147]C. A. Evans, V. D. Jovanovic, D. Indjin, Z. Ikonic, and P. Harrison, IEEE J. Quantum Electron. **42**, 857 (2006).

[148]Y. B. Shi, Z. Aksamija, and I. Knezevic, J. Comput. Electron. **11**, 144 (2012).

[149]C. Jacoboni and P. Lugli, *The Monte Carlo Method for Semiconductor Device Simulation* (Springer, Wien, 1989).

[150]M. Moško, A. Mošková, and V. Cambel, Phys. Rev. B **51**, 16860 (1995).

[151]R. Nelander and A. Wacker, J. Appl. Phys. **106**, 063115 (2009).

[152]J. T. Lü and J. C. Cao, Appl. Phys. Lett. **89**, 211115 (2006).

[153]M. Moško and A. Mošková, Semicond. Sci. Technol. **9**, 478 (1994).

[154]S.-C. Lee and I. Galbraith, Phys. Rev. B **59**, 15796 (1999).

[155]T. Kubis, C. Yeh, and P. Vogl, Phys. Status Solidi C **5**, 232 (2008).

[156]T. Kubis, S. R. Mehrotra, and G. Klimeck, Appl. Phys. Lett. **97**, 261106 (2010).

[157]C. Deutsch, H. Detz, T. Zederbauer, A. M. Andrews, P. Klang, T. Kubis, G. Klimeck, M. E. Schuster, W. Schrenk, G. Strasser, and K. Unterrainer, Opt. Express **21**, 7209 (2013).

[158]H. Brooks, Phys. Rev. **83**, 879 (1951).

[159]E. Conwell and V. F. Weisskopf, Phys. Rev. **77**, 388 (1950).

[160]J. B. Krieger and T. Meeks, Phys. Rev. B **8**, 2780 (1973).

[161]D. Chattopadhyay and H. J. Queisser, Rev. Mod. Phys. **53**, 745 (1981).

[162]H. Sakaki, T. Noda, K. Hirakawa, M. Tanaka, and T. Matsusue, Appl. Phys. Lett. **51**, 1934 (1987).

[163]R. F. Schnabel, R. Zimmermann, D. Bimberg, H. Nickel, R. Lösch, and W. Schlapp, Phys. Rev. B **46**, 9873 (1992).

[164]P. Roblin, R. C. Potter, and A. Fathimulla, J. Appl. Phys. **79**, 2502 (1996).

[165]R. Lake, G. Klimeck, R. C. Bowen, and D. Jovanovic, J. Appl. Phys. **81**, 7845 (1997).

[166]B. R. Nag, Semicond. Sci. Technol. **19**, 162 (2004).

[167]S. Tsujino, A. Borak, E. Müller, M. Scheinert, C. V. Falub, H. Sigg, D. Grützmacher, M. Giovannini, and J. Faist, Appl. Phys. Lett. **86**, 062113 (2005).

[168]A. E. Asch and G. L. Hall, Phys. Rev. **132**, 1047 (1963).

[169]S. R. Mehrotra, A. Paul, and G. Klimeck, Appl. Phys. Lett. **98**, 173503 (2011).

[170]B. Boykin, G. Klimeck, R. C. Bowen, and F. Oyafuso, Phys. Rev. B **66**, 125207 (2002).

[171]G. Klimeck, R. C. Bowen, T. B. Boykin, C. Salazar-Lazaro, T. A. Cwik, and A. Stoica, Superlattices Microstruct. **27**, 77 (2000).

[172]S. Krishnamurthy, A. Sher, and A.-B. Chen, Appl. Phys. Lett. **47**, 160 (1985).

[173]P. Harrison, D. Indjin, and R. W. Kelsall, J. Appl. Phys. **92**, 6921 (2002).

[174]V. Spagnolo, G. Scamarcio, H. Page, and C. Sirtori, Appl. Phys. Lett. **84**, 3690 (2004).

[175]M. S. Vitiello, G. Scamarcio, V. Spagnolo, B. S. Williams, S. Kumar, Q. Hu, and J. L. Reno, Appl. Phys. Lett. **86**, 111115 (2005).

[176]R. C. Iotti and F. Rossi, Rep. Prog. Phys. **68**, 2533 (2005).

[177]C. Jirauschek, Opt. Express **18**, 25922 (2010).

[178]B. S. Williams, S. Kumar, Q. Hu, and J. L. Reno, Opt. Express **13**, 3331 (2005).

[179]T. Unuma, M. Yoshita, T. Noda, H. Sakaki, and H. Akiyama, J. Appl. Phys. **93**, 1586 (2003).

[180]J. Schwinger, J. Math. Phys. **2**, 407 (1961).

[181]L. P. Kadanoff and G. Baym, *Quantum Statistical Mechanics* (W. A. Benjamin, Inc., Menlo Park, California, 1962).

[182]L. V. Keldysh, Sov. Phys. JETP **20**, 1018 (1965).

[183]B. K. Nikolić, S. Souma, L. P. Zrbo, and J. Sinova, Phys. Rev. Lett. **95**, 046601 (2005).

[184]K. Suzuki and S. Kurihara, e-print arXiv:cond-mat/0611013v2.

[185]T. Kubis and P. Vogl, Phys. Status Solidi C **5**, 290 (2008).

[186]Y. Xu, J.-S. Wang, W. Duan, B.-L. Gu, and B. Li, Phys. Rev. B **78**, 224303 (2008).

[187]T. Yamamoto and K. Watanabe, Phys. Rev. Lett. **96**, 255503 (2006).

[188]M. Luisier, Phys. Rev. B **86**, 245407 (2012).

[189]M. Luisier, A. Schenk, and W. Fichtner, J. Appl. Phys. **100**, 043713 (2006).

[190]V. N. Do, P. Dollfus, and V. L. Nguyen, J. Appl. Phys. **100**, 093705 (2006).

[191]Z. Chen, J. Wang, B. Wang, and D. Y. Xing, Phys. Lett. A **334**, 436 (2005).

[192]Y. Ke, K. Xia, and H. Guo, Phys. Rev. Lett. **100**, 166805 (2008).

[193]A. Bulusu and D. G. Walker, ASME J. Heat Transfer **129**, 492 (2007).

[194]S. Hong, V. Diep, S. Datta, and Y. P. Chen, Phys. Rev. B **86**, 085131 (2012).

[195]A. Wacker, Phys. Rep. **357**, 1 (2002).

[196]M. Lazzeri, S. Piscanec, F. Mauri, A. C. Ferrari, and J. Robertson, Phys. Rev. Lett. **95**, 236802 (2005).

[197]H. Li and X. Q. Zhang, Phys. Lett. A **372**, 4294 (2008).

[198]T. Sato, K. Shizu, T. Kuga, K. Tanaka, and H. Kaji, Chem. Phys. Lett. **458**, 152 (2008).

[199]P. Damle, A. W. Ghosh, and S. Datta, Chem. Phys. **281**, 171 (2002).

[200]T. Kubis and P. Vogl, Laser Phys. **19**, 762 (2009).

[201]S. Birner, C. Schindler, P. Greck, M. Sabathil, and P. Vogl, J. Comput. Electron. **8**, 267 (2009).

[202]J. Z. Huang, W. C. Chew, J. Peng, C.-Y. Yam, L. J. Jiang, and G.-H. Chen, IEEE Trans. Electron Devices **60**, 2111 (2013).

[203]S. R. Mehrotra, S. Kim, T. Kubis, M. Povolotskyi, M. S. Lundstrom, and G. Klimeck, IEEE Trans. Electron Devices **60**, 2171 (2013).

[204]Z. Jiang, M. A. Kuroda, Y. Tan, D. M. Newns, M. Povolotskyi, T. B. Boykin, T. Kubis, and G. Klimeck, Appl. Phys. Lett. **102**, 193501 (2013).

[205]K. S. Thygesen and A. Rubio, Phys. Rev. B **77**, 115333 (2008).

[206]S.-C. Lee and A. Wacker, Phys. Rev. B **66**, 245314 (2002).

[207]N. Vukmirović, D. Indjin, Z. Ikonić, and P. Harrison, IEEE Photonics Technol. Lett. **20**, 129 (2008).

[208]L. Zeng, Y. He, M. Povolotsky, X. Liu, G. Klimeck, and T. Kubis, J. Appl. Phys. **113**, 213707 (2013).

[209]H. Haug and A.-P. Jauho, *Quantum Kinetics in Transport and Optics of Semiconductors* (Springer, Berlin, 1996).

[210]A. L. Fetter and J. D. Walecka, *Quantum Theory of Many-Particle Systems* (McGraw-Hill, Inc., New York, 1971).

[211]G. D. Mahan, *Many-Particle Physics*, edited by J. T. Devreese, R. P. Evrard, S. Lundquist, G. D. Mahan, and N. H. March (Plenum Press, New York, 2000).

[212]J. D. Jackson, *Classical Electrodynamics* (Wiley & Sons, New York, 1999).

[213]D. Z.-Y. Ting, E. T. Yu, and T. C. McGill, Phys. Rev. B **45**, 3583 (1992).

[214]S.-C. Lee, F. Banit, M. Woerner, and A. Wacker, Phys. Rev. B **73**, 245320 (2006).

[215]N. Vukmirović, Z. Ikonić, D. Indjin, and P. Harrison, Phys. Rev. B **76**, 245313 (2007).

[216]E. M. Lifschitz and L. P. Pitajewski, *Physical Kinetics*, Course of Theoretical Physics, Vol. 10 (Elsevier, Oxford, 2008).

[217]C. Jacoboni and L. Reggiani, Rev. Mod. Phys. **55**, 645 (1983).

[218]B. K. Ridley, *Quantum Processes in Semiconductors* (Oxford Science Publications, Oxford, 1982).

[219]T. Kubis, "Quantum transport in semiconductor nanostructures," Ph.D. thesis (Technische Universität München, 2009); available at http://nanohub.org/resources/8612.

[220]A. Svizhenko and M. P. Anantram, Phys. Rev. B **72**, 085430 (2005).

[221]S. Yamakawa, H. Ueno, K. Taniguchi, C. Hamaguchi, K. Miyatsuji, K. Masaki, and U. Ravaioli, J. Appl. Phys. **79**, 911 (1996).

[222]G. H. Kruithof, T. M. Klapwijk, and S. Bakker, Phys. Rev. B **43**, 6642 (1991).

[223]L. Hedin, Phys. Rev. **139**, A796 (1965).

[224]U. von Barth and B. Holm, Phys. Rev. B **54**, 8411 (1996).

[225]B. Holm and U. von Barth, Phys. Rev. B **57**, 2108 (1998).

[226]S. Agarwal, M. Povolotskyi, T. Kubis, and G. Klimeck, J. Comput. Electron. **9**, 252 (2010).

[227]M. Büttiker, Phys. Rev. B **33**, 3020 (1986).

[228]H. Mera, M. Lannoo, C. Li, N. Cavassilas, and M. Bescond, Phys. Rev. B **86**, 161404 (2012).

[229]G. Klimeck, R. Lake, C. L. Fernando, R. C. Bowen, D. Blanks, M. Leng, T. Moise, Y. C. Kao, and W. R. Frensley, in Proceedings of the International Conference on Quantum Devices and Circuits: Alexandria, Egypt, 4–7 June 1996, edited by K. Ismail, S. Bandyopadhyay, and J. P. Leburton (Alexandria, Egypt, 1996).

[230]M. Luisier and G. Klimeck, Phys. Rev. B **80**, 155430 (2009).

[231]Landolt-Börnstein, edited by O. Madelung, *Semiconductors: Intrinsic Properties of Group IV Elements and III–V, II–VI and I–VII Compounds* (Springer, Berlin, 1987).

[232]A. M. Andrews, A. Benz, C. Deutsch, G. Fasching, K. Unterrainer, P. Klang, W. Schrenk, and G. Strasser, Mater. Sci. Eng., B **147**, 152 (2008).

[233]K. Fujita, M. Yamanishi, S. Furuta, K. Tanaka, T. Edamura, T. Kubis, and G. Klimeck, Opt. Express **20**, 20647 (2012).

[234]H. Yasuda, T. Kubis, P. Vogl, N. Sekine, I. Hosako, and K. Hirakawa, Appl. Phys. Lett. **94**, 151109 (2009).

[235]A. Gordon and D. Majer, Phys. Rev. B **80**, 195317 (2009).

[236]Z. Aksamija and U. Ravaioli, J. Appl. Phys. **105**, 083722 (2009).

[237]A. Matyas, P. Lugli, and C. Jirauschek, Appl. Phys. Lett. **102**, 011101 (2013).

[238]I. Bhattacharya, C. W. I. Chan, and Q. Hu, Appl. Phys. Lett. **100**, 011108 (2012).

[239]J. S. Bhat, S. S. Kubakaddi, and B. G. Mulimani, J. Appl. Phys. **72**, 4966 (1992).

[240]A. Wacker, G. Bastard, F. Carosella, R. Ferreira, and E. Dupont, Phys. Rev. B **84**, 205319 (2011).

[241]A. Mátyás, T. Kubis, P. Lugli, and C. Jirauschek, Physica E **42**, 2628 (2010).

[242]W. Freeman and G. Karunasiri, Phys. Rev. B **85**, 195326 (2012).

[243]F. Sudradjat, W. Zhang, K. Driscoll, Y. Liao, A. Bhattacharyya, C. Thomidis, L. Zhou, D. J. Smith, T. D. Moustakas, and R. Paiella, J. Appl. Phys. **108**, 103704 (2010).

[244]C. Gmachl, H. M. Ng, and A. Y. Cho, Appl. Phys. Lett. **77**, 334 (2000).

[245]V. D. Jovanović, D. Indjin, Z. Ikonić, and P. Harrison, Appl. Phys. Lett. **84**, 2995 (2004).

[246]E. Bellotti, K. Driscoll, T. D. Moustakas, and R. Paiella, J. Appl. Phys. **105**, 113103 (2009).

[247]G. Sun, R. A. Soref, and J. B. Khurgin, Superlattices Microstruct. **37**, 107 (2005).

[248]H. Yasuda, T. Kubis, I. Hosako, and K. Hirakawa, J. Appl. Phys. **111**, 083105 (2012).

[249]S. A. Lynch, R. Bates, D. J. Paul, D. J. Norris, A. G. Cullis, Z. Ikonic, R. W. Kelsall, P. Harrison, D. D. Arnone, and C. R. Pidgeon, Appl. Phys. Lett. **81**, 1543 (2002).

[250]G. Dehlinger, L. Diehl, U. Gennser, H. Sigg, J. Faist, K. Ensslin, D. Grützmacher, and E. Müller, Science **290**, 2277 (2000).

[251]Z. Ikonić, P. Harrison, and R. W. Kelsall, J. Appl. Phys. **96**, 6803 (2004).

[252]Z. Ikonić, R. W. Kelsall, and P. Harrison, Phys. Rev. B **69**, 235308 (2004).

[253]L. Friedman, G. Sun, and R. A. Soref, Appl. Phys. Lett. **78**, 401 (2001).

[254]K. Driscoll and R. Paiella, J. Appl. Phys. **102**, 093103 (2007).

[255]L. Lever, A. Valavanis, C. A. Evans, Z. Ikonić, and R. W. Kelsall, Appl. Phys. Lett. **95**, 131103 (2009).

[256]M. de Seta, G. Capellini, Y. Busby, F. Evangelisti, M. Ortolani, M. Virgilio, G. Grosso, G. Pizzi, A. Nucara, and S. Lupi, Appl. Phys. Lett. **95**, 051918 (2009).