Semiconductor Quantum Plasmonics

Angela Vasanelli ®,[1,*] Simon Huppert ®,[2,†] Andrew Haky,[1] Thibault Laurent,[2] Yanko Todorov ®,[1] and Carlo Sirtori ®[1]

[1]*Laboratoire de Physique de l'Ecole normale supérieure, ENS, Université PSL, CNRS, Sorbonne Université,*
Université de Paris, 75005 Paris, France
[2]*Laboratoire Matériaux et Phénomènes Quantiques, CNRS—UMR7162, Université de Paris, 75013 Paris, France*

We investigate the frontier between classical and quantum plasmonics in highly doped semiconductor layers. The choice of a semiconductor platform instead of metals for our study permits an accurate description of the quantum nature of the electrons constituting the plasmonic response, which is a crucial requirement for quantum plasmonics. Our quantum model allows us to calculate the collective plasmonic resonances from the electronic states determined by an arbitrary one-dimensional potential. Our approach is corroborated with experimental spectra, realized on a single quantum well, in which higher order longitudinal plasmonic modes are present. We demonstrate that their energy depends on the plasma energy, as is also the case for metals, but also on the size confinement of the constituent electrons. This work opens the way toward the applicability of quantum engineering techniques for semiconductor plasmonics.

The properties of plasmons in nanostructures can be profoundly modified whenever their elementary constituents, electrons and photons, enter the quantum regime [1–5]. One of the pioneering works in quantum plasmonics is the demonstration that the optical resonances of localized surface plasmons can be strongly affected by the electronic confinement for metallic nanoparticle sizes of the order of 10 nm [6]. In this case, the Drude model fails in describing the optical response of the electron gas, and a quantum treatment or nonlocal electromagnetic models [7] must be considered. Size confinement also strongly affects the optical properties of polar materials [8] and has required the use of computational models based on perturbative density functional theory [9] or the use of a nonlocal dielectric response that goes beyond the Lorentz model to describe the material resonances [10]. Another fundamental quantum effect, tunneling, has been shown to strongly impact the properties of the plasmons beyond the classical treatment [11–13].

In this Letter, we investigate the frontier between classical and quantum plasmonics by studying the effect of a confining potential on volume plasmons. We observe that size confinement gives rise to several longitudinal modes with quantized wave vectors at different energies, as a particle in a quantum well. We demonstrate that these modes do not obey the Lindhard formula, which holds for thin metallic films. Nonlocality of volume plasmons in semiconductor layers clearly appears to be related to electronic size confinement and is explained by using a quantum model that constructs the plasmon modes directly from the confined states of the constituent electrons.

Figure 1 summarizes the properties of classical and quantum plasmons in a highly doped semiconductor.

A sketch of the system is presented in Fig. 1(a). It is composed of a highly doped semiconductor layer with a thickness smaller than the plasma wavelength in the material. The oscillation of the free electrons in the doped layer excited by an incident electromagnetic field results in an optically active collective mode of the system: a plasmon confined in the doped layer with a dipole moment along the growth direction z. The absorption spectrum, sketched in Fig. 1(a), presents a single Lorentzian resonance, centered at the plasma energy, with a quality factor on the order of 10 to 20 [14]. This resonance, called the Berreman mode [14–18], can be simulated with the Drude model, which describes the isotropic permittivity of the doped semiconductor, and also by taking into account the finite thickness of the layer through standard electromagnetic simulations. This system is thus an example of *classical plasmonics*. Berreman modes, which were first observed in thin metallic films [15,16,19,20], have recently raised considerable attention because at the plasma frequency the real part of the dielectric permittivity is zero. For this reason, the Berreman mode is also referred to as an epsilon-near-zero (ENZ) mode [21,22]. Hyperbolic metamaterials obtained by alternating ENZ and dielectric layers have also been demonstrated [23]. In the classical description, the plasma frequency is independent from the plasmon wave vector $\bar{k}$, as sketched in Fig. 1(b). As a consequence, the dielectric function is local, i.e., it only depends on the photon frequency and not on the wave vector.

Quantum effects appear in semiconductors when the thickness of the layer is smaller than the de Broglie wavelength of electrons. In this case, size confinement along the z direction gives rise to quantized energy levels. Figure 1(c) sketches a doped semiconductor quantum well (QW) with three occupied confined states. In this case,
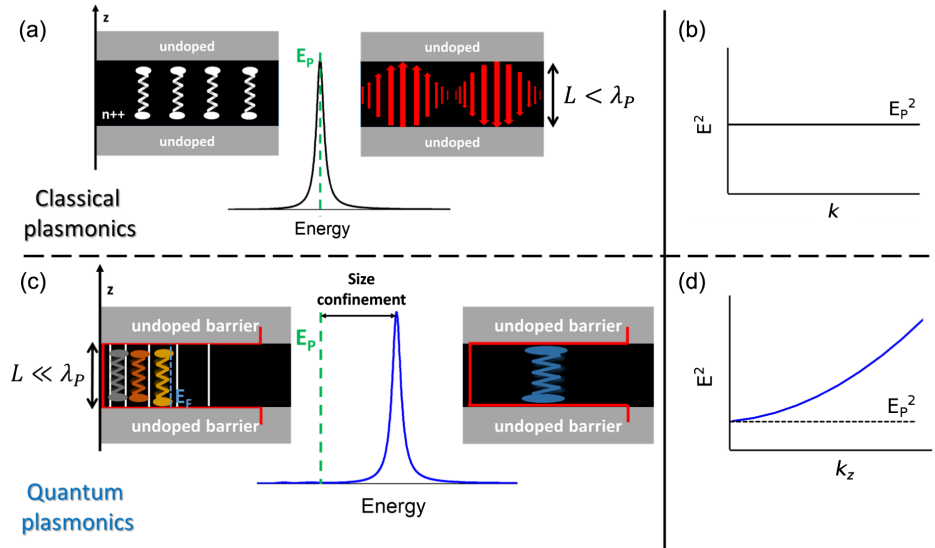
FIG. 1. (a) Sketch of a doped semiconductor layer embedded between two undoped layers. If the thickness L is smaller than the plasma wavelength in the material $\lambda_P = 2\pi c/\Omega_P$, it is possible to optically excite a Berreman mode at $E_P = \hbar\Omega_P$. (b) Dispersion of a volume plasmon in a classical model. (c) Effect of size confinement in the optical response of a semiconductor QW with three occupied electronic states. The optical response of the system is a collective resonance at a greater energy than E_P due to size confinement contribution. (d) Longitudinal dispersion of a volume plasmon taking into account the material band structure and the Fermi distribution of electrons.

there are three main optically active transitions, represented as three different sets of harmonic oscillators along z. Dipole-dipole interaction between these optically active transitions gives rise to a collective mode of the system, a confined plasmon, with an energy that is higher than the plasma energy of the electron gas [14,24,25] and higher than the energy of the individual electronic transitions. The Drude model thus fails to describe the collective optical properties of confined electrons, while quantum [22,25,26] or nonlocal semiclassical [27] models correctly reproduce the experimental absorption spectra.

Optical experiments conducted on metal foils of a few nm thickness [19,20,28] have shown the existence of higher order longitudinal modes known as Tonks–Dattner modes [20,29,30], whose energies are described by the Lindhard formula: $\Omega_{k_z}^2 = \Omega_p^2 + (3/5)v_F^2 k_z^2$, where Ω_p is the bulk plasma frequency, v_F the Fermi velocity, and k_z the z component of the plasmon wave vector. This dispersion relation, sketched in Fig. 1(d), requires a quantum treatment of electrons in metals and a semiclassical treatment of light-matter interaction [31].

In order to investigate the effect of a confining potential on the Berreman mode and on the higher order longitudinal plasmons, we have chosen to work within a semiconductor platform. Indeed, the quantum nature of the electrons is described straightforwardly in a semiconductor through the envelope function approximation that takes into account size confinement and tunneling [32], while in metals a quantum description requires numerical approaches. Our sample is a doped 100 nm GaInAs/AlInAs layer with electronic density $N_v = 7.5 \times 10^{18}$ cm^{-3}, embedded

between two AlInAs barriers. We have performed angle resolved emission experiments under thermal excitation of the plasmons through the application of an in-plane current, as in Ref. [33]. Figure 2 shows normalized thermal emission spectra (in red) at different angles, measured at room temperature. Several peaks are clearly observed in the
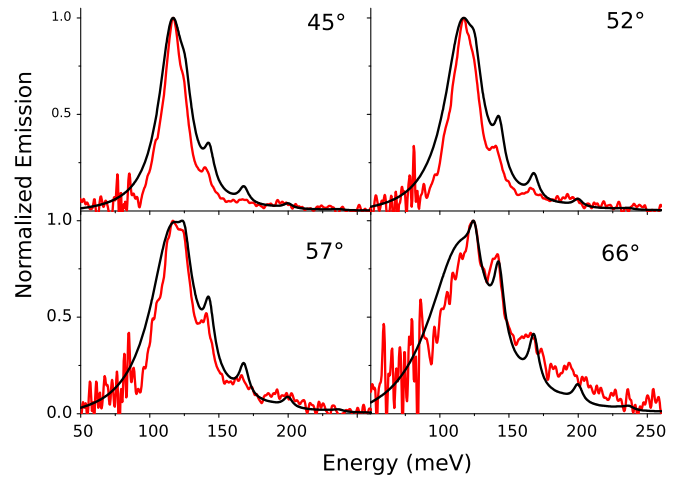

FIG. 2. Measured (red lines) and simulated (black lines) emission spectra under thermal excitation of the plasmons for different values of the internal angle of light propagation. The thermal emission spectra have been simulated by solving quantum Langevin equations in the input-output formalism [34]. The plasmon eigenmodes have been calculated from a full numerical diagonalization accounting for the finite barrier height as well as for nonparabolicity effects [35].

spectra. When increasing the angle, the main plasmon resonance, the Berreman mode, becomes broader and weaker due to the increase of the radiative decay rate of the plasmon [33]. Simultaneously, the higher energy resonances, the Tonks–Dattner modes, become more and more visible.

The energy position of all the plasmon modes and their angle dependent radiative broadening are very well reproduced by our quantum model (black lines in Fig. 2), which is based on the dipole representation of the light-matter interaction in the Coulomb gauge [25,26,34,36]. In this model, the Hamiltonian of the electron gas in the QW is written in terms of the excitations between confined states as

$$H = \sum_{\alpha} \hbar \omega_{\alpha} b_{\alpha}^{\dagger} b_{\alpha}$$
$$+ \frac{e^2}{2\epsilon_0 \epsilon_s} \sum_{\alpha,\beta} S_{\alpha,\beta} \sqrt{\Delta N_{\alpha} \Delta N_{\beta}} (b_{\alpha}^{\dagger} + b_{\alpha})(b_{\beta}^{\dagger} + b_{\beta}). \quad (1)$$

Here the indices α, β run over all transitions between confined states in the QW; the operators $b_{\alpha}^{\dagger}, b_{\alpha}$ are the creation and annihilation operators of the transition α of energy $\hbar \omega_{\alpha}$ and ΔN_{α} the associated population difference; ϵ_s is the background dielectric constant. The coupling coefficients $S_{\alpha,\beta}$ are expressed as

$$S_{\alpha\beta} = \frac{1}{\hbar \omega_{\alpha}} \frac{1}{\hbar \omega_{\beta}} \left(\frac{\hbar^2}{2m^*} \right)^2 \int_{-\infty}^{+\infty} dz \xi_{\alpha}(z) \xi_{\beta}(z), \quad (2)$$

with m^* the electron effective mass. The coupling coefficients between the electronic transitions $S_{\alpha,\beta}$ are proportional to the overlap between the a.c. microcurrent functions ξ_{α} describing the electronic transitions $\alpha = i \rightarrow i+j$ and defined as [36]

$$\xi_{\alpha}(z) \equiv \xi_{i \rightarrow i+j}(z) = \psi_i(z) \frac{\partial \psi_{i+j}(z)}{\partial z} - \psi_{i+j}(z) \frac{\partial \psi_i(z)}{\partial z}, \quad (3)$$

where ψ_i is the envelope function of the confined electronic state with quantum number i. The integral of each microcurrent function is proportional to the optical dipole of the corresponding transition [26,36]. Note that in a square QW the transitions with a nonzero dipole are only those with odd j.

Figure 3 focuses on the construction of quantum confined plasmons from single electron transitions in a QW with infinite barriers and thickness L. In this case, the microcurrent associated with the transition $i \rightarrow i+j$ writes as

$$\xi_{i \rightarrow i+j}(z) = \frac{\pi}{L^2} \left[j \sin \left(\frac{2i+j}{L} \pi z \right) - (2i+j) \sin \left(\frac{j\pi z}{L} \right) \right]. \quad (4)$$

Examples of such functions for optically active transitions $i \rightarrow i+j$, with odd j, are plotted in Fig. 3(a), ($j = 1$), and

3(b), ($j = 3$), for different values of i. From both panels one can see that, for a fixed j, when increasing the value of the index i, the shape of the functions representing the microcurrents approaches $\sin(j\pi z/L)$. As a consequence, in the limit $i \gg 1$, when a large number of subbands are occupied, the microcurrents $\{\xi_{i \rightarrow i+j}\}_{i,j}$ are mutually orthogonal for different j. In this approximation, the coupling coefficients between microcurrents are given by $S_{i \rightarrow i+j, i \rightarrow i+j'} = L/(2\pi^2 jj')\delta_{j,j'}$. The matrix of the coupling coefficients is block-diagonal: all the electronic transitions with the same j contribute to the same plasmon mode. In other words, the Hamiltonian Eq. (1) can be independently diagonalized on the subspaces relative to transitions $i \rightarrow i+j$ for fixed j. For each subspace of index j, light couples with the plasmon corresponding to the highest frequency eigenmode. Its frequency Ω_j can be found, after Bogoliubov transformation within the associated subspace, by calculating the zeros of the following determinant [26,27,36]:

$$\Delta_j(\omega) = 1 - \frac{2e^2}{\hbar \epsilon_0 \epsilon_s} \sum_{i=1}^{N_{occ}} \frac{S_{i \rightarrow i+j, i \rightarrow i+j} \Delta N_{i \rightarrow i+j} \omega_{i \rightarrow i+j}}{\omega^2 - \omega_{i \rightarrow i+j}^2}$$
$$= 1 - \sum_{i=1}^{N_{occ}} \frac{\omega_{Pi,i+j}^2}{\omega^2 - \omega_{i \rightarrow i+j}^2}, \quad (5)$$

with $\omega_{Pi,i+j}$ the plasma frequency associated with the transition $i \rightarrow i+j$.

Figure 3(c) presents the microcurrents associated with the confined plasmon modes issued from the subspaces of index $j = 1, 3, 5$. These microcurrents present the same symmetry as those of the electronic transitions $i \rightarrow i+j$: they vary as $\sin(j\pi z/L)$ with a quantized wave vector $k_z = j\pi/L$. The oscillator strength of the plasmon mode Ω_j concentrates the interaction with light of all $i \rightarrow i+j$ transitions, and it is given, for odd values of j, by $4L\hbar N_v/(m^*\pi^2 j^2)$ [25,26]. Figure 3(c) also presents the charge distributions oscillating at frequency Ω_j, which are proportional to the derivative of the microcurrents [26]. We can clearly see that the fundamental mode $j = 1$ corresponds to a dipole along z. This mode is thus the analogue of a Berreman mode in a metallic thin film. The modes with $j > 1$ and j odd are higher order longitudinal confined plasmons characterized by a quantized wave vector. They correspond to the Tonks–Dattner modes observed in the spectra in Fig. 2.

The zeros of the determinant, Eq. (5), can be analytically calculated in two different regimes: the *metallic* limit where a very large number of electronic states are occupied in an infinite potential well, and the *semiconductor* limit that describes QWs with only few tens of electronic subbands occupied. In the *metallic* limit, when the number of occupied subbands N_{occ} is such that $j^2/N_{occ} \ll 1$, our model demonstrates the Lindhard formula (see the Supplemental Material [37]). However, in the QW that we have experimentally investigated we are far from the
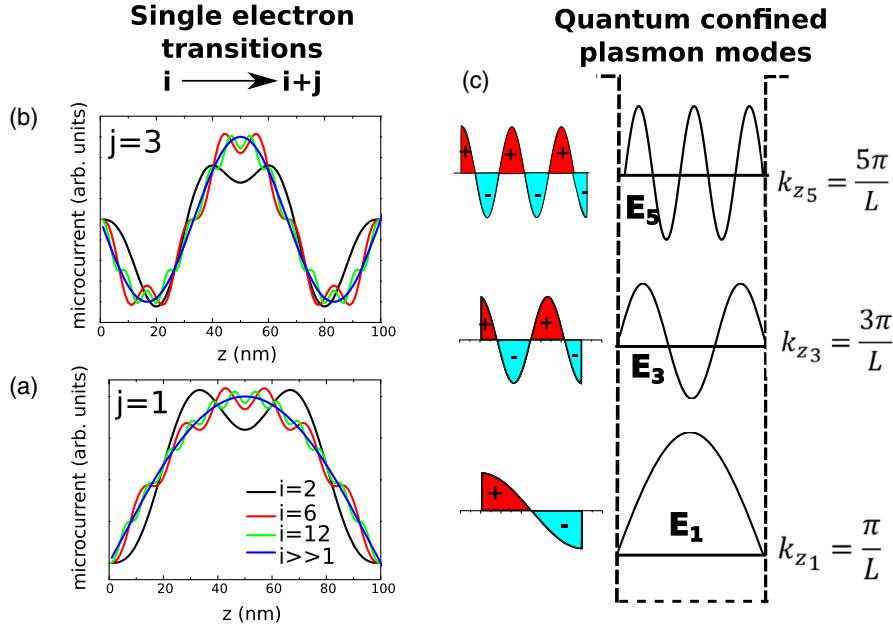
FIG. 3.   (a),(b) Normalized microcurrents $\xi_{i \to i+j}(z)$ calculated for an infinite barrier QW for $j = 1$ (a) and $j = 3$ (b) and $i = 2$ (black), $i = 6$ (red), $i = 12$ (green) and in the limit of $i \gg 1$ (blue). (c) Schematic representation of a square QW for the plasmon modes (dashed line). We have plotted in this effective potential the plasmon microcurrents (black continuous lines) of the $j = 1$, $j = 3$, and $j = 5$ modes, with j the quantum number of the confined longitudinal wave vector. The charge density distributions are also plotted for each longitudinal mode.

metallic limit, as $N_{\text{occ}} = 20$, and we observe higher order modes up to $j = 9$; thus, $j^2/N_{\text{occ}} \gg 1$. In our system, the energy separation between successive confined states is approximately constant due to finite barrier and band nonparabolicity effects. In this case, we can introduce an average energy separation between the confined states, $E_0 = \hbar\omega_0$, calculated as the conduction band offset V_b divided by the total number of confined states in the QW N_{tot}: $E_0 = V_b/N_{\text{tot}}$. The zeros of the determinant, Eq. (5), can be again calculated analytically, observing that $\Omega_P^2 = \sum_{i=1}^{N_{\text{occ}}} \omega_{Pi,i+j}^2$ [14]. The resulting plasmon mode frequencies are accurately described by

$$\Omega_j^2 = \Omega_p^2 + \omega_0^2 j^2. \tag{6}$$

From this equation, it clearly appears that in the *semiconductor* limit the observation of higher order plasmon modes is intimately related to the existence of an energy separation between the electronic states, E_0, induced by size confinement.

In Fig. 4 we compare the results of equation (6) with our data and numerical simulations. Figure 4(a) presents in red symbols the squared plasmon energies $E_j^2 = (\hbar\Omega_j)^2$ extracted from 15 emission spectra measured at angles between 23° and 83° as a function of the index j of the plasmon. The red line in Fig. 4 shows the calculated dispersion following Eq. (6), which is in excellent agreement with the experimental results. Note that this dispersion has been calculated with no free parameters by using the calculated plasma energy $E_P = 114$ meV,

which takes into account band nonparabolicity [14] and the confinement energy $E_0 = 17$ meV. From the excellent agreement between the experimental and the theoretical results, we can infer that, notably, size confinement still plays a nonnegligible role in a QW with $L = 100$ nm, and it is at the origin of the observed nonlocality effects [10].

In Figs. 4(b) and 4(c), we verify the validity of Eq. (6) by varying the QW thickness for $N_v = 7.5 \times 10^{18}$ cm^{-3} [panel (b)] and the electronic density for $L = 100$ nm [panel (c)]. The squared plasmon frequencies are numerically calculated with our full quantum model (bullets) and compared to the results of Eq. (6) (lines). Note that all the dispersions plotted in panel (b) have been calculated with the same value of the plasma energy $E_P = 114$ meV, while those in panel (c) have been obtained with $E_0 = 17$ meV independently on the electronic density.

Having confirmed our microscopic approach in a square QW through the comparison of experimental and theoretical results, we discuss in the last part of this Letter how electronic confinement and tunneling can be used as a degree of freedom to engineer the plasmonic resonances (Fig. 5). The starting point is a GaInAs layer of 54 nm thickness, sandwiched between two AlInAs barriers. The electronic density per unit volume in the GaInAs layer is $N_v = 2 \times 10^{19}$ cm^{-3}. Figure 5 presents the absorptivity spectrum simulated at 45° (black line), showing the Tonks–Dattner resonances as previously discussed. The inset of the figure shows the corresponding band diagram, where we also plotted the square moduli of the electronic envelope functions and the position of the Fermi energy (black
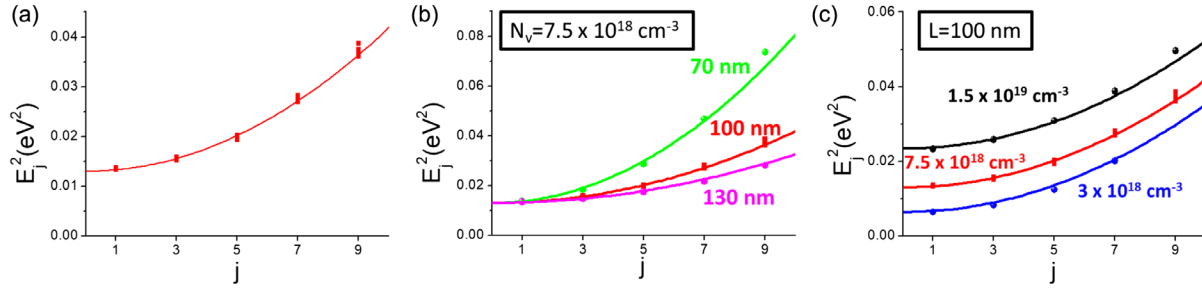
FIG. 4. (a) Squared energy of the plasmon modes extracted from the measured thermal emission spectra (red squares) plotted as a function of the quantum number j of the corresponding confined plasmon. The line shows the result of Eq. (6). (b),(c) Squared plasmon frequencies numerically calculated with our full quantum model (bullets) compared to the results of Eq. (6) (lines). In panel (b), the electronic density has been set to $N_v = 7.5 \times 10^{18}$ cm^{-3}, while the QW thickness is varied and in panel (c) $L = 100$ nm and the electronic density is varied. The experimental results are also reported for comparison (squares).

horizontal line). We now insert in the GaInAs layer six identical AlInAs barriers of 1.5 nm thickness such that the structure is now composed of tunnel coupled asymmetric QWs as shown in the inset of Fig. 5. The electronic structure is profoundly modified by the presence of the barriers, resulting in the formation of several minibands. In particular, as the tunnel coupled QWs have different sizes, we now have optically active transitions between the states of the ground miniband and those of the second excited one (i.e., $j = 2$, which are forbidden in a single QW) [38]. The electronic density per unit volume is kept equal to that of the single GaInAs layer, $N_v = 2 \times 10^{19}$ cm^{-3}. The absorptivity spectrum, including the collective effects, is presented in red. This spectrum is completely different with respect to that obtained in the single layer case, proving that also collective excitations can be engineered by a judicious
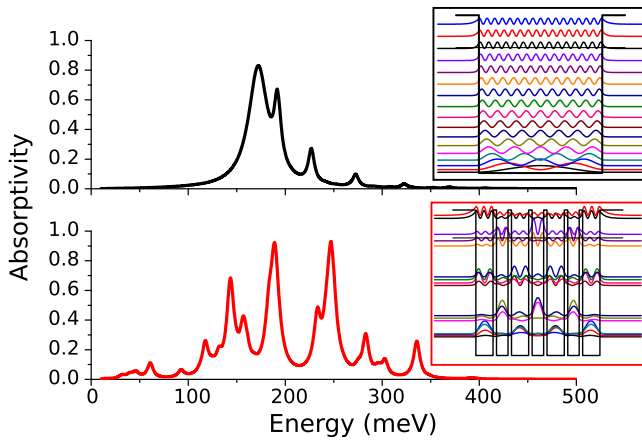

FIG. 5. Upper panel: Calculated absorptivity for a square GaInAs highly doped semiconductor layer with thickness 54 nm and electronic density $N_v = 2 \times 10^{19}$ cm^{-3}. Lower panel: Calculated absorptivity for a system of asymmetric tunnel coupled GaInAs/AlInAs quantum wells with the same electronic density as the 54 nm structure. The conduction band profile and the square moduli of the wave functions for the two structures are presented as insets. The black horizontal lines indicate the Fermi energy.

size confinement of the single electron states. Not only can many collective resonances be observed, but their oscillator strengths are distributed differently between them. In particular, the lowest energy collective mode is no longer the mode with the highest absorptivity.

This Letter clarifies the link between the single electron wave functions and their collective response to the light. In that sense, it opens new degrees of freedom for engineering *ad hoc* plasmonic resonances in which the energy position and oscillator strength are determined by shaping single electron wave functions. Our theoretical approach is well-suited for this goal, as it goes well beyond the Drude model, which is commonly used in semiconductor plasmonics [39], and the Lindhard formula, which applies to a free electron gas in a metallic thin film. Our model thus allows quantum engineering techniques to enter the field of semiconductor plasmonics. The ultimate goal is a complete three-dimensional shaping of the dielectric function, which fully exploits all the degrees of freedom offered by epitaxial growth and nanofabrication techniques in order to fabricate artificial ENZ materials for negative refraction [40] and high harmonic generation [41] exploiting nonlinear optical properties [42].

*Corresponding author.
angela.vasanelli@ens.fr
†Present address: Institut des Nanosciences de Paris, CNRS, Sorbonne Université, 75005 Paris, France.

[1] W. P. Halperin, Rev. Mod. Phys. **58**, 533 (1986).
[2] M. S. Tame, K. R. McEnery, K. Özdemir, J. Lee, S. A. Maier, and M. S. Kim, Nat. Phys. **9**, 329 (2013).
[3] J. M. Fitzgerald, P. Narang, R. V. Craster, S. A. Maier, and V. Giannini, Proc. IEEE **104**, 2307 (2016).

[4] F. J. Garcia de Abajo, Nature (London) **483,** 417 (2012).

[5] M. I. Stockman, K. Kneipp, S. I. Bozhevolnyi, S. Saha, A. Dutta, J. Ndukaife, N. Kinsey, H. Reddy, U. Guler, V. M. Shalaev *et al.*, J. Opt. **20,** 043001 (2018).

[6] J. A. Scholl, A. L. Koh, and J. A. Dionne, Nature (London) **483,** 421 (2012).

[7] J. M. McMahon, S. K. Gray, and G. C. Schatz, Phys. Rev. Lett. **103,** 097403 (2009).

[8] D. C. Ratchford, C. J. Winta, I. Chatzakis, C. T. Ellis, N. C. Passler, J. Winterstein, P. Dev, I. Razdolski, J. R. Matson, J. R. Nolen *et al.*, ACS Nano **13,** 6730 (2019).

[9] T. R. Paudel and W. R. L. Lambrecht, Phys. Rev. B **80,** 104202 (2009).

[10] C. R. Gubbin and S. De Liberato, Phys. Rev. X **10,** 021027 (2020).

[11] K. J. Savage, M. M. Hawkeye, R. Esteban, A. G. Borisov, J. Aizpurua, and J. J. Baumberg, Nature (London) **491,** 574 (2012).

[12] W. Zhu, R. Esteban, A. G. Borisov, J. J. Baumberg, P. Nordlander, H. J. Lezec, J. Aizpurua, and K. B. Crozier, Nat. Commun. **7,** 11495 (2016).

[13] S. F. Tan, L. Wu, J. K. Yang, P. Bai, M. Bosman, and C. A. Nijhuis, Science **343,** 1496 (2014).

[14] B. Askenazi, A. Vasanelli, A. Delteil, Y. Todorov, L. C. Andreani, G. Beaudoin, I. Sagnes, and C. Sirtori, New J. Phys. **16,** 043029 (2014).

[15] R. A. Ferrell, Phys. Rev. **111,** 1214 (1958).

[16] A. J. McAlister and E. A. Stern, Phys. Rev. **132,** 1599 (1963).

[17] B. Harbecke, B. Heinz, and P. Grosse, Appl. Phys. A **38,** 263 (1985).

[18] W. D. Newman, C. L. Cortes, J. Atkinson, S. Pramanik, R. G. DeCorby, and Z. Jacob, ACS Photonics **2,** 2 (2015).

[19] A. R. Melnyk and M. J. Harrison, Phys. Rev. Lett. **21,** 85 (1968).

[20] I. Lindau and P. Nilsson, Phys. Scr. **3,** 87 (1971).

[21] S. Vassant, J.-P. Hugonin, F. Marquier, and J.-J. Greffet, Opt. Express **20,** 23971 (2012).

[22] I. Montano, S. Campione, J. F. Klem, T. E. Beechem, O. Wolf, M. B. Sinclair, and T. S. Luk, Sci. Rep. **8,** 16694 (2018).

[23] A. J. Hoffman, L. Alekseyev, S. S. Howard, K. J. Franz, D. Wasserman, V. A. Podolskiy, E. E. Narimanov, D. L. Sivco, and C. Gmachl, Nat. Mater. **6,** 946 (2007).

[24] T. Ando, A. B. Fowler, and F. Stern, Rev. Mod. Phys. **54,** 437 (1982).

[25] A. Delteil, A. Vasanelli, Y. Todorov, C. Feuillet Palma, M. Renaudat St-Jean, G. Beaudoin, I. Sagnes, and C. Sirtori, Phys. Rev. Lett. **109,** 246808 (2012).

[26] G. Pegolotti, A. Vasanelli, Y. Todorov, and C. Sirtori, Phys. Rev. B **90,** 035305 (2014).

[27] F. Alpeggiani and L. C. Andreani, Phys. Rev. B **90,** 115311 (2014).

[28] A. R. Melnyk and M. J. Harrison, Phys. Rev. B **2,** 835 (1970).

[29] D. A. Hart and N. L. Oleson, J. Appl. Phys. **40,** 4541 (1969).

[30] R. S. Fletcher, X. L. Zhang, and S. L. Rolston, Phys. Rev. Lett. **96,** 105003 (2006).

[31] F. Wooten, *Optical Properties of Solids* (Academic Press, New York, 1972).

[32] P. Harrison, *Quantum Wells, Wires and Dots: Theoretical and Computational Physics of Semiconductor Nanostructures* (John Wiley & Sons, Ltd, New York, 2005).

[33] T. Laurent, Y. Todorov, A. Vasanelli, A. Delteil, C. Sirtori, I. Sagnes, and G. Beaudoin, Phys. Rev. Lett. **115,** 187402 (2015).

[34] S. Huppert, A. Vasanelli, G. Pegolotti, Y. Todorov, and C. Sirtori, Phys. Rev. B **94,** 155418 (2016).

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

[36] Y. Todorov and C. Sirtori, Phys. Rev. B **85,** 045304 (2012).

[37] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.125.187401 for the demonstration of the Lindhard formula for $j^2/N_{\rm occ} \ll 1$ and of Eq. (6) in the semiconductor limit.

[38] C. Sirtori, F. Capasso, D. L. Sivco, S. N. G. Chu, and A. Y. Cho, Appl. Phys. Lett. **59,** 2302 (1991).

[39] T. Taliercio and P. Biagioni, Nanophotonics **8,** 949 (2019).

[40] A. Hierro, M. Montes Bajo, M. Ferraro, J. Tamayo-Arriola, N. Le Biavan, M. Hugues, J. M. Ulloa, M. Giudici, J.-M. Chauveau, and P. Genevet, Phys. Rev. Lett. **123,** 117401 (2019).

[41] Y. Yang, J. Lu, A. Manjavacas, T. Luk, H. Liu, K. Kelley, J.-P. Maria, E. L. Runnerstrom, M. Sinclair, S. Ghimire, and I. Brener, Nat. Phys. **15,** 1022 (2019).

[42] M. Z. Alam, I. De Leon, and R. W. Boyd, Science **352,** 795 (2016).