Editors' Suggestion

Atomic-Scale Insights into Semiconductor Heterostructures: From Experimental Three-Dimensional Analysis of the Interface to a Generalized Theory of Interfacial Roughness Scattering

T. Grange,[1,†] S. Mukherjee,[2,†] G. Capellini,[3,4,*] M. Montanari,[4] L. Persichetti,[4] L. Di Gaspare,[4] S. Birner,[1] A. Attiaoui,[2] O. Moutanabbir,[2] M. Virgilio,[5] and M. De Seta[4]

[1] *Nextnano GmbH, Garching b., München 85748, Germany*

[2] *Department of Engineering Physics, École Polytechnique de Montréal, Montreal, Quebec, Canada*

[3] *IHP – Leibniz-Institut für innovative Mikroelectronik, Im Technologiepark 25, 15236 Frankfurt (Oder), Germany*

[4] *Dipartimento di Scienze, Università degli Studi Roma Tre, 00146, Roma, Italy*

[5] *Dipartimento di Fisica "E. Fermi", Università di Pisa, Largo Pontecorvo 3, 56127 Pisa, Italy*

(Received 5 January 2020; revised manuscript received 3 February 2020; accepted 1 April 2020; published 23 April 2020)

In this manuscript, we develop a generalized theory for the scattering process produced by interface roughness on charge carriers that is suitable for any semiconductor heterostructure. By exploiting our experimental insights into the three-dimensional atomic landscape of Ge/Ge-Si heterointerfaces obtained by atom probe tomography, we are able to define the full set of interface parameters relevant to the scattering potential, including both the in-plane and axial correlation inside real diffuse interfaces. Our experimental findings indicate a partial coherence of the interface roughness along the growth direction within the interfaces. We show that it is necessary to include this feature, previously neglected by theoretical models, when heterointerfaces characterized by finite interface widths are taken into consideration. To show the relevance of our generalized scattering model in the physics of semiconductor devices, we implement it in a nonequilibrium Green's function simulation platform to assess the performance of a Ge/Si-Ge-based terahertz quantum cascade laser.

DOI: 10.1103/PhysRevApplied.13.044062

I. INTRODUCTION

Nowadays, the boundary between fundamental research and technology in the field of semiconductor science is becoming more and more blurred, as witnessed by the increasingly shorter time required to bring innovative materials and processes to mass production [1]. As an example, "exotic" quantum effects can now be exploited in nanoscale devices to improve their performance or to add functionalities. In this scenario, the interface in semiconductor heterostructures plays an ever-greater role and its characterization down to the atomic scale is becoming more relevant. Indeed, a well-defined and sharp interface is a prerequisite for a number of different physical systems leveraging on two-dimensional (2D) effects, such as solid-state-based qubits [2], thermoelectric devices [3,4], and photonic devices [5,6], as well as for more conventional architectures, such as field-effect transistors [7–9]. Furthermore, in the sub-10-nm CMOS technology nodes, interfaces are critical in developing high-quality nanosheet channels [1].

The interdiffusion of atoms and 2D or three-dimensional (3D) island formation [10] makes a "real" heterointerface deviate from an "ideal" flat atomic plane in two important aspects: (i) the average composition profile is smeared out along the growth (axial) direction, known as interface broadening and usually quantified by the interface width $\mathcal{L}$; (ii) the isocompositional surfaces exhibit in-plane fluctuations of their axial position, which are responsible for the so-called interface roughness (IFR).

The interface broadening and roughness influence the optoelectronic properties in different ways and thus, they are usually modeled separately. Interface broadening results in the broadening of the potential experienced by charge carriers in the z direction, i.e., perpendicular to the interface plane. This can be accounted for by replacing the ideal one-dimensional (1D) boxlike (abrupt) potential profile with the real broadened z-dependent potential profile in the one-dimensional Schrödinger equation. In many situations, electronic wave functions and energies are not severely affected by smeared-out interfaces, and thus, their effect can be neglected or easily taken into account when designing the active-layer region [11,12]. By contrast, the interface roughness, which breaks the in-plane invariance,

*giovanni.capellini@uniroma3.it

†These authors contributed equally to this work.

has to be treated as a perturbing potential term responsible for scattering among different eigenstates [13–15]. This scattering channel is extensively modeled only for abrupt interfaces [13,14,16], with the roughness described as an in-plane $z(x,y)$ fluctuation of the interface position. This fluctuation distribution is typically characterized by its root mean square (rms), Δ, and in-plane correlation length, $\Lambda_{||}$. Recently, Valavanis *et al.* [17] generalized the IFR modeling to treat diffuse interfaces as well. Nevertheless, this model is limited to the case of a perfectly correlated roughness along the growth direction, implying that the out-of-plane fluctuations of all isocomposition surfaces, defining the diffused interface region at each (x,y) point, are equal. It is clear that this approximation is not valid when the correlation length of fluctuations along z (axial correlation length, $\Lambda_{\perp}$) is comparable to or smaller than that of the interface width $\mathcal{L}$.

Pulsed laser-assisted atom probe tomography (APT) has recently emerged as the ideal technique for evaluating the interface width $\mathcal{L}$ [18]. Furthermore, APT makes it possible to reconstruct atom-by-atom the crystal structure in three dimensions and, potentially, to quantify other properties of buried interfaces, such as the rms roughness Δ, the in-plane correlation length [19] $\Lambda_{||}$, and the axial correlation length $\Lambda_{\perp}$.

Motivated by recent interest in Ge/Si-Ge-based devices [20–22], here we show how it is possible to measure all four interface parameters ($\mathcal{L}$, Δ, $\mathcal{L}_{||}$, and now $\Lambda_{\perp}$ as well), in a sample consisting of a stack of strain-compensated $Ge/Ge_{0.8}Si_{0.2}$ asymmetric coupled quantum wells (ACQWs). Our data demonstrate the excellent interface quality, in terms of small interface width ($\mathcal{L} = 1.16$ nm) and interface rms roughness (Δ=0.18 nm). Interestingly, we find that the interface roughness is vertically correlated along the growth axis on a length $\Lambda_{\perp} = 0.26$ nm smaller than that of the interfacial width. This outcome prompts us to develop a general theory for interface roughness scattering, which accounts for both the finite correlation length along the growth direction and the interference effects associated with the interaction among different interfaces. We show that the effect of a finite axial correlation length has a large impact on the IFR scattering rate and that knowledge of the IFR at the atomic scale and its proper modeling are essential when assessing the performance of quantum devices. Indeed, we incorporate our findings into a nonequilibrium Green's function (NEGF) simulation platform [21] to model the optoelectronic properties of an *n*-type Ge/Si-Ge terahertz (THz) quantum cascade laser (QCL). We choose this device, which can be considered a stack of different ACQWs, as the ideal case study for assessing the relevance of our generalized model, given that heterointerfaces play a very important role in its operating principle. We find that the finite axial correlation length critically affects the QCL gain. Interestingly, for the set of interface parameters ($\mathcal{L}$, $\Delta\mathcal{L}_{||}\Lambda_{\perp}$), which we measure in our sample, we can predict a room-temperature operation of the device.

II. METHODS

The $Ge/Ge_{0.8}Si_{0.2}$ ACQW sample is grown by ultrahigh-vacuum chemical vapor deposition (UHV CVD) at 500 °C, using germane and silane, without carrier gases, on Si(001) substrates. The ACQW stack is deposited on top of a Si-Ge reverse-graded (RS) virtual substrate (VS), ending with a 1.5-μm-thick constant-composition fully relaxed $Ge_{0.87}Si_{0.13}$. To avoid plastic relaxation accompanied by lattice defects in the ACQW region, the composition of the RG VS is chosen to achieve strain-compensation conditions in the ACQW, featuring tensile-strained $Ge_{0.8}Si_{0.2}$ barriers and compressively strained Ge well layers. Each of the 20 modules is designed to have a thick ($d_{TW} = 12$ nm) and a narrow ($d_{NW} = 5$ nm) Ge well, coupled to a d_{TB}=2.3 nm $Ge_{0.8}Si_{0.2}$ tunnel barrier. Another d_{SB}=20-nm-thick $Ge_{0.8}Si_{0.2}$ layer is used to separate each module, so that the nominal period thickness of the structure is 39.3 nm. More details about the growth and X-ray diffraction structural characterization of ACQW samples are reported in Ref. [23]. Sample preparation for high-resolution scanning transmission electron microscopy (HR STEM) analysis and APT analysis is done using a Helios Nanolab 650 dual channel (Ga^+ ion column for milling, SEM column for imaging) focused-ion beam (Dual-FIB) microscope, using the standard lamella lift-out technique. HR STEM analysis is conducted using a double cross-section-corrected FEI Titan microscope, operated at 200 kV. The beam converge angle is 19.1 mrad. Aberration-corrected magnetic lenses help to make the probes of the order of 1–2 Å in diameter with beam currents of 200 pA. A CEOS CESCOR corrector is used to yield a resolution of 0.8 Å. The images are recorded using a high-angle annular dark field (HAADF) detector, and the data are processed using digital micrograph GMS3 software. As for the APT experiments, a stable and layer-by-layer evaporation of atoms is achieved by focusing a picosecond-pulsed ultraviolet laser on the FIB-defined nanotips. The laser energy, pulse frequency, and evaporation rate (ions/pulse) are maintained at 5.0–10.0 pJ, 250–500 kHz, and 0.5–1.5, respectively. The constant shank-angle-based 3D reconstructions are performed using the Cameca's integrated visualization and analysis software (IVAS) package. In an APT setup, atoms are evaporated as cations, in layer-by-layer fashion, starting from the apex of the tip, while the software rearranges these atoms in the same sequence as they are evaporated (detected) to create the 3D atomistic reconstruction.

The NEGF simulations are performed using the nextnano.QCL package, which calculates quantum transport through layered heterostructures. The scattering processes are treated within the self-consistent Born

approximation, including phonons, charged impurities, electron-electron, interface roughness, and alloy disorder [21,24,25]. We neglect hot-phonon effects, assuming Bose-Einstein statistics at the lattice temperature. Moreover, aiming to reduce the computational time, the electron-electron interaction is calculated by exploiting a mean-field approximation, as discussed in Ref. [21], where the interested reader can find a detailed description of the model and its approximations. The gain is calculated self-consistently within the linear response theory following Ref. [26]. The graded composition profile of the QCL is calculated by convoluting the nominal squared profile with a Gaussian function of full width at half maximum (FWHM) $\mathcal{L}$ [11]. As detailed in the Appendixes, the interface roughness self-energy is derived according to the present generalized theory of interface roughness.

III. RESULTS AND DISCUSSION

A. Experimental determination of interface parameters in Ge/Si-Ge ACQWs

The HAADF STEM image acquired on the ACQW stack is shown in Fig. 1(a). Images recorded at progressively higher magnifications are highlighted by color-coded rectangular boxes and are displayed as insets. We can clearly distinguish stacking along the growth direction of 20 identical periods separated by a $Ge_{0.8}Si_{0.2}$ layer (darker gray). Line scans (averaged over 100 pixels) are recorded using the GSM3 software to extract the intensity variation in the HAADF STEM images across the entire thickness of the ACQW stack. We measure the mean thicknesses of the wells and barriers, by fitting the intensity profile with Gaussian functions (not shown), to be $d_{SB} = 19.60\,\text{nm}$ ($\pm 1.3\,\%$) for the Si-Ge spacer barrier;

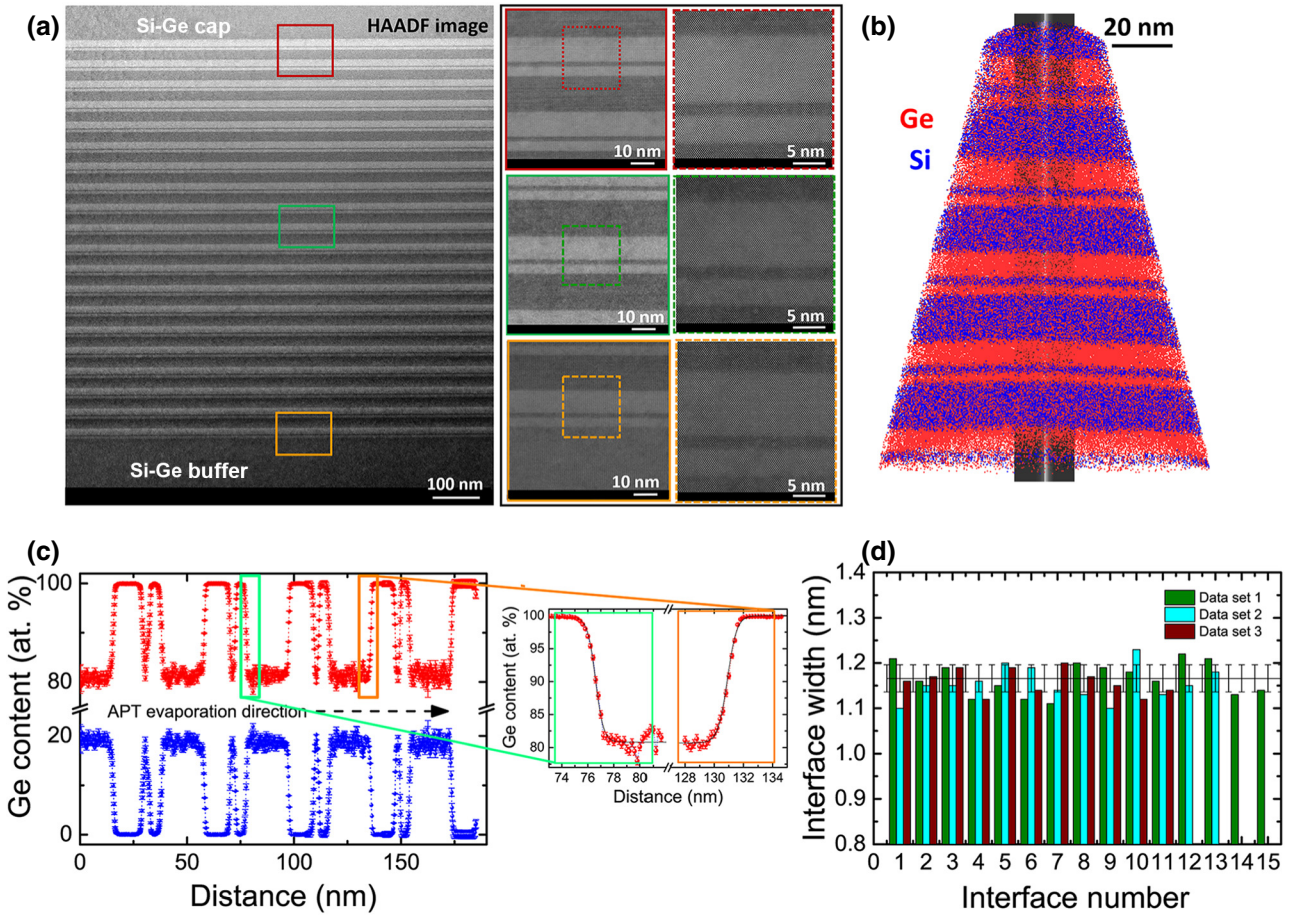

FIG. 1. (a) Left: HAADF STEM image showing all 20 periods of the Ge/Si-Ge ACQW, parts of the Si-Ge capping layer, and the Si-Ge buffer in the VS. Right: images recorded at progressively higher magnification from different sample regions. (b) 3D atom-by-atom reconstruction of the Ge/Si-Ge ACQW (data set 1). (c) 1D concentration profile of Si and Ge (at a fixed bin width of 0.2 nm) along the direction of APT evaporation. Data are extracted from the 30-nm-wide cylinder coaxial to the tip and highlighted in black in (b). Inset: Details of falling (green) and rising (orange) Ge concentration profiles. Black solid lines in the insets represent the fitting function [Eq. (1)] used to extract the interface widths. (d) Extracted values of $\mathcal{L}$ for all interfaces and three different APT data sets of the sample. Mean value of $\mathcal{L}$ is shown as black horizontal line together with error bars representing the standard deviation of the measurements on all interfaces.

$d_{TW} = 12.22$ nm ($\pm 1.4\%$) and $d_{NW} = 5.15$ nm ($\pm 1.2\%$) for the thick and narrow Ge wells, respectively; and $d_{TB} = 2.48$ nm ($\pm 1.2\%$) for the Si-Ge tunneling barrier, resulting in an average total period thickness of 39.5 nm($\pm 5.1\%$) (the uncertainties represent the standard deviation of the thicknesses measured over 20 periods of the ACQW). HAADF STEM analysis confirms, within experimental uncertainty, the thickness reproducibility of all individual layers along the stack and the good match to nominal values.

Furthermore, the absence of any misfit dislocations and extended defects in the HAADF STEM images of Fig. 1(a) highlights the good crystalline quality of the sample and the coherent nature of the heterointerfaces.

The atom-by-atom 3D reconstruction of four periods of the sample (data set 1 of 3) is shown in Fig. 1(b), with atom evaporation proceeding from the top to the bottom of the tip, i.e., in the opposite direction with respect to the growth direction. The second of the three APT data sets is shown in Fig. S1 within the Supplemental Material [27], along with a representative mass spectrum. To ensure higher accuracy of the atomic scale, only atoms evaporated from a 30-nm-wide cylinder coaxial to the tip [see Fig. 1(b)] are considered in the analysis. In fact, the uniformity in the layer-by-layer evaporation is maximum for atoms located at the center of the hemispherical tip [28].

Figure 1(c) shows the 1D concentration of Si and Ge atoms along the APT evaporation direction. The mean Ge concentration within the Si-Ge barrier layers is found to be 80.0 %($\pm 0.6\%$), which is in perfect agreement with the nominal value. The 1D concentration profile is used to estimate the interface width $\mathcal{L}$. To this aim, in the inset of Fig. 1(c), we show a magnified view of the Ge concentration profile at two Ge→Si-Ge and Si-Ge→Ge heterointerfaces, corresponding to the marked green and orange rectangular boxes, respectively, in the plot of the 1D concentration profile. The raw data are fitted using the error function [11]

$$c(z) = c_0 + d_0 \mathrm{erf}\left[2\sqrt{\ln(2)}(z - z_0)/\mathcal{L}\right], \qquad (1)$$

where c_0 and d_0 are an offset and scale parameters, respectively, which are introduced to achieve the correct concentrations on both side of the heterointerface, and z_0 defines the position of its center. The interfacial width $\mathcal{L}$ is defined here as the FWHM of the derivative of the fitted concentration profile and corresponds to the interface width over which the concentration changes from 12% to 88% of the plateau value. For the two examples of the rising and falling Ge concentration profiles shown in the inset of Fig. 1(c), the values of $\mathcal{L}$ extracted from the fit are 1.16 and 1.15 nm, respectively.

Figure 1(d) shows the $\mathcal{L}$ values measured from all interfaces present in the three investigated data sets, resulting in an average value of 1.16 nm ($\pm 3.7\%$). Since often in the literature a sigmoidal function of the concentration profile is used as an alternative fitting profile, we also fit our data to the function $c_s(z) = c_0 + d_0/[1 + e^{-(z_0 \pm z)/L}]$, finding an average value of L=0.288 nm. The quantity $4L = 1.15$ nm corresponds to the length over which the concentration changes from 12% to 88% of the plateau value of the logistic function and can be compared to the interface width $\mathcal{L}$. Remarkably, the interface width obtained here is only slightly larger than that measured from low Ge content Si/Si-Ge multilayers in Ref. [18], using the same technique. The value of $\mathcal{L}$ we report here is also more than a factor of about 2.5 times smaller than those measured in $Ge/Si_{0.2}Ge_{0.8}$ multiquantum well samples grown by reduced pressure CVD and in plasma-enhanced CVD reactors [29,30]. This suggests that UHV CVD might be better suited for realizing sharper interfaces in high-Ge-content heterostructures with layer thicknesses in the order of a few nanometers. Interestingly, we do not observe any systematic difference between the measured value of $\mathcal{L}$ on the Ge→Si-Ge and Si-Ge→Ge interfaces. This observation differs from that in previous reports on both low- [18,31] and high-Ge-content [30] Si/Si-Ge heterostructures. In fact, in these early reports, larger widths are measured for the interface featuring a Ge content decreasing along the growth direction with respect to the opposite case, as one expects considering the tendency to surface segregation of Ge atoms during the Si overlayer deposition. We speculate that this discrepancy may be related to specific peculiarities of the UHV CVD growth process, such as the relatively short precursor residential time in the reactor. Finally, as evident from Fig. 1(d), we notice that $\mathcal{L}$ is also independent of both the thickness of the layers and the position within the whole stack. This evidence suggests that the adopted growth temperature of 500 °C does not induce relevant intermixing in buried interfaces during the subsequent steps of the growth process [32].

After having characterized the interface width, we now exploit the fully 3D character of our analysis for assessing the roughness Δ and the coherence along both the parallel and axial directions. To this aim, from the 3D atomic distribution shown, for instance, in the left plot of Fig. 2(a), we preliminarily determine the isocompositional surfaces at a Ge concentration of 90%, i.e., halfway between the barrier and well compositions [right plot in Fig. 2(a)]. Their average positions along the growth direction define a set of z_0 values, each of them corresponding to the position of a given heterointerface. This allows the calculation of the height fluctuation $h(\vec{\rho})$ of the 90% isocompositional surfaces, with respect to their z_0, as a function of the in-plane coordinates $\vec{\rho} = (x, y)$. As an example, we report in the inset of Fig. 2(b) the $h(\vec{\rho})$ distribution of one 90% isocompositional interface [blue arrows in the right plot of Fig. 2(a)]. The height-height correlation function $H(\tau)$ is then experimentally determined as the squared difference of $h(\vec{\rho})$ calculated at two points separated by the in-plane

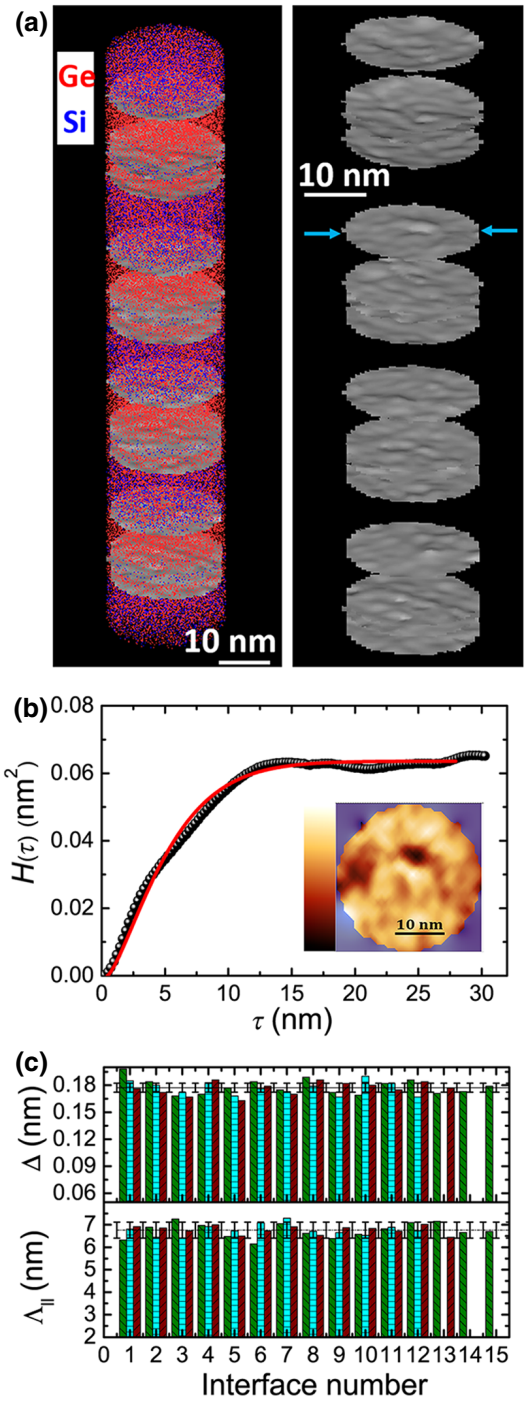

FIG. 2. (a) Left: 3D reconstruction of all atoms located within the black cylinder of diameter 30 nm, shown in Fig. 1(b). Right: Corresponding 90% Ge content isocompositional surfaces between Ge and Si-Ge layers. (b) Height-height correlation function $H(\tau)$ as a function of the length of the in-plane vector $\vec{\tau}$ measured on the interface marked by the arrow in (a). Inset: Color-coded height distribution of the same interface: color bar range is $[-0.38; +0.47]$ nm. (c) Extracted values of Δ and $\Lambda_\parallel$ of all interfaces, for the three different APT data sets. We plot the mean values of the measurement parameters as black horizontal lines, together with error bars representing the standard deviation of the measurements on all interfaces.

vector $\vec{\tau}$ and averaged in the in-plane direction

$$H(\tau) = \left\langle |h(\vec{\rho}) - h(\vec{\rho}+\vec{\tau})|^2 \right\rangle_{\vec{\rho}} = 2\left\langle |h(\vec{\rho})|^2 \right\rangle_{\vec{\rho}} - 2\left\langle |h(\vec{\rho})h(\vec{\rho}+\vec{\tau})|^2 \right\rangle_{\vec{\rho}}. \quad (2)$$

In the right-hand side of Eq. (2), the first term represents the IFR mean squared roughness $\Delta^2 = \left\langle |h(\vec{\rho})|^2 \right\rangle_{\vec{\rho}}$. Under the assumption of in-plane rotation and translation invariance, the second term depends only on the modulus of $\vec{\tau}$, and then can be written as

$$\left\langle |h(\vec{\rho})h(\vec{\rho}+\vec{\tau})|^2 \right\rangle_{\vec{\rho}} = \Delta^2 C_\parallel(\tau), \quad (3)$$

where the dimensionless quantity $C_\parallel(\tau)$ is usually referred to as the IFR in-plane correlation function.

In Fig. 2(b), we plot the functional behavior of $H(\tau)$ associated with the exemplificative surface shown in its inset. By assuming a Gaussian decay of the in-plane correlation function [13–17,19,33],

$$C_\parallel(\tau) = e^{-(\tau/\Lambda_\parallel)^2}, \quad (4)$$

from Eq. (4), we can estimate the Δ and $\Lambda_\parallel$ parameters by fitting the measured $H(\tau)$ to the function $H_{\text{fit}}(\tau) = 2\Delta^2[1 - e^{-(\tau/\Lambda_\parallel)^2}]$. In this case, we obtain $\Delta = 0.18$ nm and $\Lambda_\parallel = 6.98$ nm.

The values of Δ and $\Lambda_\parallel$ obtained for all of the 90% isocompositional surface across the different APT data sets are highly uniform, as evident in Fig. 2(c). The mean values of Δ and $\Lambda_\parallel$ are found to be 0.18 nm ($\pm 5\%$) and 6.9 nm ($\pm 4\%$), respectively.

On the other hand, as already pointed out, for a complete characterization of IFR in diffuse interfaces, the measurement of the roughness correlation length along the growth direction is also required. Therefore, we set up a procedure for the quantification of the axial correlation length $\Lambda_\perp$ from APT data.

For this purpose, we "slice" the diffuse interfaces into seven isoconcentration surfaces, which are defined at different Ge concentrations in the 87%–93% range (outside this range of concentration values, no meaningful isocompositional surfaces can be obtained). For each isocompositional surface, we calculate their average positions, $z_i = 1, \ldots, 7$, obtaining values in perfect agreement with the 1D composition previously displayed in Fig. 1(c).

Subsequently, the height function $h_i(\vec{\rho})$ associated with each isocompositional surface is calculated and used to generate the differential maps $w_{ij}(\vec{\rho})$ defined by

$$w_{ij}(\vec{\rho}) = h_i(\vec{\rho}) - h_j(\vec{\rho}), \quad (5)$$

where the (i,j) indices correspond to isocompositional surface pairs belonging to the same diffuse interface. Their values are chosen to have a data set featuring increasing values of $|\Delta z_{ij}| = |z_i - z_j|$, corresponding to the following couples of isocompositional surfaces: $(i,j) = (90.5\%, 90\%)$, (91%, 90%), (91%, 89%), (92%, 88%), and (93%, 87%). In Fig. 3(a), we show the measured values, $\Delta_{ij} = \sqrt{\langle {w_{ij}}^2(\vec{\rho})\rangle}$, of the rms roughness of the $w_{ij}(\vec{\rho})$ maps for all interfaces across the different APT data sets for (i,j) corresponding to (91%, 89%). As an example, in Fig. 3(b), we show representative differential surface maps $w_{ij}(\vec{\rho})$ for this (i,j) pair [the Δ_{ij} values extracted from all available $w_{ij}(\vec{\rho})$ maps for other (i,j) values are shown in Fig. S2 within the Supplemental Material [27]]. From these data,

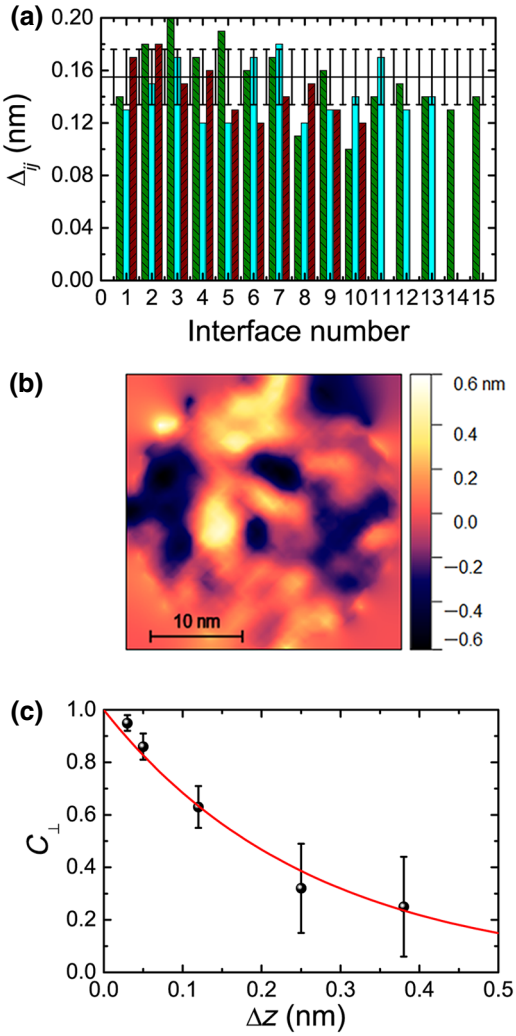

FIG. 3. (a) Measured Δ_{ij} of the $w_{ij}(\vec{\rho})$ differential map with $i = 91\%, j = 89\%$ calculated for all interfaces across the different APT data sets. (b) Representative differential map $w_{ij}(\vec{\rho})$ for $i = 91\%$ and $j = 89\%$ isocompositional surfaces. (c) Data points (spheres) for the axial correlation function fitted to an exponential decay function $C_\perp(\Delta z) = e^{-\Delta z/\Lambda_\perp}$ (line).

we can determine the axial correlation function

$$C_\perp(|\Delta z|) = \langle h(\vec{\rho}, z) h(\vec{\rho}, z + \Delta z)\rangle_{\vec{\rho}} / \Delta^2. \tag{6}$$

To this aim, we write the measured ${\Delta_{ij}}^2$ as

$$\begin{aligned} {\Delta_{ij}}^2 &= \langle {w_{ij}}^2(\vec{\rho})\rangle = \left\langle [h_i(\vec{\rho}) - h_j(\vec{\rho})]^2 \right\rangle \\ &= \langle {h_i}^2(\vec{\rho})\rangle + \langle {h_j}^2(\vec{\rho})\rangle - 2\langle h_i(\vec{\rho})\, h_j(\vec{\rho})\rangle \\ &= 2\Delta^2[1 - C_\perp(|z_i - z_j|)], \end{aligned} \tag{7}$$

where we assume that the considered isocompositional surfaces feature the same Δ value. This relation is used to plot the experimental $C_\perp(\Delta z_{ij})$ data points in Fig. 3(c). These values are fitted to an exponential decay $e^{-\Delta z/\Lambda_\perp}$ function from which we estimate the value $\Lambda_\perp$=0.26 nm(±7.5%). Remarkably, this quantity is smaller than that of the mean interface width $\mathcal{L} = 1.16$ nm, meaning that the usually made assumption of fully correlated roughness within each diffuse interface is not valid [17]. Based on this result, the next section elaborates a general theoretical framework to calculate the impact of a finite value of $\Lambda_\perp$ on the IFR scattering rate.

B. Theory of IFR scattering rate for diffuse interfaces with finite axial correlation length

The impact of roughness on electron states in multilayer structures is usually theoretically addressed in the framework of perturbation theory, introducing the perturbing potential $\delta V_{\mathrm{IFR}}(\vec{\rho}, z)$, which represents the difference between the total 3D potential $V(\vec{\rho}, z)$ and the 1D average potential $\bar{V}(z) = V(\vec{\rho}, z)_{\vec{\rho}}$, where $\langle\rangle_{\vec{\rho}}$ denotes averaging over the in-plane coordinate $\vec{\rho}$. Hence, $\delta V_{\mathrm{IFR}}(\vec{\rho}, z)$ represents fluctuations of the potential that break the in-plane translational invariance [34]. Consequently, $\delta V_{\mathrm{IFR}}(\vec{\rho}, z)$ can couple eigenstates of the unperturbed Hamiltonian featuring different in-plane vectors and z-dependent envelope functions, i.e., $\varphi_\alpha(z)e^{i\vec{k}.\vec{\rho}}$ and $\varphi_\beta(z)e^{i(\vec{k}+\vec{q}).\vec{\rho}}$.

According to the first-order perturbation theory, the rate of elastic scattering events from the initial state $|\alpha, \vec{k}\rangle$ to any available final state belonging to the β subband is given by

$$\hbar\Gamma_{\alpha,\beta,k} = 2\pi \sum_{\vec{q}} |\delta V_{\alpha\beta q}|^2 \delta\left(E_{\alpha\beta} + \frac{\hbar^2 k^2}{2m_\parallel^*} - \frac{\hbar^2|\vec{k} + \vec{q}|^2}{2m_\parallel^*}\right), \tag{8}$$

where k and q are the modulus of the initial and exchanged momentum, respectively; $m_\parallel^*$ is the in-plane effective mass; $E_{\alpha\beta} = E_\alpha - E_\beta$ is the subband energy separation; and $\delta V_{\alpha\beta q} = \langle \alpha, \vec{k}\, |\delta V_{IFR}| \beta, \vec{k} + \vec{q}\rangle$ is the IFR scattering matrix element, which is a k-independent quantity. Assuming (i)

abrupt interfaces ($\mathcal{L} \to 0$), (ii) the absence of correlation among the different heterointerfaces, and (iii) a Gaussian shape of the two-point correlation function for $h(\vec{\rho})$, this matrix element is (see Refs. [13,14])

$$|\delta V^{(0)}_{\alpha\beta q}|^2 = \pi \Delta^2 \Lambda_\parallel^2 e^{-\frac{\Lambda_\parallel^2 q^2}{4}} V_0^2 \sum_\eta |\varphi_\beta(z_0^{(\eta)})|^2 |\varphi_\alpha(z_0^{(\eta)})|^2. \tag{9}$$

In Eq. (9), the sum extends over the different uncorrelated interfaces η located along the growth direction at z_0^η, and V_0 is the value of the associated band offset, which is assumed to be constant for all interfaces. The resulting physical picture is that the IFR perturbation can be regarded as a contact potential active at the interface planes, and therefore, sensitive to the local amplitude of the wave functions. The square modulus of its matrix element decays exponentially with the square modulus of the exchanged momentum q, with a characteristic scale given by $1/\Lambda_\parallel$.

As already mentioned in Sec. I, Valavanis *et al.* [17] relaxed the hypothesis of interface abruptness with the scope of investigating Si-Ge multilayer systems, which typically feature non-negligible interface widths $\mathcal{L}$, in the 0.9–1.5 nm range [18,31]. Remarkably, the relevance of proper modeling of diffuse interfaces has recently emerged also in the realm of III-V material systems [11,35]. In the absence of experimental inputs, Valavanis *et al.* assume that (i) the roughness is fully correlated along the growth direction within each diffuse interface, (ii) the separation between interfaces is much larger than that of the interdiffusion length, and (iii) the IFRs of different interfaces are fully uncorrelated (i.e., $C_\perp(|z_0^{(\eta)} - z_0^{(\eta')}|) = 0$ for $\eta \neq \eta'$). Within these assumptions, the interface broadening is found to have a very limited impact on the IFR scattering rate up to $\mathcal{L} \leq 3$ nm. However, we point out here that not only may assumption (i) not hold ubiquitously, as directly demonstrated by our experimental findings, but also that assumptions (ii) and (iii) may not be applicable when studying tunneling structures with a barrier thickness comparable to the interface width and/or to the axial correlation length. This implies that in many multilayer materials embedded in real-word devices these working hypotheses are not satisfied.

To overcome these limitations and based on the experimental observations outlined in Sec. III A, we develop here a more general theoretical framework to describe the IFR scattering in a generic sequence of diffuse interfaces with a finite correlation length along the axial direction. In the presence of diffuse interfaces, the axial correlation can couple the perturbing potential associated with different interfaces and, as a consequence, their joint action cannot be described in terms of a sum of independent scattering sources, as in Eq. (9). As detailed in the Appendix A, in this general case, the scattering rate depends on the 3D correlation function of the interface roughness. Assuming that this function can be factorized into in an in-plane and an axial part, we obtain

$$|\delta V_{\alpha\beta q}|^2 = \pi \Delta^2 \Lambda_\parallel^2 e^{-\frac{\Lambda_\parallel^2 q^2}{4}} \int dz_1 \int dz_2 \; \varphi^*_\beta(z_1)\varphi_\alpha(z_1) \times \varphi^*_\alpha(z_2)\varphi_\beta(z_2) \frac{\partial \bar{V}}{\partial z}(z_1) \frac{\partial \bar{V}}{\partial z}(z_2) C_\perp(|z_2 - z_1|), \tag{10}$$

where the z integrals extend over the whole heterostructure stack. As such, we notice that Eq. (10), being formulated in terms of a generic potential profile $\bar{V}(z)$ defined throughout the sample, can be applied without artificially describing $\bar{V}(z)$ as a set of different contributions coming from partially overlapping "blurred" interfaces [36].

To gain an insight into the impact of the correlation effect due to a finite value of $C_\perp$, we calculate the scattering rate $\Gamma_{\alpha\beta k}$ from an initial state $|\alpha, k\rangle$ to a different subband, which is obtained by summing over all available wave vectors $\overrightarrow{k_f} = \overrightarrow{k} + \vec{q}$ in the final subband

$$\Gamma_{\alpha\beta k} = \frac{m^*_\parallel}{\pi \hbar^3} \int_0^\pi d\theta |\delta V_{\alpha\beta q}|^2, \tag{11}$$

where we take into account energy conservation to express the modulus of the exchanged wave vector $\overrightarrow{q}$, as a function of the scattering angle θ only.

We now disregard the correlation that possibly exists between different interfaces, considering one heterointerface only or, alternatively, we can suppose that the axial correlation function $C_\perp$ is negligible when calculated at the minimum heterointerface distance featured by the multilayer stack. Under the additional hypothesis that the variations of the wave functions at the interface width scale are small, it can be easily proven (see Appendix B) that $\Gamma_{\alpha\beta k} = F\Gamma^{(0)}_{\alpha\beta k}$ where

$$F = \frac{1}{V_0^2} \int dz_1 \int dz_2 \; \frac{\partial \bar{V}}{\partial z}(z_1) \frac{\partial \bar{V}}{\partial z}(z_2) C_\perp(|z_2 - z_1|), \tag{12}$$

and $\Gamma^{(0)}_{\alpha\beta k}$ is the scattering rate calculated according to Eq. (9). Aiming at a quantitative estimation of the F value, we calculate the right-hand side of Eq. (12), in the case of an exponentially decaying axial correlation function and assuming an error-function profile for the diffuse interface, as suggested by the fit of our APT experimental data. Under these conditions, we obtain that F is controlled by the dimensionless parameter $\mathcal{L}/\Lambda_\perp$, according to

$$F = \exp\left(\frac{\mathcal{L}^2}{16\ln(2)\Lambda_\perp^2}\right)\left[\mathrm{erf}\left(-\frac{\mathcal{L}}{4\ln(2)\Lambda_\perp}\right) + 1\right]. \tag{13}$$

Interestingly, for our measured values of $\mathcal{L} = 1.16$ nm and $\Lambda_\perp = 0.26$ nm, we find F=0.35, which means that the IFR scattering rate is reduced by a factor of about 3 times, with respect to the value predicted for an abrupt interface with the same rms roughness.

We can also define an equivalent rms roughness for abrupt interfaces as $\Delta_{eq}^2 = \Delta^2 F$. The value $F = 0.35$ calculated in this work gives an equivalent rms roughness of $\Delta_{eq} = 0.106$ nm, instead of $\Delta = 0.18$ nm.

This reduced effectiveness of the IFR potential can be explained by considering that the interface does not behave as a single scattering center, but, instead, acts as a collection of different, only partially correlated, scattering centers. We notice that, in line with Eq. (12), $F \sim 0$ when $\Lambda_\perp \ll \mathcal{L}$. On the contrary, for $\Lambda_\perp \gg \mathcal{L}$, $F \sim 1$ and, consequently, the abrupt interface limit is recovered.

C. Impact of IFR scattering rate on the performances of a QCL device

As previously discussed in Sec. I, a thorough theoretical and experimental understanding of heterointerfaces can greatly help in designing and modeling innovative devices. As a relevant example, we discuss here the impact of IFR scattering on the performances of an *n*-type Ge/Si-Ge quantum cascade laser grown on top of a Si(001) substrate, featuring a composition profile quite similar to that of the ACQW samples investigated here. This choice is motivated by the fact that QCL systems are a prototypical case study to assess the role of IFR scattering [6,37,38], owing to the inherently large number of heterointerfaces featured by this class of device, and by the technological relevance of a Si-based THz source, suitable for room-temperature operation [20,21].

In Fig. 4, we show the *L*-point band-edge energy profile together with the relevant states, calculated in the envelope function approximation, considering both abrupt (dashed lines) and diffuse (solid lines) interfaces, for which, guided by our experimental results, we assume $\mathcal{L} = 1.16$ nm. From Fig. 4, we predict that this moderate degree of interdiffusion has a very limited impact on the wave functions of the low-energy states controlling the QCL carrier dynamics. Actually, to recover the same optical properties predicted for the abrupt heterointerface condition, the effect of interface broadening on the gain can be compensated for by an appropriate fine-tuning of the layer thicknesses and compositions (see Fig. S3 within the Supplemental Material [27]).

In Fig. 5(a), we report the IFR scattering rate from an initial state in the ULL subband with zero in-plane momentum to the LLL subband calculated as a function of $\Lambda_\parallel$, setting the values $\mathcal{L} = 1.16$ nm, $\Lambda_\perp = 0.26$ nm, and $\Delta = 0.18$ nm, which are obtained from APT data. The scattering rate tends to zero when both $\Lambda_\parallel = 0$ and $\Lambda_\parallel$ has a maximum of $\Lambda_\parallel = 5.6$ nm, which is unfortunately quite close to the value of 6.9 nm measured in our ACQW samples (vertical dashed line). To understand the reason for the occurrence of a peak in Fig. 5(a) at about 5.6 nm, we note that for an in-plane mass of 0.32 m_0 (associated with the effective mass tensor of Ge(001) at the *L* point) and an energy separation between the two laser levels of 16 meV, the modulus of the wave vector exchanged in the IFR elastic scattering process is $q \sim 0.35$ nm^{-1}. Considering now the functional dependence on $\Lambda_\parallel$ in Eq. (10), the scattering rate is expected to have a maximum when $2/\Lambda_\parallel = q$, corresponding to $\Lambda_\parallel = 5.7$ nm when the exchanged momentum, q, is 0.35 nm^{-1}, in great agreement with the observed peak value in Fig. 5(a). The results in Fig. 5(a) suggest that a reduction of the IFR scattering rate can be achieved by mismatching q and $2/\Lambda_\parallel$ scales: to this aim, one can either act on the laser-level separation or modify the in-plane correlation length at the growth stage, which is clearly a more challenging task.

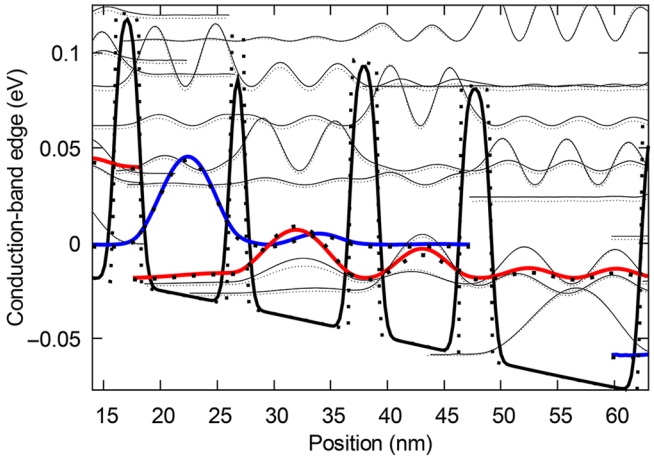

FIG. 4. *L*-valley band-edge and electronic states of a THz *n*-type $Ge/Ge_{0.75}Si_{0.25}$ QCL with abrupt interfaces (dashed lines) and with interdiffusion (solid lines). Squared wave functions associated with the upper (ULL, blue) and lower laser levels (LLL, red) are highlighted.

The dependence of the IFR scattering rate between the two laser levels on the axial correlation length obtained from our model, setting $\mathcal{L} = 1.16$ nm, $\Lambda_\parallel = 6.9$ nm, and $\Delta = 0.18$ nm, as experimentally determined, is displayed by the blue curve in Fig. 5(b); the dashed vertical line is drawn at $\Lambda_\perp = 0.26$ nm, as in our sample. For comparison, the horizontal lines represent the IFR scattering rate obtained using the abrupt interface model of Eq. (9) (red line) and the perfectly correlated diffuse interface model developed by Valavanis *et al.* [17] (green line), both neglecting possible coupling among different interfaces. Notably, the deviation of the IFR scattering rate obtained from the Valavanis model, with respect to that predicted by the abrupt interface model, is due to the merging of the rising and falling concentration profiles in

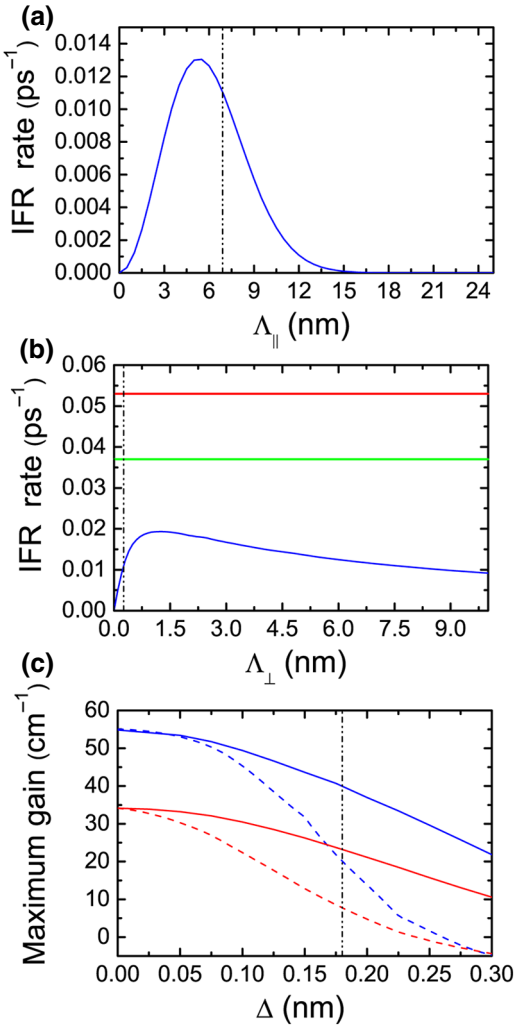

FIG. 5. IFR scattering rate (blue curve) between the QCL laser levels calculated with our model (a) as a function of in-plane correlation length $\Lambda_{\parallel}$ setting $\mathcal{L} = 1.16$ nm, $\Lambda_{\perp} = 0.26$ nm, and $\Delta = 0.18$ nm; and (b) as a function of axial correlation length $\Lambda_{\perp}$ setting $\mathcal{L} = 1.16$ nm, $\Lambda_{\parallel} = 6.9$ nm, and $\Delta = 0.18$ nm. In (b), reference values corresponding to the IFR scattering rate calculated with the same values of $\Lambda_{\parallel}$and Δ using the abrupt interface model ($\mathcal{L} = 0$) and the Valavanis model for diffuse interfaces [17] ($\mathcal{L} = 1.16$ nm) are displayed in red and green, respectively. (c) Maximum QCL gain as a function of the rms interface roughness Δ for lattice temperatures of 10 K (blue) and 300 K (red), by taking into account interdiffusion (solid lines) or neglecting it (dashed lines). Calculations are performed with $\mathcal{L} = 1.16$ nm, $\Lambda_{\parallel} = 6.9$ nm, and $\Lambda_{\perp} = 0.26$ nm. In (a)–(c), vertical lines mark the values measured by APT on our samples ($\Lambda_{\parallel} = 6.9$ nm, $\Lambda_{\perp} = 0.26$ nm, $\Delta = 0.18$ nm).

the diffuse interfaces defining the tunnel barrier located between the ULL and the LLL, which lowers the potential barrier height, and hence, slightly suppresses the scattering rate with respect to the atomically sharp case (see dashed and continuum band-edge profiles in Fig. 4). For the QCL structure considered here, this deviation emerges despite $\mathcal{L} < 3$ nm, and thus, in a range where, according to Ref. [17], the impact of interface broadening on the IFR scattering rate should be negligible. This apparent contradiction is because our QCL structure features thin tunneling barriers, with a thickness in the order of $\mathcal{L}$, and thus, does not satisfy the assumption of interface separations larger than $\mathcal{L}$ considered in Ref. [17].

Our model (blue curve) indicates that the IFR scattering rate, in the range explored in Fig. 5(b), has a nonmonotonic behavior, which is an increasing (decreasing) function for small (large) values of $\Lambda_{\perp}$, with a maximum value achieved at $\Lambda_{\perp} \sim 1.2$ nm. To understand this functional behavior, we first notice that a multilayer system is characterized by two length scales: (i) the interface width $\mathcal{L}$ and (ii) the minimum distance ($d_{\min}$) separating two distinct interface planes featuring non-negligible amplitudes of both interacting wave functions. In our QCL design and for the considered states, $d_{\min}$ corresponds to the thickness of the tunneling barrier separating the two wells that host the laser levels ($d_{\min} = 1.4$ nm). This value is approximately equal to $\mathcal{L} = 1.16$ nm (see Fig. 4). Consequently, the two length scales in our case are practically coincident.

To understand the rising part of the blue curve, we observe that, for $\Lambda_{\perp} << d_{\min}$, the coupling among different interfaces is negligible, and therefore, the double spatial integral of Eq. (10) can be split into a discrete sum of terms, each representing the contribution from an isolated interface, analogously to the case of Eq. (9). In other words, in this regime, $C_{\perp}(|z_2 - z_1|)$ is not negligible only when z_1 and z_2 are in the range of neighbors of the same interface. Consequently, the spatial derivative $(\partial V/\partial z)(z_1)(\partial V/\partial z)(z_2)$ and the products of the two wave functions $\varphi_{\alpha}(z_1)\varphi^*_{\alpha}(z_2)$ and $\varphi_{\beta}(z_1)\varphi^*_{\beta}(z_2)$ entering the integral in Eq. (10) have positive values.

This condition, together with the monotonic increase of $C_{\perp}(|z_2 - z_1|)$ with $\Lambda_{\perp}$, produces the rise of the IFR scattering rate for $\Lambda_{\perp} << d_{\min}$. When $\Lambda_{\perp}$ approaches $d_{\min}$, the correlation between interfaces starts to become relevant, since mixed terms in the double spatial integral, associated with the coupling of different interfaces, are no longer suppressed by the $C_{\perp}(|z_2 - z_1|)$ factor. Interference effects between the different interfaces then come into play, with a positive or negative contribution that, in general, depends on the particular potential shape and on the considered wave functions. In the case of the QCL of Fig. 4, considering the IFR scattering between the ULL and the LLL subbands, the interference effects are mainly due to the interaction between the two interfaces that define the tunneling barrier at a position of about 27 nm. In such a case, when z_1 and z_2 belong to the two consecutive interfaces, the product $(\partial V)/(\partial z)(z_1)(\partial V)/(\partial z)(z_2)$ is a negative number, while the $\varphi_{\alpha}(z_1)\varphi^*_{\alpha}(z_2)\varphi_{\beta}(z_1)\varphi^*_{\beta}(z_2)$ product has a positive value. This explains why an increase of the axial correlation $\Lambda_{\perp}$beyond 1.2 nm triggers a decrease of the IFR rate. Notably, the IFR scattering rate predicted by

our model never reaches the limiting value of the Valavanis model (green curve), since the two assumption of the Valavanis model, i.e., (i) $\Lambda_\perp \ll d_{\min}$ (uncoupled interfaces) and (ii) $\Lambda_\perp >> \mathcal{L}$ (perfect correlation within each interface), cannot be simultaneously fulfilled in our QCL design for which $\mathcal{L} \approx d_{\min}$. Remarkably, at the value $\Lambda_\perp = 0.26$ nm experimentally evaluated on the grown structure, we obtain an IFR scattering rate lower by a factor of about 3 times with respect to the value predicted by the Valavanis model. It is interesting to note that this suppression factor matches the analytical estimation of the F value in Eqs. (12) and (13), which results in a value of 0.35 for the IFR parameters measured on our sample.

The main message of Fig. 5(b) is thus that the parameter $\Lambda_\perp$ has a crucial effect on the IFR rate and should be properly considered when designing tunneling heterostructure devices. In fact, two effects simultaneously affect scattering in tunneling structures: the imperfectly correlated roughness within the single interface and the coupling of the IFR potential associated with different interfaces. Therefore, for fixed $\Lambda_\perp$, changes in the multilayer thicknesses and/or in the wave function profile can result in relevant variation of the IFR scattering rate.

Finally, for the QCL structure of Fig. 4, we quantitatively compare the maximum gain as a function of Δ calculated using the generalized model of IFR scattering described here (solid lines) with that predicted by the abrupt uncorrelated interface approach (dashed lines), as done in Refs. [21] and [39] [Fig. 5(c)]. For each model, the gain is obtained from NEGF simulations at lattice temperatures of 10 K (blue) and 300 K (red). Again, the calculations are performed by setting the experimentally determined interface parameters for $\mathcal{L}$, $\Lambda_\parallel$, and $\Lambda_\perp$. We observe that, for the case of our model with diffuse interfaces, the gain is much more robust against an increase of Δ, due to the diminished detrimental impact of the IFR scattering rate, as noted when discussing Fig. 5(b). Interestingly, for the APT measured value of $\Delta = 0.18$ nm (vertical dashed line), our model predicts gain values more than double that of the abrupt case. To appreciate the relevance of this result, we observe that, for a realistic value of optical losses (i.e., waveguide plus facet losses) of about 20 cm^{-1} [21], laser action at room temperature can be expected.

IV. CONCLUSIONS

By exploiting the elemental-resolved resolution of APT at the atomic scale, we perform a complete characterization of the interface properties in UHV CVD grown Ge/Si-Ge quantum heterostructures, obtaining the full set of interface parameters, including width, rms roughness, and in-plane and axial correlation of the heterointerfaces. Our results demonstrate the excellent interface quality, in terms of small interface width ($\mathcal{L} = 1.16$ nm) and interface rms roughness ($\Delta = 0.18$ nm). We also find that the interface roughness is correlated along the growth axis on a length ($\Lambda_\perp = 0.26$ nm) smaller than that of the interfacial width $\mathcal{L}$. Such a partial coherence of roughness along the growth direction is neglected by available models of the IFR scattering. To obtain a realistic picture consistent with the measured structural data, we therefore develop a generalized theory for IFR scattering of carriers in semiconductor heterostructures featuring a non-negligible interface width. With our measured value of the axial correlation length, this model predicts a reduction of IFR scattering by a factor of about 3 times, compared with that corresponding to an abrupt interface featuring the same rms IFR value and a perfect axial correlation within the interface width.

Finally, by using NEGF simulations, we predict the impact of such a reduced IFR scattering rate on the maximum gain of a Ge/Si-Ge-based quantum cascade laser structure, featuring the same material parameters measured in our ACQW structures. The estimated maximum gain is found to be double that obtained using oversimplified models. We underline that the technique of extracting the interfacial parameters demonstrated in this work, in the case of the Ge-Si heterostructures, can be extended to a variety of other material systems, such as III-V multiquantum wells, group IV dielectric interfaces, III-V dielectric interfaces, if clean and high-quality APT maps of these interfaces are achieved. Furthermore, our modeling of scattering by nonabrupt interfaces can be extended to the calculation of roughness-limited mobilities in two-dimensional electronic systems.

ACKNOWLEDGMENTS

This work is supported by the European Union Horizon 2020 program under Grant No. 766719 -FLASH Project. The work carried out in Montreal is supported by the Canada Research Chairs, NSERC (CRD, SPG, and Discovery Grants), Defence Canada (Innovation for Defence Excellence and Security, IDEaS), and PRIMA Québec. Fruitful discussions with Giacomo Scalari, Jerome Faist, and Douglas J. Paul are gratefully acknowledged.

APPENDIX A: THEORY OF INTERFACE ROUGHNESS FOR DIFFUSE INTERFACES: GENERALIZED MODEL

We consider the perturbing potential due to interface roughness. We note $V(\vec{\rho}, z)$ is the 3D potential in the presence of interface roughness, with $\vec{\rho} = (x, y)$ as the in-plane coordinates. The averaged potential along the growth axis is

$$\bar{V}(z) = V(\vec{\rho}, z)_{(\vec{\rho})}, \tag{A1}$$

where $\langle \ldots \rangle_{(\vec{\rho})}$ denotes the in-plane averaging. This 1D potential $\bar{V}(z)$ can be treated exactly by being included in

the 1D Schrödinger equation of the heterostructure. The perturbating potential describing IFR is

$$\delta V_{\mathrm{IFR}}(\vec{\rho},z) = V(\vec{\rho},z) - \bar{V}(z). \tag{A2}$$

For each coordinate, $h(\vec{\rho},z)$ is the height deviation of the isoconcentration surface from its nominal position. We assume h to be small with respect to the interfacial width $\mathcal{L}$. In first order, the IFR potential is

$$\delta V_{\mathrm{IFR}}(\vec{\rho},z) = -\frac{\partial \bar{V}}{\partial z}(z)\, h(\vec{\rho},z). \tag{A3}$$

We consider two eigenstates, $|\alpha,\ \vec{k}\rangle$ and $|\beta,\ \vec{k}+\vec{q}\rangle$, of the nonperturbated heterostructure potential, the wave functions of which can be written $\varphi_\alpha(z)e^{i\vec{k}.\vec{\rho}}/\sqrt{S}$ and $\varphi_\beta(z)e^{i(\vec{k}+\vec{q}).\vec{\rho}}/\sqrt{S}$, respectively; S is a normalization surface. The IFR potential generates the following coupling between these two states:

$$\begin{aligned}\delta V_{\alpha\beta\vec{q}} &= \langle \alpha,\ \vec{k}|\delta V_{\mathrm{IFR}}|\beta,\vec{k}+\vec{q}\rangle \\ &= -\int dz\ \varphi_\alpha^*(z)\varphi_\beta(z)\frac{\partial \bar{V}}{\partial z}(z)\int \frac{d^2\vec{\rho}}{S}\ e^{i\vec{q}.\vec{\rho}}\ h(z,\vec{\rho}).\end{aligned} \tag{A4}$$

Within Fermi's golden rule, the IFR scattering rate from the initial state $|\alpha,\ \vec{k}\rangle$ to the continuum formed by the subband $|\beta,\vec{k}+\vec{q}\rangle\rangle$ is obtained by summing over the exchanged in-plane momentum $\vec{q}$:

$$\hbar\Gamma_{\alpha,\beta,k} = 2\pi \sum_{\vec{q}} |\delta V_{\alpha\beta\vec{q}}|^2 \delta\left(E_{\alpha\beta} + \frac{\hbar^2 k^2}{2m_\parallel^*} - \frac{\hbar^2|\vec{k}+\vec{q}|^2}{2m_\parallel^*}\right), \tag{A5}$$

in which we have to evaluate the squared matrix element:

$$\begin{aligned}|\delta V_{\alpha\beta\vec{q}}|^2 = \int dz_1 \int dz_2\ \varphi_\beta^*(z_1)\varphi_\alpha(z_1)\ \varphi_\alpha^*(z_2)\varphi_\beta(z_2) \\ \times \frac{\partial \bar{V}}{\partial z}(z_1)\frac{\partial \bar{V}}{\partial z}(z_2)\int \frac{d^2\overrightarrow{\rho_1}}{S}\ \int \frac{d^2\overrightarrow{\rho_2}}{S} e^{i\vec{q}.(\overrightarrow{\rho_1}-\overrightarrow{\rho_2})} \\ \times h(z_1,\overrightarrow{\rho_1})h(z_2,\overrightarrow{\rho_2}).\end{aligned} \tag{A6}$$

The 3D correlation function of the roughness is defined as

$$\begin{aligned}\int \frac{d^2\overrightarrow{\rho_0}}{S} h(z_0,\overrightarrow{\rho_0})h(z_0+z,\overrightarrow{\rho_0}+\overrightarrow{\rho}) : \\ = \langle h(z_0,\overrightarrow{\rho_0})h(z_0+z,\overrightarrow{\rho_0}+\overrightarrow{\rho})\rangle_{\overrightarrow{\rho_0}},\end{aligned} \tag{A7}$$

which is assumed to be invariant by translation along the z axis. We calculate Δ as the rms of the interface roughness:

$$\Delta^2 = \langle h(z,\vec{\rho}\ )h(z,\overrightarrow{\rho})\rangle_{\vec{\rho}}. \tag{A8}$$

We assume that this 3D correlation function can be factorized into an axial (1D) and an in-plane (2D) correlation function:

$$\begin{aligned}\frac{\langle h(z_1,\overrightarrow{\rho_1})h(z_2,\overrightarrow{\rho_2})\rangle_{\overrightarrow{\rho_1}}}{\Delta^2} = \frac{\langle h(z_1,\overrightarrow{\rho_1}\ )h(z_1,\overrightarrow{\rho_2})\rangle_{\overrightarrow{\rho_1}}}{\Delta^2} \\ \times \frac{\langle h(z_1,\overrightarrow{\rho_2})h(z_2,\overrightarrow{\rho_2})\rangle_{\overrightarrow{\rho_2}}}{\Delta^2},\end{aligned} \tag{A9}$$

where the right-hand side is composed of an in-plane correlation function

$$C_\parallel(|\overrightarrow{\rho}\ |) = \frac{\langle h(z,\overrightarrow{\rho_0})h(z,\overrightarrow{\rho_0}+\overrightarrow{\rho}\ )\rangle_{\overrightarrow{\rho_0}}}{\Delta^2}, \tag{A10}$$

and an axial correlation function

$$C_\perp(|z|) = \frac{\langle h(z_0,\overrightarrow{\rho_0}\ )h(z_0+z,\overrightarrow{\rho_0})\rangle_{\overrightarrow{\rho_0}}}{\Delta^2}, \tag{A11}$$

which are assumed to depend only on the norm of $\overrightarrow{\rho}$ and z, respectively. While the in-plane correlation function is typically introduced to describe the roughness of abrupt interfaces, the introduction of an axial (i.e., out-of-plane) correlation function is here needed for the generalization to diffuse interfaces. Physically, this correlation function describes how the shape of isoconcentration surfaces are correlated along the z coordinate, i.e., how the roughness at a position z_2 is reminiscent of the roughness at a position z_1.

The matrix element squared can now be expressed as a function of these two correlation functions:

$$\begin{aligned}|\delta V_{\alpha\beta q}|^2 = \Delta^2 C_\parallel(|\overrightarrow{q}\ |)\int dz_1 \int dz_2\ \varphi_\beta^*(z_1)\varphi_\alpha(z_1)\varphi_\alpha^*(z_2) \\ \times \varphi_\beta(z_2)\frac{\partial \bar{V}}{\partial z}(z_1)\frac{\partial \bar{V}}{\partial z}(z_2)C_\perp(|z_2-z_1|)\ ,\end{aligned} \tag{A12}$$

where the Fourier transform of the in-plane correlation function is defined as

$$C_\parallel(|\overrightarrow{q}\ |) = \int \frac{d^2\vec{\rho}}{\mathrm{S}} e^{i\vec{q}.\overrightarrow{\rho}}\, C_\parallel(|\overrightarrow{\rho}\ |). \tag{A13}$$

The in-plane correlation function extracted from the experimental data is fitted using a Gaussian shape:

$$C_\parallel(\rho) = e^{\left(-\rho/\Lambda_\parallel\right)^2}, \tag{A14}$$

while the axial correlation function is fitted using a monoexponential:

$$C_\perp(z) = e^{-z/\Lambda_\perp}. \tag{A15}$$

Plugging in these specific forms of the correlation functions leads to

$$|\delta V_{\alpha\beta\vec{q}}|^2\ = \frac{\pi\Lambda_\parallel^2}{S} e^{-\Lambda_\parallel^2 q^2/4}|\delta V_{\alpha\beta}^{(z)}|^2\ ,$$

$$|\delta V^{(z)}_{\alpha\beta}|^2 = \Delta^2 \int dz_1 \int dz_2\ \varphi^*_\beta(z_1)\varphi_\alpha(z_1)\varphi^*_\alpha(z_2)\varphi_\beta \times (z_2)\frac{\partial \bar{V}}{\partial z}(z_1)\frac{\partial \bar{V}}{\partial z}(z_2)e^{-|z_2-z_1|/\Lambda_\perp}, \quad \text{(A16)}$$

Finally, the scattering rate from the initial state $|\alpha,\ \vec{k}\rangle$ to the subband β is

$$\Gamma_{\alpha\beta k} = \frac{\Lambda^2_\parallel m^*_\parallel}{\hbar^3}|\delta V^{(z)}_{\alpha\beta}|^2 \int_0^\pi d\theta e^{-\frac{\Lambda^2_\parallel q^2(\theta)}{4}}, \quad \text{(A17)}$$

where the exchanged momentum $\vec{q}$ fulfils the following energy conservation:

$$E_\alpha + \frac{\hbar^2|\vec{k}|^2}{2m^*_\parallel} = E_\beta + \frac{\hbar^2|\vec{k}+\vec{q}|^2}{2m^*_\parallel}. \quad \text{(A18)}$$

APPENDIX B: REDUCTION FACTOR WITH RESPECT TO ABRUPT INTERFACE

We consider below the case where the interfaces are well separated with respect to the interfacial width $\mathcal{L} \ll |z_{i+1} - z_i|$ and the axial correlation length $\Lambda_\perp \ll |z_{i+1} - z_i|$. In addition, we assume that the variations of the wave functions are small at the scale of the interface. We consider the error function shape for the potential across each interface: in the vicinity of the interface i, assuming that the first interface i=1 corresponds to a potential profile with negative slope, the electronic potential has the form

$$\bar{V}(z) = v_0 \pm (-1)^i \frac{V_0}{2}\sum_i \left(1 + \text{erf}\left[\frac{(z-z_i)}{L}\right]\right), \quad \text{(B1)}$$

where V_0 is the potential offset across each interface, and $\mathcal{L} = 2\sqrt{\ln(2)}L$ is the interfacial width. By evaluating the z integral, we find

$$|\delta V_{\alpha\beta\vec{q}}|^2 = \Delta^2\pi\ \frac{\Lambda^2_\parallel}{S}e^{-\Lambda^2_\parallel q^2/4}F\text{V}_0^2\sum_i \varphi^*_\beta(z_i) \times \varphi_\alpha(z_i)\ \varphi^*_\alpha(z_i)\varphi_\beta(z_i), \quad \text{(B2)}$$

$$F = \int dz_1 \int dz_2\ \frac{\partial V}{\partial z}(z_1)\frac{\partial V}{\partial z}(z_2)C_\perp(|z_2 - z_1|) = \exp\left(\frac{L^2}{4\Lambda^2_\perp}\right)\left[\text{erf}\left(-\frac{L}{\Lambda_\perp}\right) + 1\right]. \quad \text{(B3)}$$

This factor F can be interpreted as a reduction factor of the IFR scattering rate with respect to the abrupt interface case. Indeed, for $\mathcal{L} \ll \Lambda_\perp$, $F = 1$ and the usual expression of IFR scattering rate for abrupt interfaces is retrieved. However, in the presence of a finite axial correlation length, the IFR scattering rate is reduced ($F < 1$). We can also define an equivalent rms roughness for abrupt interfaces as

$$\Delta^2_{\text{eq}} = \Delta^2 F. \quad \text{(B4)}$$

The value $F = 0.35$ calculated in this work gives an equivalent rms roughness of $\Delta_{\text{eq}} = 0.106\,\text{nm}$, instead of $\Delta = 0.18\,\text{nm}$.

Notably, this equivalent rms roughness can be used as long as the interfaces are well separated ($\mathcal{L} \ll |z_{i+1} - z_i|$). This is not the case for the thinnest barriers studied in the present work, where the effective height of the barrier is reduced.

APPENDIX C: SELF-ENERGIES FOR DIFFUSE INTERFACES IN THE NONEQUILIBRIUM GREEN'S FUNCTIONS FORMALISM

Within the NEGF formalism, we express here the self-energy corresponding to the generalized model of IFR scattering presented above. This IFR self-energy is calculated within the self-consistent Born approximation. Its lesser component is

$$\Sigma^<_{\alpha\beta}(E,\vec{k}) = \sum_{\mu,\nu,\vec{q}} \delta V_{\alpha\mu\vec{q}}\delta V^*_{\beta\nu\vec{q}}\ G^<_{\mu\nu}(E,\vec{k}+\vec{q}).$$

As derived for the scattering rates, it can be shown similarly that the product of the coupling terms can be expressed as a function of the in-plane and axial correlation functions:

$$\delta V_{\alpha\mu\vec{q}}\delta V^*_{\beta\nu\vec{q}} = \Delta^2 C_\parallel(|\overrightarrow{q}|)\int dz_1 \int dz_2 \varphi^*_\alpha(z_1)\varphi_\mu(z_1)\varphi^*_\nu(z_2) \times \varphi_\beta(z_2)\frac{\partial\ \bar{V}}{\partial\ z}(z_1)\frac{\partial\ \bar{V}}{\partial\ z}(z_2)C_\perp(|z_2 - z_1|).$$

Notably, when a localized basis is used for the NEGF calculation, only the terms $\mu = \alpha$ and $\nu = \beta$ are nonzero.

[1] International Roadmap for Devices and Systems, www.irds.ieee.org2017.

[2] D. Sabbagh, N. Thomas, J. Torres, R. Pillarisetty, P. Amin, H. C. George, K. Singh, A. Budrevich, M. Robinson, D. Merrill, *et al.*, Quantum Transport Properties of Industrial $^{28}Si/^{28}SiO_2$, Phys. Rev. Appl. **12**, 014013 (2019).

[3] L. Thumfart, J. Carrete, B. Vermeersch, N. Ye, T. Truglas, J. Feser, H. Groiss, N. Mingo, and A. Rastelli, Thermal transport through Ge-rich Ge/Si superlattices grown on Ge(001), J. Phys. D: Appl. Phys. **51**, 014001 (2018).

[4] Z. Aksamija and I. Knezevic, Thermal conductivity of $Si_{1-x}Ge_x/Si_{1-y}Ge_y$ superlattices: Competition between interfacial and internal scattering, Phys. Rev. B **88**, 155318 (2013).

[5] F. Capasso and A. Y. Cho, Bandgap engineering of semiconductor heterostructures by molecular beam epitaxy: Physics and applications, Surf. Sci. **299-300**, 878 (1994).

[6] Y. Chiu, Y. Dikmelik, P. Q. Liu, N. L. Aung, J. B. Khurgin, and C. F. Gmachl, Importance of interface roughness induced intersubband scattering in mid-infrared quantum cascade lasers, Appl. Phys. Lett. **101**, 171117 (2012).

[7] A. Pirovano, A. L. Lacaita, G. Ghidini, and G. Tallarida, On the correlation between surface roughness and inversion layer mobility in Si-MOSFETs, IEEE Electron Device Lett. **21**, 34 (2000).

[8] Y. Zhao, H. Matsumoto, T. Sato, S. Koyama, M. Takenaka, and S. Takagi, A novel characterization scheme of Si/SiO_2 interface roughness for surface roughness scattering-limited mobilities of electrons and holes in unstrained- and strained-Si MOSFETs, IEEE Trans. Electron Devices **57**, 2057 (2010).

[9] S. M. Goodnick, D. K. Ferry, C. W. Wilmsen, Z. Liliental, D. Fathy, and O. L. Krivanek, Surface roughness at the Si(100)-SiO_2 interface, Phys. Rev. B **32**, 8171 (1985).

[10] E. Luna, Á Guzmán, A. Trampert, and G. Álvarez, Critical Role of Two-Dimensional Island-Mediated Growth on the Formation of Semiconductor Heterointerfaces, Phys. Rev. Lett. **109**, 126101 (2012).

[11] X. Lü, L. Schrottke, E. Luna, and H. T. Grahn, Efficient simulation of the impact of interface grading on the transport and optical properties of semiconductor heterostructures, Appl. Phys. Lett. **104**, 232106 (2014).

[12] A. Y. Song, R. Bhat, A. A. Allerman, J. Wang, T.-Y. Huang, C.-E. Zah, and C. F. Gmachl, Quantum cascade emission in the III-nitride material system designed with effective interface grading, Appl. Phys. Lett. **107**, 132104 (2015).

[13] T. Unuma, T. Takahashi, T. Noda, M. Yoshita, H. Sakaki, M. Baba, and H. Akiyama, Effects of interface roughness and phonon scattering on intersubband absorption linewidth in a GaAs quantum well, Appl. Phys. Lett. **78**, 3448 (2001).

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

[15] M. Califano, N. Q. Vinh, P. J. Phillips, Z. Ikonić, R. W. Kelsall, P. Harrison, C. R. Pidgeon, B. N. Murdin, D. J. Paul, P. Townsend, *et al.*, Interwell relaxation times in *p*-Si/SiGe asymmetric quantum well structures: Role of interface roughness, Phys. Rev. B **75**, 045338 (2007).

[16] H. Sakaki, T. Noda, K. Hirakawa, M. Tanaka, and T. Matsusue, Interface roughness scattering in GaAs/AlAs quantum wells, Appl. Phys. Lett. **51**, 1934 (1987).

[17] A. Valavanis, Z. Ikonić, and R. W. Kelsall, Intersubband carrier scattering in *n*- and *p*-Si/SiGe quantum wells with diffuse interfaces, Phys. Rev. B **77**, 075312 (2008).

[18] O. Dyck, D. N. Leonard, L. F. Edge, C. A. Jackson, E. J. Pritchett, P. W. Deelman, and J. D. Poplawsky, Accurate quantification of Si/SiGe interface profiles via atom probe tomography, Adv. Mater. Interfaces **4**, 1700622 (2017).

[19] S. Mukherjee, A. Attiaoui, M. Bauer, and O. Moutanabbir, 3-D atomic mapping of interfacial roughness and its spatial correlation length in sub-10nm superlattices, ACS Appl. Mater. Interfaces **12**, 1728 (2020).

[20] D. J. Paul, The progress towards terahertz quantum cascade lasers on silicon substrates, Laser Photonics Rev. **4**, 610 (2010).

[21] T. Grange, D. Stark, G. Scalari, J. Faist, L. Persichetti, L. Di Gaspare, M. De Seta, M. Ortolani, D. J. Paul, G. Capellini, *et al.*, Room temperature operation of n-type Ge/SiGe terahertz quantum cascade lasers predicted by non-equilibrium green's functions, Appl. Phys. Lett. **114**, 111102 (2019).

[22] N. W. Hendrickx, D. P. Franke, A. Sammak, G. Scappucci, and M. Veldhorst, Fast two-qubit logic with holes in germanium, Nature **577**, 487 (2020).

[23] C. Ciano, M. Virgilio, M. Montanari, L. Persichetti, L. Di Gaspare, M. Ortolani, L. Baldassarre, M. H. Zoellner, O. Skibitzki, G. Scalari, *et al.*, Control of Electron-State Coupling in Asymmetric Ge/SiGe Quantum Wells, Phys. Rev. App. **11**, 014003 (2019).

[24] T. Grange, Contrasting influence of charged impurities on transport and gain in terahertz quantum cascade lasers, Phys. Rev. B **92**, 241306 (2015).

[25] T. Grange, Electron transport in quantum wire superlattices, Phys. Rev. B **89**, 165310 (2014).

[26] F. Banit, S. C. Lee, A. Knorr, and A. Wacker, Self-consistent theory of the gain linewidth for quantum-cascade lasers, Appl. Phys. Lett. **86**, 041108 (2005).

[27] See the Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevApplied.13.044062 for additional structural data and for the evaluation of the effect of composition grading on the calculated gain of a quantum cascade laser.

[28] S. Koelling, M. Gilbert, J. Goossens, A. Hikavyy, O. Richard, and W. Vandervorst, High depth resolution analysis of Si/SiGe multilayers with the atom probe, Appl. Phys. Lett. **95**, 144106 (2009).

[29] A. Bashir, K. Gallacher, R. W. Millar, D. J. Paul, A. Ballabio, J. Frigerio, G. Isella, D. Kriegner, M. Ortolani, J. Barthel, *et al.*, Interfacial sharpness and intermixing in a Ge-SiGe multiple quantum well structure, J. Appl. Phys. **123**, 035703 (2018).

[30] Y. Yamamoto, O. Skibitzki, M. A. Schubert, M. Scuderi, F. Reichmann, M. Zöllner, M. De Seta, G. Capellini, and B. Tillack, Ge/SiGe multi quantum well fabrication by using reduced pressure chemical vapor deposition, Jpn. J. Appl. Phys. **59**, SGGK10 (2020).

[31] D. J. Norris, A. G. Cullis, T. J. Grasby, and E. H. C. Parker, Investigation of nanoscale Ge segregation in p-channel SiGe/Si field effect transistor structures by electron energy loss imaging, J. Appl. Phys. **86**, 7183 (1999).

[32] M. De Seta, G. Capellini, L. Di Gaspare, F. Evangelisti, and F. D'Acapito, Freezing shape and composition of Ge/Si(001) self-assembled islands during silicon capping, J. Appl. Phys. **100**, 093516 (2006).

[33] T. Gredig, E. A. Silverstein, and M. P. Byrne, Height-Height correlation function to determine grain size in iron phthalocyanine thin films, J. Phys.: Conf. Ser. **417**, 012069 (2013).

[34] T. Ando, A. B. Fowler, and F. Stern, Electronic properties of two-dimensional systems, Rev. Mod. Phys. **54**, 437 (1982).

[35] X. Lü, E. Luna, L. Schrottke, K. Biermann, and H. T. Grahn, Determination of the interface parameter in terahertz quantum-cascade laser structures based on transmission electron microscopy, Appl. Phys. Lett. **113**, 172101 (2018).

[36] Although a gaussian function is normally used to describe the in-plane roughness correlation of Si-Ge and III-V interfaces (see Refs. [13,14,17]), some authors have reported an exponential decay function for describing other heterointerfaces (see e.g., Yoshinobu *et al.*, J. of Vac. Sci. & Technol. B **13**, 1630 (1995)). In this case, the expression for the height-height correlation function is $H(\tau) = 2\Delta^2[1 - e^{-(\tau/\Lambda_\parallel)}]$ and, consequently, Eq. (10) has to be replaced by $|\delta V_{\alpha\beta q}|^2 = \pi\Delta^2 \frac{2\Lambda_\parallel^2}{(1+\Lambda_\parallel^2 q^2)^{3/2}} \int dz_1 \int dz_2\ \varphi_\beta^*(z_1)\varphi_\alpha(z_1)\ \varphi_\alpha^*(z_2)\varphi_\beta(z_2) \frac{\partial \tilde{V}}{\partial z}(z_1) \frac{\partial \tilde{V}}{\partial z}(z_2) C_\perp(|z_2 - z_1|)$.

[37] Y. V. Flores, S. S. Kurlov, M. Elagin, M. P. Semtsiv, and W. T. Masselink, Leakage current in quantum-cascade lasers through interface roughness scattering, Appl. Phys. Lett. **103**, 161102 (2013).

[38] Y. V. Flores and A. Albo, Impact of interface roughness scattering on the performance of GaAs/$Al_xGa_{1-x}As$ terahertz quantum cascade lasers, IEEE J. Quantum Electron. **53**, 1 (2017).

[39] Notably, the Si-Ge QCL design proposed here is slightly different from that investigated in Ref. [21], since our structure is optimized to maximize the room-temperature gain in the presence of diffuse interfaces.