PHYSICAL REVIEW B **99**, 115405 (2019)

Significance of the nonlocal optical response of metal nanoparticles in describing the operation of plasmonic lasers

Dasuni Lelwala Gamacharige,[1,*] Sarath D. Gunapala,[2] Mark I. Stockman,[3] and Malin Premaratne[1,†]
[1] *Advanced Computing and Simulation Laboratory (AχL), Department of Electrical and Computer Systems Engineering, Monash University, Clayton, Victoria 3800, Australia*
[2] *Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, USA*
[3] *Department of Physics and Astronomy, Georgia State University, Atlanta, Georgia 30303 USA*

(Received 17 December 2018; revised manuscript received 14 February 2019; published 6 March 2019)

A metal nanoparticle (MNP) coupled to a quantum emitter (QE) is a versatile composite nanostructure with unique chemical and physical properties which has been studied intensively, owing to its vast range of promising applications in nanoscience and nanotechnology. When pumped into a higher gain level, an MNP-QE composite nanostructure functions as a nanoplasmonic counterpart of a conventional laser, which is capable of operating at subwavelengths. The theory of plasmonic lasers is hitherto developed on the local optical response of the MNP, disregarding the nanoscale effects of its free electrons. In this paper, we perform a comprehensive quantum mechanical analysis of a complex MNP-QE composite nanostructure, capturing the size-dependent nonclassical effects through the nonlocal optical response of the MNP. Our study reveals that the nonlocal correction introduces significant deviations to the plasmon statistics of the hybrid particle suggested by the local calculations, becoming more prominent when the number of QEs coupled to the MNP increases. Furthermore, for the typical material parameter values used in the literature, we observed the initiation of quenching effects at lower pumping rates than suggested by the local response formalism. In essence, nonlocally assessed plasmon statistics of MNP-QE composite nanostructures demand the concerted coupling of an even higher number of QEs to compensate the deterioration in coherence and to sustain lasing.

DOI: 10.1103/PhysRevB.99.115405

I. INTRODUCTION

In the visible frequency regime, noble metal nanoparticle (MNP)–quantum emitter (QE) composite nanostructures generate interesting optical responses that are tunable, exploiting the physical and optical properties of the particles and the surrounding medium. Such MNPs possess a remarkable ability for retaining the optical energy confined within, owing to the excitation of localized surface plasmons [1–3]. On the other hand, QEs exhibit optoelectronic properties which are adjustable with respect to their composition, size, and shape [4–6]. MNP-QE composite nanostructures (or hybrids), as well as their individual constituents, hold a vast range of fascinating interdisciplinary applications including bio-optics [7–10], nanoelectronic devices [11,12], resonance energy transfer [13–15], and superradiance [16–19]. Furthermore, researchers have conceptualized [20], formulated [21–26], and demonstrated [27] the possibility of realizing a plasmonic laser using stimulated emission of surface plasmons (SPs), which could overcome the half-wavelength diffraction limit imposed on conventional lasers, and thus, bolstering the miniaturization of devices into the nanoscale.

Similar to its macroscale counterpart, a plasmonic laser consists of a resonator which produces SP modes and a gain medium which amplifies the SPs. Metal nanostructures of noble metals, graphene, plasmene, or molybdenum disulfide generally act as the resonator, while QEs including quantum dots, quantum wells, dye molecules, or carbon nanotubes act as the gain medium [20,23–25,28,29]. The gain elements are generally pumped incoherently, and they relax back to form excitons, which would donate energy into the SP mode. The gain medium self-sustains its excitation via the highly localized electric field and stimulated emission, establishing and sustaining coherent operation of the plasmonic laser.

Being the building block of plasmonic lasers, an MNP-QE composite nanostructure forms a quantum mechanical system which interacts with its environment and is characterized by mixed states that are a statistical ensemble of pure states. Concomitant to the short-lived femtosecond range lifetime of the plasmons, a strong plasmon excitation delivered by the concerted coupling of a substantial amount of emitters is pivotal to the realization of a plasmonic laser. Due to the exponentially growing computational complexity of the quantum mechanical analysis of hybrid systems with many emitters, researchers have mainly focused on a classical or semiclassical description of these devices. Yet, the size, shape, and the structure of the particles greatly affect their optical properties in the nanoscale, making an accurate quantum mechanical description crucial in comprehending and interpreting their dynamics. A numerically exact quantum mechanical analysis of a shell of up to 100 QEs coupled to a spherical silver MNP has been studied recently [30,31] and an undemanding analytical solution for such a system, based on the plasmon

---

[*] dasuni.lelwalagamacharige@monash.edu
[†] malin.premaratne@monash.edu

reduced density matrix (RDM) approximation has also been suggested [32].

MNPs and MNP-QE composite nanostructures are often studied under the local response approximation (LRA) [33–35], where the nanoscale effects of the metal's free electron gas are neglected and the dielectric function of the metal is assumed to be dependent only on the angular frequency of oscillation. Even though LRA formalism successfully explains numerous plasmonic phenomena, it fails to predict or rationalize some crucial experimental observations such as plasmonic resonance blueshift or the linewidth broadening [2,36–38] observed with decreasing particle sizes. Furthermore, it has been shown that the dynamics of nanoparticle dimers [39,40] and nanogaps in plasmonic structures [41–43], whose dimensions are comparable to the electron mean-free path can be better explained through the intrinsic nonlocality of the dielectric response of the plasmonic material [41,44]. The nonlocal optical response of a metal implies that the polarization field at a point depends on the electric field at that point, as well as on the electric field in its neighborhood [41,45]. In the Fourier space, it is represented through the spatial dispersion or the wave-vector dependence of the metal's dielectric response. In an attempt to postulate these inherent size-dependent effects, several theories including first-principle approaches based on time-dependent density functional theory (TDDFT) [40,46,47] and configuration interaction formalism [48] have been developed.

The significance of quantum effects like electron tunneling and nonlocal screening at subnanometer gaps in plasmonic structures are well captured by TDDFT methods in good agreement with measurements [49]; thus, widely embraced in the field of plasmonics. Even though some strong approximations such as time-dependent local density approximation (TDLDA) allow TDDFT methods to accommodate relatively large nanosystems [39,50] than more atomistic expositions, being a volume-dependent technique, the associated computational cost and the excessive memory requirements of TDDFT approach scales swiftly with the system dimensions, hindering its feasibility beyond small particle sizes [51,52]. In response, a less demanding semiclassical quantum corrected model (QCM) which combines the electron tunneling with LRA formalism has been developed [53]. Intuitively, for interparticle separations greater than the atomic scale tunneling threshold distance, QCM and LRA approaches yield identical results [40]. Alternatively, simpler, yet scrupulous theories have been built on nonlocal hydrodynamic Drude model (NLHDM) [54,55], which are capable of an analytical characterization of quantum nonlocal effects. However, it is important to note the fact that the more complicated phenomena in the plasmonic metal, such as the inhomogeneous equilibrium electron density due to Friedel oscillations, electron spill-out, or changes in the electronic band structure [36,39] are not accounted in any analytical nonlocal model, causing their predictions to still be slightly deviated from the experimental observations.

NLHDM formalism appraises the nonlocal response of a metal, via the existence of longitudinal waves and it has been unified into generalized nonlocal optical response (GNOR) theory [56], where both the quantum effects of the propagating pressure waves and the classically induced charge diffusion kinetics are treated on an equal footing. While NLHDM-based methods are demonstrated to be consistent with TDDFT outcomes outside the tunneling regime [40,56], they surpass these *ab initio* methods in analytical capabilities and numerical efficiency as well as in their ability to account for retardation effects. It has been shown that nonlocal correction to the dipolar Mie coefficients derived in a retarded framework is identical to that based on GNOR theory [36]. Moreover, the GNOR model is accented to be capable in accounting for size-dependent Landau damping in particle shapes beyond the spheres [36], and in comparison to the NLHDM formalism, it is suggested that the GNOR theory better captures the resonance shifts, linewith broadening, and amplitude scaling visible in the Reyleigh scattering spectra [57,58]. Furthermore, it is demonstrated that the properties of molecular fluorescence enhancement in the vicinity of metallic nanostructures predicted by the GNOR formalism are in better agreement with the experimental observations than its NLHDM counterparts [59]. Nevertheless, all the nonlocal or quantum informed models employ approximations to certain extents, hence a template for accessing their performance has been suggested recently [44].

Nonlocal effects of single MNPs [54,55,60], dimers [36,39,49,52,53,56,61], MNPs on substrates [62,63], and MNP-QE hybrids with a single emitter [57,64,65] have been thoroughly studied. It has been attempted to derive the nonlocal polarizability for spherical MNPs, followed by a multipolar description in Refs. [60,66]. The influence of nonlocal effects on Förster resonance energy transfer between two molecules placed in the vicinity of an MNP has also been studied in Ref. [13]. However, the significance of the nonlocal effects on MNP-QE composite nanostructures with multiple emitters is yet to be studied. In this paper, we peruse the influence of nonlocal optical response of a spherical MNP on a composite nanostructure with a cluster of identical QEs positioned in a ring around the MNP. This paper provides a comprehensive quantum mechanical elucidation of a tunable plasmonic laser, with physical insights into the plasmon statistics of the system.

This paper is organized as follows. We first present an overview of the MNP-QE composite nanostructure in Sec. II, outlining the total energy of the system through its Hamiltonian. Then in Sec. III, we discuss how the GNOR theory quantifies the nonlocal effects arising from the free-electron gas of the MNP, to emphasize the impact of particle size on the MNP's plasmonic resonance, decay rate, and the coupling with the QEs, as opposed to the existing LRA-based explanation. In Sec. IV, we derive the equation of motion (EoM) for the complete system density matrix and simplify it with RDM approximation in Sec. IV C, to study the plasmon statistics of the system. In Sec. IV D, we explain the importance of and provide the expressions for the observables of interest of the system, in the context of plasmonic lasers. Furthermore, we introduce two measures to quantify the difference between the local and nonlocal calculations of those observables. Then in Sec. V, we numerically analyze the discussed nonlocal model for a silver MNP, varying its radius, number of coupled QEs, pumping rate, and the dielectric constant of the surrounding medium, and present the results in contrast to the LRA approach, discussing in detail the implications of the nonlocal response of the MNP on the dynamics of plasmonic lasers. We then discuss in Sec. V B, how the presented model accommodates the possible differences between the QEs and alternative structures for MNP-QE hybrids. In Sec. VI, we conclude the paper, summarizing the physical insights gained through the proposed approach.

II. OVERVIEW OF THE MODEL

We begin our description with a general subwavelength MNP-QE composite nanostructure; a spherical MNP surrounded with QEs and immersed in a homogeneous dielectric bath with real positive relative permittivity $\epsilon_{\rm b}$ (see Fig. 1). A set of identical QEs are placed along the MNP's equator in the $x-y$ plane, such that their dipole moments align with the z direction. The dipole approximation is verified to be reasonably accurate for spherical or elliptical particles with dimensions below 100 nm [33], allowing us to treat the QEs as point dipoles. The scheme is considered to be highly symmetric, with QEs coupling identically to the MNP's dipole plasmon mode and the MNP's dipole moment pointing in the z direction. Later, in Sec. V B, we discuss how our model accommodates the breaking of the indistinguishability of QEs and their symmetric placement around the MNP. We neglect the retardation effects in our description, given that the nonretarded limit for the noble metals is extended to larger particle sizes, generally beyond 10 nm radii [36].

System Hamiltonian

In our model, QEs are treated as two-level systems [22,67,68] with basis quantum states $|a\rangle$, where a denotes the energy level of interest, such that $a = g$ for the ground state and $a = e$ for the excited state. The related wave functions for the nth QE is denoted by $\varphi_{\rm na}$. The QEs are excited to higher energy states by an external electromagnetic field and they form excitons with zero ground-state energy and transition frequency $\omega_{\rm qe}^{\rm n}$ between the states $|e\rangle$ and $|g\rangle$. The excitonic coupling resulting from the Coulomb interactions among the electrons and the nuclei of different QEs can be safely neglected as we maintain an inter-QE spacing greater than 1 nm in our model [69].

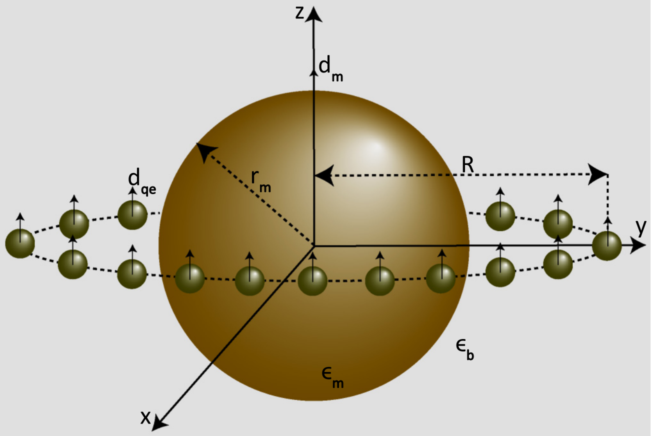

FIG. 1. Schematic of a spherical MNP coated with identical QEs placed around its equatorial orbit of radius R in $x-y$ plane. The dielectric function of the plasmonic metal is given by $\epsilon_{\rm m}$ and the entire composite nanostructure is immersed in a homogeneous bath with a real positive relative permittivity $\epsilon_{\rm b}$. Dipole moment of the MNP ($d_{\rm m}$) as well as that of all the QEs ($d_{\rm qe}$) are aligned along the z axis.

The Hamiltonian of the localized SPs in the spherical MNP is derived by associating each mode with a quantum harmonic oscillator. The presence of strong interband effects in noble metals dampen the higher order modes, ensuring that only the dipole plasmon mode is in resonance [36]. Consequently, our system reduces to an $N_{\rm qe}$ number of two-level QEs interacting with a single cavity mode, which is well described by the closed Dicke model [70] given by,

$$\hat{H}_{\rm sys} = \underbrace{\sum_{\rm n}^{N_{\rm qe}} \hbar\omega_{\rm qe}^{\rm n}\hat{\sigma}_{\rm n}^{\dagger}\hat{\sigma}_{\rm n}}_{\hat{H}_{\rm qe}} + \underbrace{\hbar\omega_{\rm pl}\hat{a}^{\dagger}\hat{a}}_{\hat{H}_{\rm pl}} + \underbrace{\sum_{\rm n}^{N_{\rm qe}} \upsilon_{\rm n}(\hat{\sigma}_{\rm n}^{\dagger}\hat{a} + \hat{\sigma}_{\rm n}\hat{a}^{\dagger})}_{\hat{H}_{\rm int}}, \quad (1)$$

where $\hat{H}_{\rm qe}$ is the Hamiltonian of the QEs and $\hat{\sigma}_{\rm n} = |\varphi_{\rm ng}\rangle\langle\varphi_{\rm ne}|$ denotes the lowering transition operator of the nth QE, whereas its hermitian conjugate $\hat{\sigma}_{\rm n}^{\dagger}$ denotes the raising transition operator. The Hamiltonian of the plasmons is given by $\hat{H}_{\rm pl}$ and the creation and annihilation operators of the plasmon field are denoted by $\hat{a}^{\dagger}$ and $\hat{a}$, respectively. The plasmon number states $|\mu\rangle$ form a complete basis set for the MNP dipole plasmons, with $\omega_{\rm pl}$ denoting the SP resonance frequency of the MNP. The interaction Hamiltonian of QEs and plasmons is denoted by $\hat{H}_{\rm int}$, with $\upsilon_{\rm n}$ denoting the coupling coefficient between the nth QE and the dipole plasmon mode. Later, in Sec. IV, we employ this Hamiltonian to arrive at the EoM which describes the time evolution of the system.

III. NONLOCAL OPTICAL RESPONSE OF METALLIC NANOSTRUCTURES

The quantum mechanical theory of plasmonic lasers is hitherto developed on the local optical response of the MNP. However, the validity range of the dipole plasmon approximation overlaps with particle sizes where the nonlocal correction becomes essential. In this section, we perform a nonlocal adjustment to the existing LRA theory, to elucidate the impact of the dimensions of the MNP.

As discussed in Sec. I, nonlocal theories of diverse sophistication levels are found in literature. Following a meticulous review, we proceed with GNOR formalism in our analysis, due to its physical transparency, analytical robustness, and numerical simplicity, delivered without compromising the accuracy within our tunneling-free, nonretarded model [36,56,57].

The nonlocal optical response of an MNP is often characterized by the wave-vector-dependent dielectric response of the metal. NLHDM formalism-based models such as GNOR method are classified as longitudinal nonlocal response models, where the transverse dielectric response of the metal $\epsilon_{\rm m}^{\rm T}(\omega)$, is given by a Drude-like dielectric function $\epsilon_{\rm m}(\omega)$, which is similar to the local response models, whereas the contribution of the free-electron response to its longitudinal dielectric function $\epsilon_{\rm m}^{\rm L}(k_{\rm L},\omega)$, is modified by a longitudinal wave-vector (k_{L})-dependent term as [45],

$$\epsilon_{\mathrm{m}}^{\mathrm{T}}(\omega) = \epsilon_{\mathrm{m}}(\omega) = \epsilon_{\mathrm{core}}(\omega) - \frac{\omega_{\mathrm{p}}^2}{\omega(\omega + i\gamma)}, \tag{2a}$$

$$\epsilon_{\mathrm{m}}^{\mathrm{L}}(k_{\mathrm{L}}, \omega) = \epsilon_{\mathrm{core}}(\omega) - \frac{\omega_{\mathrm{p}}^2}{\omega(\omega + i\gamma) - \zeta^2 k_{\mathrm{L}}^2}, \tag{2b}$$

where ω_{p} and γ are the bulk plasmon frequency and the relaxation constant of the bulk material, respectively. In the presence of interband effects that are inherent to the plasmonic metals, the response of the bound electrons $\epsilon_{\mathrm{core}}(\omega)$ can be calculated using the bulk dielectric function of the metal, whereas it can safely be assumed as 1 for the metals with negligible interband contribution [71]. For $\omega \gg \gamma$, $\zeta^2 = (3/5)\upsilon_{\mathrm{F}}^2$ with υ_{F} denoting the Fermi velocity of the metal.

Nonlocal correction to the MNP-QE coupling term

In an attempt to apprehend the impact of the optical response of the MNP on its plasmonic resonance frequency, decay rate, and the coupling coefficient between the MNP and the QEs, it is quantified that the nonlocal response of the MNP manifests itself into the MNP's polrizability, as a rescaling to either the metal's Drude-like dielectric function from ϵ_{m} to $\epsilon_{\mathrm{m}}^{\mathrm{NL}} = \epsilon_{\mathrm{m}}(1 + \delta_{\mathrm{NL}})^{-1}$ or the bath permittivity from ϵ_{b} to $\epsilon_{\mathrm{b}}^{\mathrm{NL}} = \epsilon_{\mathrm{b}}(1 + \delta_{\mathrm{NL}})$, with δ_{NL} being the nonlocal correction [63]. This redefines the Frölich condition for plasmon resonance frequency ω_{pl} as,

$$\mathrm{Re}[\epsilon_{\mathrm{m}}(\omega_{\mathrm{pl}})] \approx -2\epsilon_{\mathrm{b}}\,\mathrm{Re}[1 + \delta_{\mathrm{NL}}]. \tag{3}$$

Consequently, this modifies the MNP-QE coupling term as [57],

$$\upsilon_{\mathrm{n}}^{\mathrm{NL}} = \upsilon_{\mathrm{n}}^{\mathrm{L}}\sqrt{\mathrm{Re}[1 + \delta_{\mathrm{NL}}]}, \tag{4}$$

where $\upsilon_{\mathrm{n}}^{\mathrm{NL}}$ denotes the nonlocally corrected coupling coefficient between the nth QE and the dipole plasmon mode. Its local counterpart $\upsilon_{\mathrm{n}}^{\mathrm{L}}$ is given by [34,57],

$$\upsilon_{\mathrm{n}}^{\mathrm{L}} = \frac{d_{\mathrm{qe}}^{\mathrm{n}}}{\hbar}\frac{(3\cos\theta_{\mathrm{n}} - 1)}{R_{\mathrm{n}}^3}\sqrt{\frac{3\hbar\eta r_{\mathrm{m}}^3}{4\pi\epsilon_0}}, \tag{5}$$

where r_{m} is the radius of the MNP and R_{n} is the distance between the centers of the MNP and nth QE. The dipole moment of the nth QE is given by d_{qe}^{n}, while ϵ_0 stands for the permittivity of the free space. The spatially constant local dielectric function of the plasmonic metal, which is identical to the transverse dielectric response in NLHDM formalism [see Eq. (2a)], is identified by $\epsilon_{\mathrm{m}}(\omega)$. The modified Planck's constant is denoted by $\hbar$ and $\eta = (d\,\mathrm{Re}[\epsilon_{\mathrm{m}}(\omega)]/d\omega)^{-1}_{(\omega=\omega_{\mathrm{pl}})}$. The angle between the direction of the dipole moment of nth QE and the line connecting its center to that of the MNP is denoted by θ_{n}.

The lifetime of SPs becomes finite due to the energy losses from the SP mode. This plasmon decay rate is given by [34],

$$\gamma_{\mathrm{pl}} \approx 2\eta\,\mathrm{Im}[\epsilon_{\mathrm{m}}(\omega_{\mathrm{pl}})]. \tag{6}$$

By applying the convection-diffusion equation and Fick's law for the current density in the hydrodynamic equation, the nonlocal correction of the GNOR model is quantified as [36],

$$\delta_{\mathrm{NL}} = \frac{\epsilon_{\mathrm{m}}(\omega) - \epsilon_{\mathrm{core}}(\omega)}{\epsilon_{\mathrm{core}}(\omega)}\frac{J_1(k_{\mathrm{L}}r_{\mathrm{m}})}{k_{\mathrm{L}}r_{\mathrm{m}}J_1'(k_{\mathrm{L}}r_{\mathrm{m}})}, \tag{7}$$

where J_1 is the Bessel function of first kind and order 1, whereas J_1' is its first-order differential with respect to the argument $k_{\mathrm{L}}r_{\mathrm{m}}$. J_1' can be found using the following identity:

$$\frac{\partial J_v(z)}{\partial z} = J_{v-1}(z) - \frac{v}{z}J_v(z). \tag{8}$$

The longitudinal wave vector k_{L}, which accounts for the nonlocal effects of the MNP, is given by [36],

$$k_{\mathrm{L}}^2 = \frac{\epsilon_{\mathrm{m}}(\omega)}{\xi^2(\omega)}. \tag{9}$$

The nonlocal parameter ξ is modeled in GNOR formalism as,

$$\xi^2(\omega) = \frac{\epsilon_{\mathrm{core}}(\omega)[\zeta^2 + D(\gamma - i\omega)]}{\omega(\omega + i\gamma)}, \tag{10}$$

with the diffusion parameter D given by [56],

$$D = \frac{4}{15}\frac{\gamma}{\omega^2 + \gamma^2}\upsilon_{\mathrm{F}}^2. \tag{11}$$

IV. QUANTUM MECHANICAL THEORY OF MNP-QE COMPOSITE NANOSTRUCTURES

The dynamics of conventional lasers are often explained through photon statistics. Equivalently, the input-output relationship of a plasmonic laser, as well as the coherence of its plasmonic field, can be learnt from the plasmon statistics. In this section, we attempt to analyze the MNP-QE composite nanostructure in Fig. 1 using the density matrix formalism and to solve the quantum master equation, which enables us to study the plasmon statistics of the system.

A. Equation of motion with dissipative correction

The MNP-QE composite nanostructure forms an open quantum system, the dynamics of which can be described using the Von-Neumann EoM for the system density matrix $\hat\rho(t)$ with dissipative corrections as follows:

$$\frac{d\hat\rho}{dt} = \frac{i}{\hbar}[\hat\rho, \hat H_{\mathrm{sys}}] - \mathcal{D}\hat\rho. \tag{12}$$

Owing to its trace preserving and completely positive nature as mandated by quantum mechanical principles, the dissipative part is often described by Lindblad form as,

$$\mathcal{D}\hat\rho = \sum_{\mathrm{m}}\frac{k_{\mathrm{m}}}{2}([\hat A_{\mathrm{m}}^\dagger\hat A_{\mathrm{m}}, \hat\rho]_+ - 2\hat A_{\mathrm{m}}\hat\rho\hat A_{\mathrm{m}}^\dagger). \tag{13}$$

The various terms in this expression can be understood as follows. For QE excited-state decay, $\hat A_{\mathrm{m}}^\dagger = \hat\sigma_{\mathrm{n}}^\dagger$ and $k_{\mathrm{m}} = \gamma_{\mathrm{qe}}^{\mathrm{n}}$ denote the spontaneous decay rate of the QE. Similarly, for plasmon decay, $\hat A_{\mathrm{m}}^\dagger = \hat a^\dagger$ and $k_{\mathrm{m}} = \gamma_{\mathrm{pl}}$ denote the common plasmon decay rate. In the cases where QE excitation is achieved via incoherent pumping, the effect of pumping is also described in Lindblad form as shown in Eq. (13), but with slightly altered notation, where $\hat A_{\mathrm{m}}^\dagger = \hat\sigma_{\mathrm{n}}$ and $k_{\mathrm{m}} = p_{\mathrm{n}}$ denote the pumping rate.

B. System density matrix

The following formalism for the system density matrix $\hat{\rho}(t)$ considers an identical set of QEs coupled to the MNP plasmons. This results in a common resonance frequency ω_{qe}, pumping rate p, decay rate γ_{qe}, and MNP-QE coupling υ for all the QEs ($\omega_{\mathrm{qe}}^{\mathrm{n}} = \omega_{\mathrm{qe}}$, $p_{\mathrm{n}} = p$, $\gamma_{\mathrm{qe}}^{\mathrm{n}} = \gamma_{\mathrm{qe}}$, and $\upsilon_{\mathrm{n}} = \upsilon$ $\forall n$). This standard treatment which incorporates the indistinguishability of the QEs has been introduced in Ref. [30] and adopted in Refs. [31,32].

We then obtain the basis states for the exciton-plasmon mixed states by the direct product of the basis states of the pure states as,

$$|\alpha\mu\rangle = |\alpha\rangle \otimes |\mu\rangle = \prod_{\mathrm{n}}^{N_{\mathrm{qe}}} \prod_{\mathrm{a=g,e}} |\varphi_{\mathrm{na}}\rangle \otimes |\mu\rangle, \tag{14}$$

where $|\alpha\rangle = \prod_{\mathrm{a=g,e}} |\varphi_{\mathrm{na}}\rangle$ and $|\mu\rangle$ are the basis states for the set of QEs and plasmons, respectively. Then, the elements of the density matrix to be determined are given by,

$$\rho_{\alpha\mu,\beta\nu}(t) = \mathrm{tr}\{\hat{\rho}(t)|\beta\nu\rangle\langle\alpha\mu|\} \equiv \langle|\beta\nu\rangle\langle\alpha\mu|\rangle. \tag{15}$$

Accordingly, we apply Eqs. (1), (13), and (15) in Eq. (12) to arrive at Eq. (16), which is the EoM for the time evolution of the complete system density matrix (the detailed derivation can be found in Appendix A). The numbers n_{cd} represent the number of QEs from product state $|\alpha\rangle$ in state φ_{c} and the number of QEs from product state $|\beta\rangle$ in state φ_{d} simultaneously.

$$\begin{aligned}\frac{\partial}{\partial t}\rho_{\mathrm{n}}^{(\mu,\nu)} = & -i(n_{\mathrm{eg}} - n_{\mathrm{ge}})\omega_{\mathrm{qe}}\rho_{\mathrm{n}}^{(\mu,\nu)} - i\omega_{\mathrm{pl}}(\mu-\nu)\rho_{\mathrm{n}}^{(\mu,\nu)} - \frac{\gamma_{\mathrm{pl}}}{2}\left[(\mu+\nu)\rho_{\mathrm{n}}^{(\mu,\nu)} - 2\sqrt{(\mu+1)(\nu+1)}\rho_{\mathrm{n}}^{(\mu+1,\nu+1)}\right]\\ & - \frac{\gamma_{qe}}{2}\left[(2n_{\mathrm{ee}} + n_{\mathrm{eg}} + n_{\mathrm{ge}})\rho_{\mathrm{n}}^{(\mu,\nu)} - 2n_{\mathrm{gg}}\rho_{\mathrm{n_{gg}-1,n_{ee}+1}}^{(\mu,\nu)}\right] - \frac{p}{2}\left[(n_{\mathrm{eg}} + n_{\mathrm{ge}} + 2n_{\mathrm{gg}})\rho_{\mathrm{n}}^{(\mu,\nu)} - 2n_{\mathrm{ee}}\rho_{\mathrm{n_{ee}-1,n_{gg}+1}}^{(\mu,\nu)}\right]\\ & + i\upsilon \sum_{\mathrm{a=\{g,e\}}} \left(\sqrt{\nu+1}n_{\mathrm{ae}}\rho_{\mathrm{n_{ae}-1,n_{ag}+1}}^{(\mu,\nu+1)} - \sqrt{\mu+1}n_{\mathrm{ea}}\rho_{\mathrm{n_{ea}-1,n_{ga}+1}}^{(\mu+1,\nu)} + \sqrt{\nu}n_{\mathrm{ag}}\rho_{\mathrm{n_{ag}-1,n_{ae}+1}}^{(\mu,\nu-1)} - \sqrt{\mu}n_{\mathrm{ga}}\rho_{\mathrm{n_{ga}-1,n_{ea}+1}}^{(\mu-1,\nu)}\right).\end{aligned} \tag{16}$$

This exact density matrix would have $4^{N_{\mathrm{qe}}} \times (N_{\mathrm{pl}}+1)^2$ elements with N_{pl} denoting the highest plasmon level under consideration. Since the matrix exponentially grows with N_{qe}, the exact computation becomes extremely computationally demanding [31,32]. Thus, claiming the homogeneity of the QEs and their symmetrical arrangement around the MNP, hereafter we adopt the plasmon RDM formalism, for a convenient analytical description.

C. Plasmon reduced-density matrix (RDM) formalism

Following the Lamb-Scully laser theory [72], a closed set of EoM can be derived for the elements of the plasmon density matrix given by [32],

$$\rho_{\mu,\nu}(t) = \mathrm{tr}\{\hat{\rho}(t)|\nu\rangle\langle\mu|\} \equiv \langle|\nu\rangle\langle\mu|\rangle. \tag{17}$$

Manipulating the general EoM given in Eq. (12), we then arrive at the following equation for the time evolution of the plasmon density matrix elements (the detailed derivation can be found in Appendix B):

$$\begin{aligned}\frac{\partial}{\partial t}\rho_{\mu\nu} = & -i\left[(\mu-\nu)\omega_{\mathrm{pl}} - i(\mu+\nu)\frac{\gamma_{\mathrm{pl}}}{2}\right]\rho_{\mu,\nu}\\ & + \gamma_{\mathrm{pl}}\sqrt{(\mu+1)(\nu+1)}\rho_{\mu+1,\nu+1}\\ & - i\upsilon\sum_{\mathrm{n}}(\sqrt{\mu+1}\rho_{\mathrm{g}\mu+1,\mathrm{e}\nu} + \sqrt{\mu}\rho_{\mathrm{e}\mu-1,\mathrm{g}\nu}\\ & - \sqrt{\nu}\rho_{\mathrm{g}\mu,\mathrm{e}\nu-1} - \sqrt{\nu+1}\rho_{\mathrm{e}\mu,\mathrm{g}\nu+1}).\end{aligned} \tag{18}$$

Equation (18) introduces a set of additional expectation values ($\rho_{\alpha\mu,\beta\nu}$), which necessitates the derivation of a separate set of equations of motion. The subsequent EoM generates a hierarchy of additional expectation values. To obtain a closed set of equations, we truncate the expectation values beyond the second order [73] to arrive at the following steady state recurrence relation $f(\mu)$ between $P_{\mu-1}$ and P_μ, which denote the population of plasmon states $|\mu-1\rangle$ and $|\mu\rangle$, respectively:

$$f(\mu) = \frac{\rho_{\mu,\mu}}{\rho_{\mu-1,\mu-1}} = \frac{P_\mu}{P_{\mu-1}} = \frac{\varrho_\mu}{\kappa_\mu + \mu\gamma_{\mathrm{pl}}}, \tag{19}$$

where

$$\kappa_\mu = \frac{N_{\mathrm{qe}}\mu l_\mu \gamma_{\mathrm{qe}}}{p + \gamma_{\mathrm{qe}} + 2\mu l_\mu}, \tag{20a}$$

$$\varrho_\mu = \frac{N_{\mathrm{qe}}\mu l_\mu p}{p + \gamma_{\mathrm{qe}} + 2\mu l_\mu}, \tag{20b}$$

$$l_\mu = \frac{\upsilon^2\left(\gamma_{\mathrm{qe}} + p + \frac{\gamma_{\mathrm{pl}}}{2\mu}\right)}{(\omega_{\mathrm{pl}} - \omega_{\mathrm{qe}})^2 + \left(\frac{\gamma_{\mathrm{qe}}+p}{2} + \frac{\gamma_{\mathrm{pl}}}{4\mu}\right)^2}. \tag{20c}$$

Both the terms κ_μ and ϱ_μ are a combination of QE decay rate γ_{qe}, QE pumping rate p, as well as l_μ, which describes the energy transfer between a QE and the plasmon states μ and $\mu-1$. Being the product of QE pumping rate p_{n} and the energy-transfer rate μl_μ into the μth excited plasmon state, divided by the total rate and summed over the QEs, we interpret ϱ_μ as a plasmon pumping rate. Similarly, κ_μ can be understood as a plasmon decay rate. Equation (19) suggests that, at the steady state, the μth excited plasmon state decays at a rate of $\kappa_\mu + \mu\gamma_{\mathrm{pl}}$ and it is compensated by pumping plasmons to the μth plasmon state at a rate of ϱ_μ.

Employing the fact that the plasmon state occupation is a probabilistic value, we then derive an expression for the population of the zeroth plasmon state P_0, which would serve as an initial conditions for the repetitive relation in Eq. (19) as,

$$P_0 = \frac{1}{1 + \sum_{l=1}^{N_{\mathrm{pl}}}\prod_{r=1}^{l} f(r)}. \tag{21}$$

D. Observable quantities for plasmon statistics

In this section, we present the expressions for the observable quantities of interest associated with the MNP-QE composite nanostructure.

The probability of finding μ number of excited plasmons in the system is signified through the plasmon state population P_μ, given in Eq. (19). For a lasing MNP-QE composite nanostructure, the said plasmon state occupation mimics a Poisson distribution characterized by the mean plasmon number [74]. In an elementary statistical sense, mean plasmon number can be calculated as,

$$\mathcal{N}_{\mathrm{pl}} = \sum_{\mu}^{N_{\mathrm{pl}}} \mu P_\mu. \tag{22}$$

Even though it is unlimited in theory, Eqs. (21) and (22) accent on a befitting upper limit for the plasmon number states (N_{pl}) for our discussion. Considering the symmetry of a Poissonian curve around its peak, we estimate N_{pl} to be twice $\mathcal{M}_{\mathrm{peak}}$, which is the plasmon state corresponding to the peak plasmon population. These relations can be expressed as [31],

$$\mathcal{M}_{\mathrm{peak}} = \frac{N_{\mathrm{qe}}(p-\gamma_{\mathrm{qe}})}{2\gamma_{\mathrm{pl}}} - \frac{(p+\gamma_{\mathrm{qe}})^2}{8\upsilon^2}, \tag{23a}$$

$$N_{\mathrm{pl}} = 2 \times \mathcal{M}_{\mathrm{peak}}. \tag{23b}$$

We then obtain the maximum pump rate $p_{\max}$, beyond which the coherent plasmon statistics begin their transition toward the chaotic regime due to the initiation of quenching effects, which act against the intended lasing action in the weak coupling regime. At pump rates greater than $p_{\max}$, the peak of the plasmon state distribution begins to shift toward the lower plasmon states. According to Eq. (23a), $p_{\max}$ can be expressed as,

$$p_{\max} = \frac{2\upsilon^2 N_{\mathrm{qe}}}{\gamma_{\mathrm{pl}}} - \gamma_{\mathrm{qe}}. \tag{24}$$

The second-order coherence function [$g^{(2)}(t)$], which is an intensity correlation function of the radiation, is an unambiguous measure of the coherence of a given radiation [75]. It is defined as the ratio of plasmon density correlation and the square of the plasmon density [35]. We express this in terms of plasmon number states as,

$$g^{(2)}(t) = \frac{\langle \hat{a}^\dagger \hat{a}^\dagger \hat{a}\hat{a}|n\rangle\langle n|\rangle(t)}{\langle \hat{a}^\dagger \hat{a}|n\rangle\langle n|\rangle(t)^2} = \frac{\sum_\mu \mu(\mu-1)P_\mu}{\left(\sum_\mu \mu P_\mu\right)^2}. \tag{25}$$

For coherent light, $g^{(2)}(0)$ approaches the Poisson limit of 1 [30,75], hence it can be used to verify the lasing operation of the MNP-QE composite nanostructure.

To demonstrate the deviation between the local and nonlocal assessments of these observables, we then define,

$$\mathcal{F}_1(A) = |A_{\mathrm{NL}} - A_{\mathrm{LRA}}|, \tag{26a}$$

$$\mathcal{F}_2^x(B) = \frac{1}{(x_2 - x_1)} \sum_{x=x_1}^{x_2} |B(x)_{\mathrm{NL}} - B(x)_{\mathrm{LRA}}|, \tag{26b}$$

where the subscripts NL and LRA denote the evaluations of the function according to GNOR and LRA formalisms, respectively. $\mathcal{F}_1(A)$ describes the absolute difference between the GNOR- and LRA-based assessments of the observable A, whereas $\mathcal{F}_2^x(B)$ describes the mean value of $\mathcal{F}_1(B)$, calculated over a range of x.

V. RESULTS AND DISCUSSION

In this section, we evaluate the impact of the nonlocal optical response of the MNP on the dynamics of the MNP-QE composite nanostructure depicted in Fig. 1, and compare the results against those generated using the local optical response of the MNP. We use a silver MNP, the dielectric response of which is calculated according to Eq. (2), using experimentally validated bulk dielectric function $\epsilon_{\exp}(\omega)$ [76] and the relation $\epsilon_{\mathrm{core}}(\omega) = \epsilon_{\exp}(\omega) + \omega_{\mathrm{p}}^2/[\omega(\omega + i\gamma)]$ [57]. The bulk parameters for silver are $\omega_{\mathrm{p}} = 8.99$ eV, $\gamma = 0.025$ eV, and $v_{\mathrm{F}} = 1.39 \times 10^6\,\mathrm{ms}^{-1}$ [36]. All the QEs are identical with $d_{\mathrm{qe}} = 0.7\,e\,\mathrm{nm}$ and $\gamma_{\mathrm{qe}} = 50\,\mu\mathrm{eV}$. For infinitesimal inter-QE interactions, we maintained not less than 1 nm separation in deciding the maximum number of QEs placed in an orbital ($N_{\mathrm{qe}}^{\max}$). Unless otherwise stated, throughout this section we use the parameters $\epsilon_{\mathrm{b}} = 3$, $R = 12.5\,\mathrm{nm}$, $p = 50\,\mathrm{meV}$, $N_{\mathrm{qe}} = N_{\mathrm{qe}}^{\max} = \lfloor 2\pi R \rfloor$ and assume MNP and QEs to be in resonance.

A. Numerical results

In Fig. 2, we demonstrate how the nonlocally appraised dynamics of a plasmonic laser would diverge from its local counterpart with the increase of the coupled QEs. For any given bath permittivity value, plasmon resonance frequency ω_{pl} of noble metals blueshifts with decreasing particle size and, as evident from Eq. (3), the shift intensifies with bath permittivity. On the other hand, the plasmon decay rate γ_{pl} shows an oscillatory behavior against the bath permittivity at different MNP radii. Extracting the commonly used parameters for plasmonic laser simulations, we used $\epsilon_{\mathrm{b}} = 3$, $r_{\mathrm{m}} = 10\,\mathrm{nm}$, $R_{\mathrm{n}} = 12.5\,\mathrm{nm}$, $p_{\mathrm{n}} = 50\,\mathrm{meV}$ for all n [30,32] and QE resonance frequency is fixed allowing a detuning of 20 meV with the locally calculated plasmon resonance.

As shown in Fig. 2(a), for all N_{qe} values, the peak of the nonlocally calculated plasmon population shifts toward lower plasmon states due to the inherent nonlocality, which results in faster decaying plasmons (due to surface-enhanced Landau damping) and a weaker coupling between the MNP and the QEs than it is suggested by the LRA approach. While Eq. (23a) predicts this behavior, the lower $\mathcal{M}_{\mathrm{peak}}$ limits the spread of the nonlocal curves and, together with the probabilistic nature of plasmon population causes a slight increase of the occupation of lower plasmon states. The said shift marginally increases with the number of coupled emitters. At higher pumping rates, the plasmon distribution spreads wider, moving toward higher plasmon states. It can be seen that the nonconformity between the two approaches intensifies at higher p values. An overestimation of mean plasmon number is evident for the LRA approach in Fig. 2(b). The local and nonlocal curves further part from each other with increasing p or N_{qe}. While Fig. 2(c) demonstrates the expected enhancement in coherence with increasing number of coupled emitters, the nonlocal curve suggests that, at smaller N_{qe}

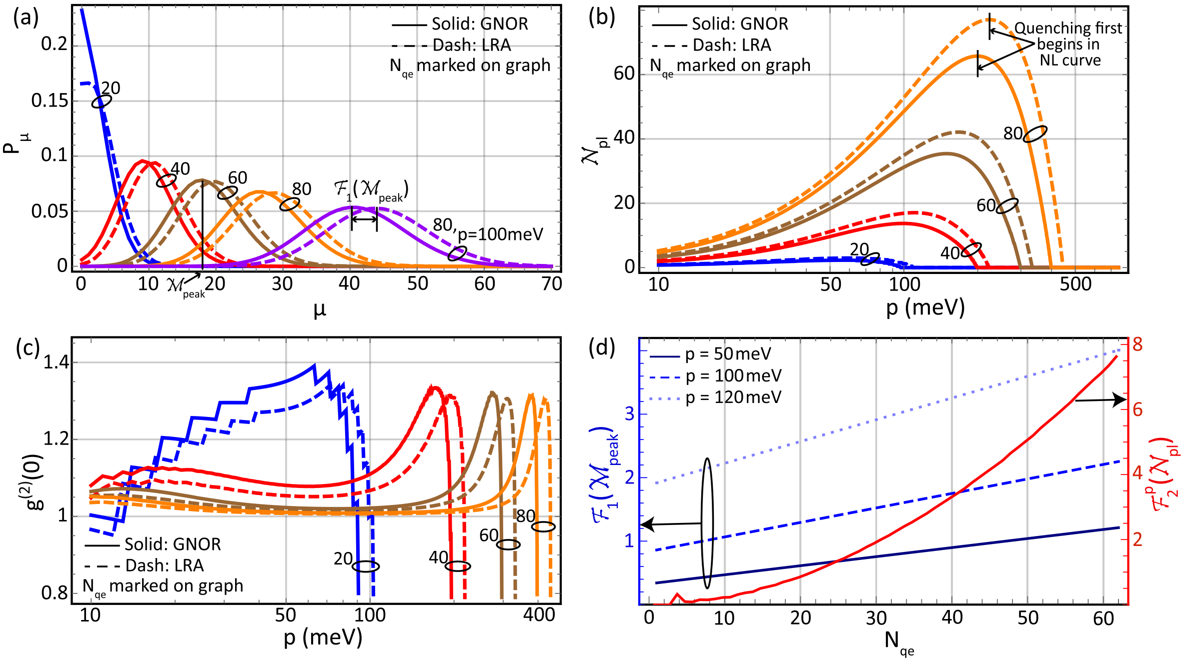

FIG. 2. Significance of nonlocal effects at different numbers of coupled QEs. QEs placed in a $R = 12.5$ nm equatorial orbital of a silver MNP of $r_m = 10$ nm, immersed in a bath of $\epsilon_b = 3$. Solid and dashed curves denote results for GNOR and LRA formalisms, respectively. (a) Plasmon state distribution. (b) Mean plasmon number against pumping rate. (c) Second-order coherence function against pumping rate. (d) Variation of $\mathcal{F}_1(\mathcal{M}_{max})$ (blue) and $\mathcal{F}_2^p(\mathcal{N}_{pl})$ (red) with the number of QEs (arrows point the curves toward their respective axes).

values, the emitted light is bunched than it is forecasted in LRA curve. This loss of coherence and the degradation of mean plasmon number at higher p are direct consequences of the quenching effect of the incoherent pump, which manifests itself into the nonlocal curve at lower pumping rates.

The difference between the GNOR and LRA calculations of $\mathcal{M}_{peak}$ [$\mathcal{F}_1(\mathcal{M}_{peak})$] and that of the input-output curve, averaged over a range of pumping rate [$\mathcal{F}_2^p(\mathcal{N}_{pl})$], is depicted in Fig. 2(d), against the number of QEs coupled to the MNP. $\mathcal{F}_1(\mathcal{M}_{peak})$ increases linearly with N_{qe}, with a greater gradient at higher pumping rates. This results in an exponential growth of $\mathcal{F}_2^p(\mathcal{N}_{pl})$ against N_{qe}, confirming the significance of nonlocal effects with increasing N_{qe}.

Due to the quenching phenomena, the plasmon statistics transition from a coherent region to a chaotic region, causing the plasmon state population distribution to alter with resemblance to a thermal distribution at higher pumping rates. This results in a leftward shift of the entire plasmon population curve. In Fig. 3(a), we explore this behavior exhibited in Eq. (24), in different bath permittivities and MNP radii. Consistent with the implications of Figs. 2(b) and 2(c), Fig. 3(a) shows that the nonlocally assessed p_{max} is lower than its LRA assessment for all r_m and ϵ_b, confirming that the nonlocal approach picks up quenching effects earlier in the pumping rate scale. A significant deviation between local and nonlocal estimations of p_{max} can be seen at lower r_m and ϵ_b. We then study the variation of $\mathcal{F}_1(p_{max})$ with the number of QEs in the hybrid system. Figure 3(b) suggests that $\mathcal{F}_1(p_{max})$ increases linearly with N_{qe} and for any N_{qe}, the difference between the local and nonlocal assessment is larger at lower r_m and ϵ_b.

In Figs. 4(a)–4(f), we analyze an MNP-QE composite nanostructure at several MNP radii and bath permittivity values extensively used in the context of plasmonic lasers. Subfigures 4(a)–4(c) are evaluated at $\epsilon_b = 3$, while Subfigures 4(d)–4(f) are evaluated at $r_m = 7$ nm. Figures 4(a) and 4(d) demonstrate the variation of mean plasmon number against

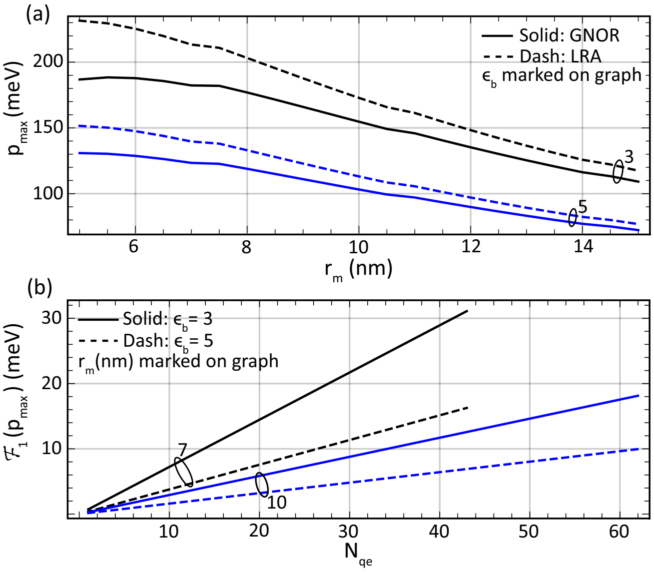

FIG. 3. Local and nonlocal evaluation of maximum pump rate before the plasmon state population switches from a Poisson to a thermal distribution. (a) Variation of maximum pump rate with MNP radius at different bath permittivities. Solid and dashed curves denote results for GNOR and LRA formalisms, respectively. (b) Variation of $\mathcal{F}_1(p_{max})$ with the number of QEs at different MNP radii and bath permittivities.

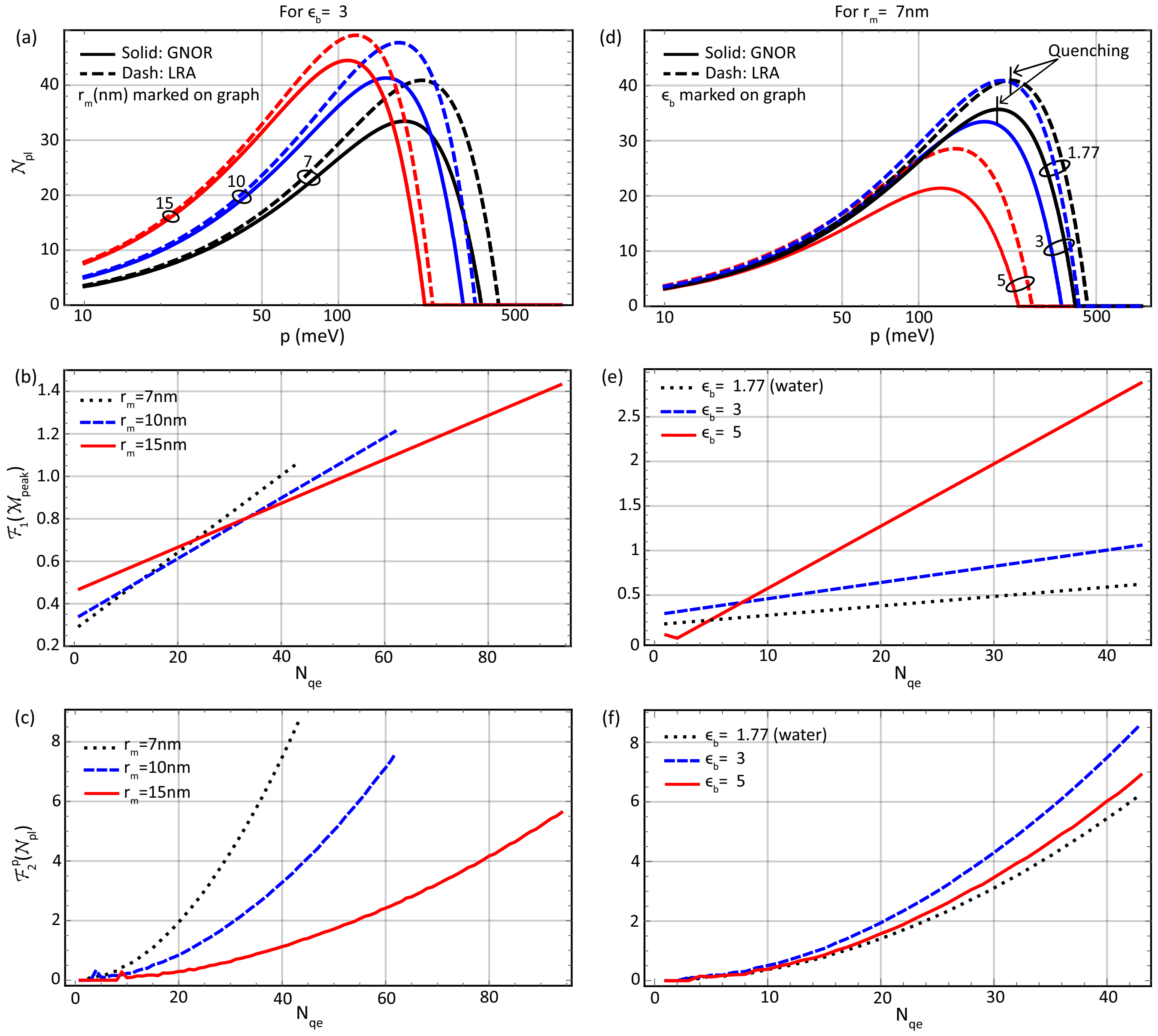

FIG. 4. Deviation of local and nonlocal dynamics of MNP-QE composite nanostructure at different MNP radii and bath permittivities. Subfigures (a), (b), and (c) are evaluated at $\epsilon_b = 3$, whereas (d), (e), and (f) are evaluated at $r_m = 7$ nm. (a), (d) Mean plasmon number against pumping rate, with $N_{\rm qe}^{\rm max}$. Solid and dashed curves denote results for GNOR and LRA formalisms, respectively. (b), (e) Variation of $\mathcal{F}_1(\mathcal{M}_{\rm max})$ with the number of QEs. (c), (f) Variation of $\mathcal{F}_2^{\rm p}(\mathcal{N}_{\rm pl})$ with the number of QEs.

the pumping rate for an MNP-QE composite nanostructure with $N_{\rm qe}^{\rm max}$ of QEs. It can be seen that the mean plasmon number is overestimated by the LRA formalism, for all the considered $r_{\rm m}$ and $\epsilon_{\rm b}$ values. In the pumping rate scale, the loss of coherence is first introduced at higher $r_{\rm m}$ and $\epsilon_{\rm b}$. This has to be understood as a sequel of increased $N_{\rm qe}^{\rm max}$ at higher $r_{\rm m}$. The GNOR approach suggests that this incoherent nature becomes visible at lower pump rates than those suggested by the LRA approach.

Subfigures 4(b) and 4(e) demonstrate the variation of $\mathcal{F}_1(\mathcal{M}_{\rm peak})$ with $N_{\rm qe}$, at different $r_{\rm m}$ and $\epsilon_{\rm b}$, respectively. While both subfigures demonstrate a linear increase of $\mathcal{F}_1(\mathcal{M}_{\rm peak})$ with $N_{\rm qe}$, it is evident that the nonlocal effects are prominent at smaller $r_{\rm m}$ and greater $\epsilon_{\rm b}$ values. The average difference between the nonlocally and locally assessed input-output curves $[\mathcal{F}_2^{\rm p}(\mathcal{N}_{\rm pl})]$ against $N_{\rm qe}$ is shown in Figs. 4(c) and 4(f). For all the $r_{\rm m}$ and $\epsilon_{\rm b}$ values, $\mathcal{F}_2^{\rm p}(\mathcal{N}_{\rm pl})$ exponentially grows with $N_{\rm qe}$, suggesting that when the MNP is coupled with many QEs, the nonlocal effects become significant even at larger $r_{\rm m}$ values. It is observable that among the considered parameter range, the nonlocal effects are least significant at $r_{\rm m} = 15$ nm and $\epsilon_{\rm b} = 1.77$ (corresponds to water); however, there still presents an indisputable nonconformity between the GNOR and LRA predictions of plasmon statistics with increasing $N_{\rm qe}$.

In Fig. 5, we study the tunability of the MNP-QE composite nanostructure with respect to the properties of the QEs as well as their separation from the surface of the MNP. Figure 5(a) shows the variation of the mean plasmon number with the detuning of resonance frequencies between the MNP and QEs. The curves are symmetric around the vertical axis as

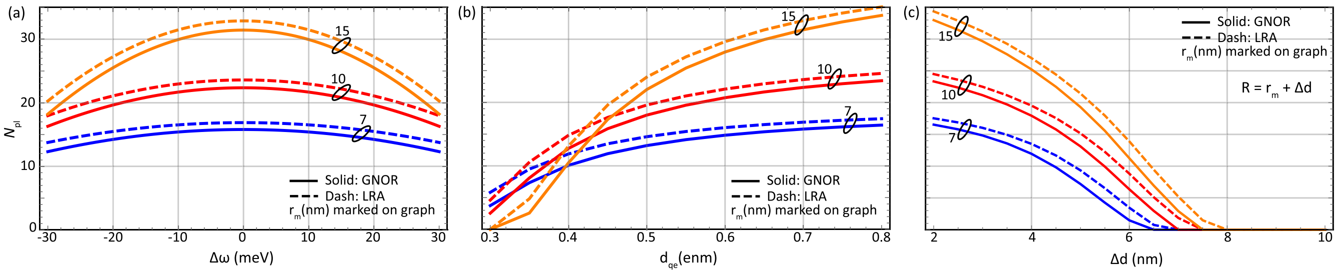

FIG. 5. Comparison of mean plasmon number variation with properties of the quantum emitters and their separation from the MNP surface at different MNP radii. Solid and dashed curves denote results for GNOR and LRA formalisms, respectively. (a) Mean plasmon number against MNP-QE resonance detuning. (b) Mean plasmon number against QE dipole moment. (c) Mean plasmon number against MNP-QE surface separation distance.

seen in Eq. (20c). Higher MNP radii are inherently associated with higher $N_{\rm qe}^{\rm max}$ and, as a result, they exhibit a drastic drop of mean plasmon number when QEs are further detuned with the MNP. According to Eqs. (4) and (5), stronger QE dipole moments strengthen the MNP-QE coupling and Fig. 5(b) confirms that this is bound to be amplified at higher MNP radii, due to the enlarged $N_{\rm qe}^{\rm max}$. We then study the variation of the mean plasmon number against the surface separation Δd between MNP and QEs, defining $R = r_{\rm m} + \Delta d$. The inverse R^3 dependence of υ appearing in Eq. (5) suggests a diminishingly weaker coupling at larger separations between the MNP and QEs, which in turn causes its output to decrease when Δd increases. The GNOR theory predicts an even weaker MNP-QE coupling than its LRA counterpart, causing the system output to be totally quenched at a distance Δd, which is smaller than its LRA prediction. Interestingly, all figures Fig. 5(a)–5(c) suggest that the gap between the local and nonlocal assessments is nearly uniform, regardless of the size of the MNP.

B. Beyond the assumptions of indistinguishability and the symmetric placement of QEs

The plasmonic laser theory we presented so far is based on the MNP-QE composite nanostructure, depicted in Fig. 1. In this section, we would discuss the robustness of the presented model against the possible dissimilarities among the QEs, as well as alternative arrangements of QEs around the MNP.

Due to the plausible minor differences arising during the manufacturing process, it is possible for the QEs to not be completely identical to each other. In such cases, we could interpret $\omega_{\rm qe}$, $\gamma_{\rm qe}$, and υ as the statistical averages of the QE resonance frequencies, decay rates, and the couplings to the MNP, respectively.

Furthermore, our model is flexible over the placement of QEs and can be extended for MNP-QE hybrid structures beyond the scheme illustrated in Fig. 1. For each nanostructure, an average value for υ can be calculated according to its geometry. For example, assuming a spherical distribution, an average MNP-QE coupling for a shell of QEs around the MNP (see Fig. 6) can be calculated as,

$$\upsilon_{\rm avg} = \frac{1}{V}\int_0^{2\pi}\int_0^{\pi}\int_{r_1}^{r_2} \upsilon_{\rm n}^{\rm NL}\, R_{\rm n}^2 \sin\theta\, dR_{\rm n} d\theta d\varphi, \tag{27}$$

where r_1, r_2, and V denote the inner and outer radii and the volume of the QE shell, respectively. We would emphasize that Eq. (27) can accommodate the variations of the magnitude as well as the orientation of $d_{\rm qe}^{\rm n}$.

VI. SUMMARY AND CONCLUSION

In this paper, we analyzed an incoherently excited equatorial ring of identical QEs, uniformly placed around a spherical MNP, in the context of plasmonic lasers. We performed a rigorous comparison between the observable quantities calculated according to the LRA and GNOR formalisms. We derived the complete EoM for the time evolution of the MNP-QE composite nanostructure and simplified it using the plasmon RDM approximation. We then introduced expressions for observables of interest of the system at steady state. Afterward, we numerically studied the plasmon state population distribution, mean plasmon number, and the second-order intensity correlation function of a MNP-QE hybrid structure containing a silver MNP at different MNP radii, bath permittivities, QE dipole moment, MNP-QE resonance detuning, and MNP-QE surface separation distance, varying the number of emitters coupled to the MNP.

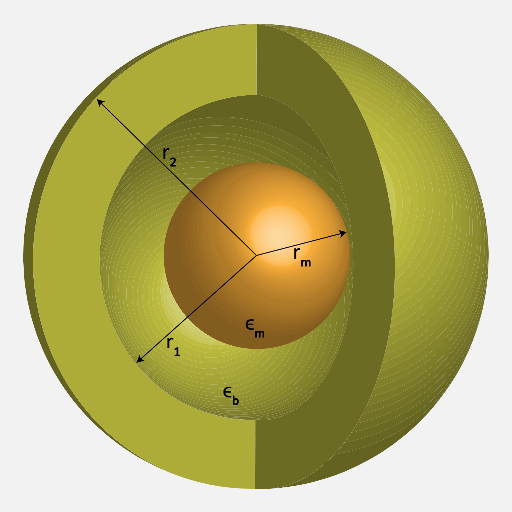

FIG. 6. An alternative arrangement of QEs around a spherical MNP. QEs are randomly placed in a concentric spherical shell with inner radius r_1 and outer radius r_2.

It is established in previous studies that the nonlocal approaches better capture the size-dependent effects of MNPs at nanoscale. In this paper, we demonstrated that the nonconformity between the local and nonlocal assessments of plasmonic statistics grows with the number of emitters, suggesting that the nonlocal screening becomes notable even at larger particle sizes when coupled with an increasing number of emitters. Furthermore, our study unveiled that, for some bath permittivities commonly used for plasmonic lasers, the quenching phenomenon resulting from the incoherent pumping of QEs becomes prominent at lower pumping rates than is suggested by the LRA-based calculations. The consequent loss of coherence of emitted light demands an even higher number of emitters than predicted by LRA methods to be coupled concertedly to a MNP in engineering plasmonic lasers. Such a collective strong coupling between the QEs and the dipole plasmon mode is achievable even at shorter surface separation distances where quenching to the nonradiative multipolar modes is expected [77]. We observed an interplay between the results for certain bath permittivity values due to the permittivities of the plasmonic metals being nonlinearly dependent on the frequency. Additionally, we studied the tunability of the presented model with respect to the optical properties of the QEs, which yielded a compelling, roughly even gap between local and nonlocal assessments regardless of the size of the MNP. Then, we discuss the flexibility of our model against nonidentical QEs and alternative arrangements of MNP-QE composite nanostructures. We emphasize that the nonlocal correction crucially alters the system dynamics within the parameter space widely adopted in describing plasmonic lasers. In essence, we stress the significance of hitherto disregarded quantum nonlocal effects in the context of plasmonic laser theory for a conceptually complete and accurate analysis, consistent with the experimental observations.

ACKNOWLEDGMENTS

D.L.G. would like to thank all the members of AχL at Monash University including H. Hapuarachchi for several illuminating discussions. The work of D.L.G. is supported by the Monash University Institute of Graduate Research and she thankfully acknowledges the computational resources granted by Monash University through the NCI NF Partner allocations for year 2018.

APPENDIX A: EoM FOR THE COMPLETE SYSTEM DENSITY MATRIX

We obtain the EoM for the complete system by applying the Hamiltonian in Eq. (1) in Eq. (12) as,

$$\frac{\partial}{\partial t}\rho_{\mathrm{n}}^{(\mu,\nu)} = \frac{-i}{\hbar}(\underbrace{[\hat{H}_{\mathrm{int}},\hat{\rho}]}_{A} + \underbrace{[\hat{H}_{\mathrm{qe}},\hat{\rho}]}_{B} + \underbrace{[\hat{H}_{\mathrm{pl}},\hat{\rho}]}_{C}) - \underbrace{\mathcal{D}\hat{\rho}}_{D}\,. \tag{A1}$$

The definition of the numbers $\{n_{\mathrm{cd}}\}$ with $c, d = g, e$ is as follows:

$$n_{\mathrm{cd}} = \sum_{r=1}^{N_{\mathrm{qe}}} \delta_{\mathrm{c},a_{\mathrm{r}}}\delta_{b_{\mathrm{r}},\mathrm{d}}, \tag{A2}$$

with a_{r} denoting the state of the rth QE. The set a_{r} and b_{r} represent the product state α and β, respectively. Thus, the numbers n_{cd} identify the number of QEs from product state $|\alpha\rangle$ in state φ_{c} and the number of QEs from product state $|\beta\rangle$ in state φ_{d} simultaneously. Then we proceed to expand the terms A-D as follows:

$$\begin{aligned}
A &= \sum_{n}^{N_{\mathrm{qe}}} \upsilon_{\mathrm{n}}\big(\langle\hat{\sigma}_{\mathrm{n}}^{\dagger}\hat{a}|\beta\nu\rangle\langle\alpha\mu|\rangle + \langle\hat{\sigma}_{\mathrm{n}}\hat{a}^{\dagger}|\beta\nu\rangle\langle\alpha\mu|\rangle - \langle|\beta\nu\rangle\langle\alpha\mu|\hat{\sigma}_{\mathrm{n}}^{\dagger}\hat{a}\rangle - \langle|\beta\nu\rangle\langle\alpha\mu|\hat{\sigma}_{\mathrm{n}}\hat{a}^{\dagger}\rangle\big)\\
&= \upsilon\sum_{n}\Big(\sqrt{\nu}\langle\hat{\sigma}_{\mathrm{n}}^{\dagger}|\beta\nu-1\rangle\langle\alpha\mu|\rangle + \sqrt{\nu+1}\langle\hat{\sigma}_{\mathrm{n}}|\beta\nu+1\rangle\langle\alpha\mu|\rangle - \sqrt{\mu}\langle\hat{\sigma}_{\mathrm{n}}^{\dagger}|\alpha\mu-1\rangle\langle\beta\nu|\rangle - \sqrt{\mu+1}\langle\hat{\sigma}_{\mathrm{n}}|\alpha\mu+1\rangle\langle\beta\nu|\rangle\Big);\\
\upsilon_{\mathrm{n}} &= \upsilon\forall n\\
&= \upsilon\sum_{n}\Big(\sqrt{\nu}\delta_{\mathrm{d},g_{\mathrm{n}}}\langle|e_{\mathrm{n}}\nu-1\rangle\langle g_{\mathrm{n}}\mu|\rangle + \sqrt{\nu+1}\delta_{\mathrm{d},e_{\mathrm{n}}}\langle|g_{\mathrm{n}}\nu+1\rangle\langle e_{\mathrm{n}}\mu|\rangle - \sqrt{\mu}\delta_{\mathrm{c},g_{\mathrm{n}}}\langle|e_{\mathrm{n}}\mu-1\rangle\langle g_{\mathrm{n}}\nu|\rangle - \sqrt{\mu+1}\delta_{\mathrm{c},e_{\mathrm{n}}}\langle|g_{\mathrm{n}}\mu+1\rangle\langle e_{\mathrm{n}}\nu|\rangle\Big)\\
&= \upsilon\sum_{\mathrm{a}=\{\mathrm{g,e}\}}\big(\sqrt{\nu}n_{\mathrm{ag}}\rho_{\mathrm{n_{ag-1},n_{ae+1}}}^{(\mu,\nu-1)} + \sqrt{\nu+1}n_{\mathrm{ae}}\rho_{\mathrm{n_{ae}-1,n_{ag}+1}}^{(\mu,\nu+1)} - \sqrt{\mu}n_{\mathrm{ga}}\rho_{\mathrm{n_{ga}-1,n_{ea}+1}}^{(\mu-1,\nu)} - \sqrt{\mu+1}n_{\mathrm{ea}}\rho_{\mathrm{n_{ea}-1,n_{ga}+1}}^{(\mu+1,\nu)}\big),
\end{aligned} \tag{A3}$$

$$\begin{aligned}
B &= \sum_{\mathrm{n}}^{N_{\mathrm{qe}}}\hbar\omega_{\mathrm{qe}}^{\mathrm{n}}\big(\langle\hat{\sigma}_{\mathrm{n}}^{\dagger}\hat{\sigma}_{\mathrm{n}}|\beta\nu\rangle\langle\alpha\mu|\rangle - \langle|\beta\nu\rangle\langle\alpha\mu|\hat{\sigma}_{\mathrm{n}}^{\dagger}\hat{\sigma}_{\mathrm{n}}\rangle\big)\\
&= \hbar\omega_{\mathrm{qe}}\sum_{\mathrm{n}}\big(\delta_{\mathrm{d},e_{\mathrm{n}}}\langle|\beta\nu\rangle\langle\alpha\mu|\rangle - \delta_{e_{\mathrm{n}},\mathrm{c}}\langle|\beta\nu\rangle\langle\alpha\mu|\rangle\big); \omega_{\mathrm{qe}}^{\mathrm{n}} = \omega_{\mathrm{qe}}\forall n\\
&= \hbar\omega_{\mathrm{qe}}\sum_{\mathrm{a}=\{\mathrm{g,e}\}}(n_{\mathrm{ae}} - n_{\mathrm{ea}})\rho_{\mathrm{n}}^{(\mu,\nu)}\\
&= \hbar\omega_{\mathrm{qe}}(n_{\mathrm{ge}} - n_{\mathrm{eg}})\rho_{\mathrm{n}}^{(\mu,\nu)},
\end{aligned} \tag{A4}$$

$$\begin{aligned} C &= \hbar\omega_{\mathrm{pl}}\big(\langle\hat{a}^\dagger\hat{a}|\beta\nu\rangle\langle\alpha\mu|\rangle - \langle|\beta\nu\rangle\langle\alpha\mu|\hat{a}^\dagger\hat{a}\rangle\big)\\ &= \hbar\omega_{\mathrm{pl}}\big(\sqrt{\nu}\langle\hat{a}^\dagger|\beta\nu-1\rangle\langle\alpha\mu|\rangle - \langle\sqrt{\mu}\hat{a}^\dagger|\alpha\mu-1\rangle\langle\beta\nu|\rangle\big)\\ &= \hbar\omega_{\mathrm{pl}}(\nu\langle|\beta\nu\rangle\langle\alpha\mu|\rangle - \langle\mu|\alpha\mu\rangle\langle\beta\nu|\rangle)\\ &= \hbar\omega_{\mathrm{pl}}(\nu-\mu)\rho_{\mathrm{n}}^{(\mu,\nu)}. \end{aligned} \tag{A5}$$

Term D includes the pumping of the QEs (D_{pumping}), the decay of the QEs (D_{qe}), and plasmons (D_{pl}):

$$D = D_{\mathrm{pumping}} + D_{\mathrm{qe}} + D_{\mathrm{pl}}, \tag{A6a}$$

$$\begin{aligned} D_{\mathrm{pumping}} &= \sum_{\mathrm{n}}^{N_{\mathrm{qe}}} \frac{p_{\mathrm{n}}}{2}\big(\langle\hat{\sigma}_{\mathrm{n}}\hat{\sigma}_{\mathrm{n}}^\dagger|\beta\nu\rangle\langle\alpha\mu|\rangle + \langle|\beta\nu\rangle\langle\alpha\mu|\hat{\sigma}_{\mathrm{n}}\hat{\sigma}_{\mathrm{n}}^\dagger\rangle - 2\langle\hat{\sigma}_{\mathrm{n}}^\dagger|\beta\nu\rangle\langle\alpha\mu|\hat{\sigma}_{\mathrm{n}}\rangle\big)\\ &= \frac{p}{2}\sum_{\mathrm{n}}\big((\delta_{g_{\mathrm{n}},\mathrm{d}} + \delta_{\mathrm{c},g_{\mathrm{n}}})\langle|\alpha\mu\rangle\langle\beta\nu|\rangle - 2\delta_{\mathrm{c},e_{\mathrm{n}}}\delta_{e_{\mathrm{n}},\mathrm{d}}\langle|g_{\mathrm{n}}\mu\rangle\langle g_{\mathrm{n}}\nu|\rangle\big); p_{\mathrm{n}} = p\forall n\\ &= \frac{p}{2}\big((2n_{\mathrm{gg}} + n_{\mathrm{eg}} + n_{\mathrm{ge}})\rho_{\mathrm{n}}^{(\mu,\nu)} - 2n_{\mathrm{ee}}\rho^{(\mu,\nu)}_{\mathrm{n_{ee}-1,n_{gg}+1}}\big), \end{aligned} \tag{A6b}$$

$$\begin{aligned} D_{\mathrm{qe}} &= \sum_{\mathrm{n}}^{N_{\mathrm{qe}}} \frac{\gamma_{\mathrm{qe}}^{\mathrm{n}}}{2}\big(\langle\hat{\sigma}_{\mathrm{n}}^\dagger\hat{\sigma}_{\mathrm{n}}|\beta\nu\rangle\langle\alpha\mu|\rangle + \langle|\beta\nu\rangle\langle\alpha\mu|\hat{\sigma}_{\mathrm{n}}^\dagger\hat{\sigma}_{\mathrm{n}}\rangle - 2\langle\hat{\sigma}_{\mathrm{n}}|\beta\nu\rangle\langle\alpha\mu|\hat{\sigma}_{\mathrm{n}}^\dagger\rangle\big)\\ &= \frac{\gamma_{\mathrm{qe}}}{2}\sum_{\mathrm{n}}\big((\delta_{e_{\mathrm{n}},\mathrm{d}} + \delta_{\mathrm{c},e_{\mathrm{n}}})\langle|\alpha\mu\rangle\langle\beta\nu|\rangle - 2\delta_{\mathrm{d},g_{\mathrm{n}}}\delta_{\mathrm{c},g_{\mathrm{n}}}\langle|e_{\mathrm{n}}\mu\rangle\langle e_{\mathrm{n}}\nu|\rangle\big); \gamma_{\mathrm{qe}}^{\mathrm{n}} = \gamma_{\mathrm{qe}}\forall n\\ &= \frac{\gamma_{qe}}{2}\big((2n_{\mathrm{ee}} + n_{\mathrm{eg}} + n_{\mathrm{ge}})\rho_{\mathrm{n}}^{(\mu,\nu)} - 2n_{\mathrm{gg}}\rho^{(\mu,\nu)}_{\mathrm{n_{gg}-1,n_{ee}+1}}\big), \end{aligned} \tag{A6c}$$

$$\begin{aligned} D_{\mathrm{pl}} &= \frac{\gamma_{\mathrm{pl}}}{2}\big(\langle\hat{a}^\dagger\hat{a}|\beta\nu\rangle\langle\alpha\mu|\rangle + \langle|\beta\nu\rangle\langle\alpha\mu|\hat{a}^\dagger\hat{a}\rangle - 2\langle\hat{a}|\beta\nu\rangle\langle\alpha\mu|\hat{a}^\dagger\rangle\big)\\ &= \frac{\gamma_{pl}}{2}\big((\mu+\nu)\rho_{\mathrm{n}}^{(\mu,\nu)} - 2\sqrt{(\mu+1)(\nu+1)}\rho_{\mathrm{n}}^{(\mu+1,\nu+1)}\big). \end{aligned} \tag{A6d}$$

We than apply Eqs. (A3)–(A6d) back into Eq. (A1) and arrive at Eq. (16).

APPENDIX B: EoM FOR THE PLASMON DENSITY MATRIX

By applying Eq. (12) on the plasmon reduced system Hamiltonian, we arrive at the following EoM:

$$\frac{\partial}{\partial t}\rho_{(\mu,\nu)} = \frac{-i}{\hbar}(\underbrace{[\hat{H}_{\mathrm{int}},\hat{\rho}]}_{E} + \underbrace{[\hat{H}_{\mathrm{pl}},\hat{\rho}]}_{F}) - \underbrace{\mathcal{D}\hat{\rho}}_{G}. \tag{B1}$$

We then expand term E as,

$$\begin{aligned} E &= \sum_{n}^{N_{\mathrm{qe}}}\upsilon_{\mathrm{n}}\big(\langle\hat{\sigma}_{\mathrm{n}}^\dagger\hat{a}|\nu\rangle\langle\mu|\rangle + \langle\hat{\sigma}_{\mathrm{n}}\hat{a}^\dagger|\nu\rangle\langle\mu|\rangle - \langle|\nu\rangle\langle\mu|\hat{\sigma}_{\mathrm{n}}^\dagger\hat{a}\rangle - \langle|\nu\rangle\langle\mu|\hat{\sigma}_{\mathrm{n}}\hat{a}^\dagger\rangle\big)\\ &= \upsilon\sum_{n}\Big(\sqrt{\nu}\langle\hat{\sigma}_{\mathrm{n}}^\dagger|\nu-1\rangle\langle\mu|\rangle + \sqrt{\nu+1}\langle\hat{\sigma}_{\mathrm{n}}|\nu+1\rangle\langle\mu|\rangle - \sqrt{\mu}\langle\hat{\sigma}_{\mathrm{n}}^\dagger|\mu-1\rangle\langle\nu|\rangle - \sqrt{\mu+1}\langle\hat{\sigma}_{\mathrm{n}}|\mu+1\rangle\langle\nu|\rangle\Big); \upsilon_{\mathrm{n}} = \upsilon\forall n\\ &= \upsilon\sum_{n}\Big(\sqrt{\nu}\langle|e_{\mathrm{n}}\nu-1\rangle\langle g_{\mathrm{n}}\mu|\rangle + \sqrt{\nu+1}\langle|g_{\mathrm{n}}\nu+1\rangle\langle e_{\mathrm{n}}\mu|\rangle - \sqrt{\mu}\langle|e_{\mathrm{n}}\mu-1\rangle\langle g_{\mathrm{n}}\nu|\rangle - \sqrt{\mu+1}\langle|g_{\mathrm{n}}\mu+1\rangle\langle e_{\mathrm{n}}\nu|\rangle\Big)\\ &= \upsilon\sum_{n}(\sqrt{\nu}\rho_{\mathrm{g}\mu,\mathrm{e}\nu-1} + \sqrt{\nu+1}\rho_{\mathrm{e}\mu,\mathrm{g}\nu+1} - \sqrt{\mu}\rho_{\mathrm{e}\mu-1,\mathrm{g}\nu} - \sqrt{\mu+1}\rho_{\mathrm{g}\mu+1,\mathrm{e}\nu}). \end{aligned} \tag{B2}$$

Term F can be expanded in terms of number states as,

$$\begin{aligned} F &= \hbar\omega_{\mathrm{pl}}\big(\langle\hat{a}^\dagger\hat{a}|\nu\rangle\langle\mu|\rangle - \langle|\nu\rangle\langle\mu|\hat{a}^\dagger\hat{a}\rangle\big)\\ &= \hbar\omega_{\mathrm{pl}}\big(\sqrt{\nu}\langle\hat{a}^\dagger|\nu-1\rangle\langle\mu|\rangle - \langle\sqrt{\mu}\hat{a}^\dagger|\mu-1\rangle\langle\nu|\rangle\big)\\ &= \hbar\omega_{\mathrm{pl}}(\nu\langle|\nu\rangle\langle\mu|\rangle - \langle\mu|\mu\rangle\langle\nu|\rangle)\\ &= \hbar\omega_{\mathrm{pl}}(\nu-\mu)\rho_{\mu,\nu}. \end{aligned} \tag{B3}$$

The plasmon decay can be accounted with Eq. (13) as,

$$G = \frac{\gamma_{\mathrm{pl}}}{2}\left([\hat{a}^\dagger\hat{a}, |\nu\rangle\langle\mu|]_+ - 2\hat{a}\langle|\nu\rangle\langle\mu|\hat{a}^\dagger\rangle\right)$$
$$= \frac{\gamma_{\mathrm{pl}}}{2}\Big((\mu+\nu)\rho_{\mu,\nu} - 2\sqrt{(\mu+1)(\nu+1)}\rho_{\mu+1,\nu+1}\Big). \tag{B4}$$

We apply Eqs. (B2)–(B4) in Eq. (B1) to arrive at Eq. (18). Similarly, we derive and simultaneously solve EoM for $\rho_{\mathrm{g}\mu,\mathrm{e}\nu-1}$, $\rho_{\mathrm{e}\mu,\mathrm{g}\nu+1}$, $\rho_{\mathrm{e}\mu-1,\mathrm{g}\nu}$ and $\rho_{\mathrm{g}\mu+1,\mathrm{e}\nu}$. We then arrive at Eq. (19) by solving Eq. (18) for the population terms ($\rho_{\mu,\mu}$) at steady state.

[1] H. A. Atwater, Sci. Am. **296**, 56 (2007).
[2] S. Lal, S. Link, and N. J. Halas, Nat. Photonics **1**, 641 (2007).
[3] M. I. Stockman, Opt. Express **19**, 22029 (2011).
[4] U. Resch-Genger, M. Grabolle, S. Cavaliere-Jaricot, R. Nitschke, and T. Nann, Nat. Methods **5**, 763 (2008).
[5] U. Woggon, *Optical Properties of Semiconductor Quantum Dots* (Springer, Berlin, 1997).
[6] Z.-J. Zhu, Y.-C. Yeh, R. Tang, B. Yan, J. Tamayo, R. W. Vachet, and V. M. Rotello, Nat. Chem. **3**, 963 (2011).
[7] H. Xu, X.-H. Wang, M. P. Persson, H. Q. Xu, M. Käll, and P. Johansson, Phys. Rev. Lett. **93**, 243002 (2004).
[8] A. O. Govorov, G. W. Bryant, W. Zhang, T. Skeini, J. Lee, N. A. Kotov, J. M. Slocik, and R. R. Naik, Nano Lett. **6**, 984 (2006).
[9] O. Chen, L. Riedemann, F. Etoc, H. Herrmann, M. Coppey, M. Barch, C. T. Farrar, J. Zhao, O. T. Bruns, H. Wei, P. Guo, J. Cui, R. Jensen, Y. Chen, D. K. Harris, J. M. Cordero, Z. Wang, A. Jasanoff, D. Fukumura, R. Reimer, M. Dahan, R. K. Jain, and M. G. Bawendi, Nat. Commun. **5**, 5093 (2014).
[10] S. Senapati, A. K. Mahanta, S. Kumar, and P. Maiti, Signal Transduction Targeted Ther. **3**, 7 (2018).
[11] S. M. Sadeghi, Phys. Rev. B **82**, 035413 (2010).
[12] P. Berini and I. De Leon, Nat. Photonics **6**, 16 (2011).
[13] H. Y. Xie, H. Y. Chung, P. T. Leung, and D. P. Tsai, Phys. Rev. B **80**, 155448 (2009).
[14] M. Lunz, V. A. Gerard, Y. K. Gun'ko, V. Lesnyak, N. Gaponik, A. S. Susha, A. L. Rogach, and A. L. Bradley, Nano Lett. **11**, 3341 (2011).
[15] J. Ren, T. Wu, B. Yang, and X. Zhang, Phys. Rev. B **94**, 125416 (2016).
[16] S. Mallawaarachchi, M. Premaratne, S. D. Gunapala, and P. K. Maini, Phys. Rev. B **95**, 155443 (2017).
[17] K. D. B. Higgins, S. C. Benjamin, T. M. Stace, G. J. Milburn, B. W. Lovett, and E. M. Gauger, Nat. Commun. **5**, 4705 (2014).
[18] S. Mallawaarachchi, S. D. Gunapala, M. I. Stockman, and M. Premaratne, Phys. Rev. B **97**, 125406 (2018).
[19] S. Mallawaarachchi, M. Premaratne, and P. K. Maini, IEEE J. Sel. Top. Quantum Electron. **25**, 7101508 (2019).
[20] D. J. Bergman and M. I. Stockman, Phys. Rev. Lett. **90**, 027402 (2003).
[21] M. I. Stockman, J. Opt. **12**, 024004 (2010).
[22] M. I. Stockman, Spaser, plasmonic amplification, and loss compensation, in *Active Plasmonics and Tuneable Plasmonic Metamaterials* (John Wiley & Sons, Inc., Hoboken, NJ, 2013).
[23] Y. Yin, T. Qiu, J. Li, and P. K. Chu, Nano Energy **1**, 25 (2012).
[24] C. Rupasinghe, I. D. Rukhlenko, and M. Premaratne, ACS Nano **8**, 2431 (2014).
[25] C. Jayasekara, M. Premaratne, S. D. Gunapala, and M. I. Stockman, J. Appl. Phys. **119**, 133101 (2016).
[26] L. Kumarapperuma, M. Premaratne, P. K. Jha, M. I. Stockman, and G. P. Agrawal, Appl. Phys. Lett. **112**, 201108 (2018).
[27] M. Noginov, G. Zhu, A. Belgrave, R. Bakker, V. Shalaev, E. Narimanov, S. Stout, E. Herz, T. Suteewong, and U. Wiesner, Nature **460**, 1110 (2009).
[28] M. Premaratne and M. I. Stockman, Adv. Opt. Photonics **9**, 79 (2017).
[29] B. Liu, W. Zhu, S. D. Gunapala, M. I. Stockman, and M. Premaratne, ACS Nano **11**, 12573 (2017).
[30] M. Richter, M. Gegg, T. S. Theuerholz, and A. Knorr, Phys. Rev. B **91**, 035306 (2015).
[31] T. Warnakula, M. I. Stockman, and M. Premaratne, J. Opt. Soc. Am. B **35**, 1397 (2018).
[32] Y. Zhang and V. May, J. Chem. Phys. **142**, 224702 (2015).
[33] S. A. Maier, *Plasmonics: Fundamentals and Applications* (Springer Science & Business Media, New York, 2007), Chap. 5.
[34] A. Ridolfo, O. Di Stefano, N. Fina, R. Saija, and S. Savasta, Phys. Rev. Lett. **105**, 263601 (2010).
[35] T. S. Theuerholz, A. Carmele, M. Richter, and A. Knorr, Phys. Rev. B **87**, 245313 (2013).
[36] S. Raza, S. I. Bozhevolnyi, M. Wubs, and N. A. Mortensen, J. Phys.: Condens. Matter **27**, 183204 (2015).
[37] W. R. Holland and D. G. Hall, Phys. Rev. Lett. **52**, 1041 (1984).
[38] J. Tiggesbäumker, L. Köller, K.-H. Meiwes-Broer, and A. Liebsch, Phys. Rev. A **48**, R1749(R) (1993).
[39] C. David and F. J. Garcia de Abajo, J. Phys. Chem. C **115**, 19470 (2011).
[40] 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).
[41] A. Pitelet, E. Mallet, R. Ajib, C. Lemaître, E. Centeno, and A. Moreau, Phys. Rev. B **98**, 125418 (2018).
[42] C. Ciracì, R. T. Hill, J. J. Mock, Y. Urzhumov, A. I. Fernández-Domínguez, S. A. Maier, J. B. Pendry, A. Chilkoti, and D. R. Smith, Science **337**, 1072 (2012).
[43] A. Wiener, A. I. Fernández-Domínguez, J. B. Pendry, A. P. Horsfield, and S. A. Maier, Opt. Express **21**, 27509 (2013).
[44] C. Tserkezis, A. T. M. Yeşilyurt, J.-S. Huang, and N. A. Mortensen, ACS Photonics **5**, 5017 (2018).
[45] M. Wubs, Opt. Express **23**, 31296 (2015).
[46] W. Ekardt and Z. Penzar, Phys. Rev. B **34**, 8444 (1986).
[47] P. Zhang, J. Feist, A. Rubio, P. García-González, and F. J. García-Vidal, Phys. Rev. B **90**, 161407 (2014).
[48] E. B. Guidez and C. M. Aikens, Nanoscale **6**, 11512 (2014).
[49] J. Zuloaga, E. Prodan, and P. Nordlander, Nano Lett. **9**, 887 (2009).
[50] E. Prodan and P. Nordlander, Nano Lett. **3**, 543 (2003).

[51] M. Pelton, J. Aizpurua, and G. Bryant, Laser Photonics Rev. **2**, 136 (2008).

[52] T. V. Teperik, P. Nordlander, J. Aizpurua, and A. G. Borisov, Opt. Express **21**, 27306 (2013).

[53] R. Esteban, A. G. Borisov, P. Nordlander, and J. Aizpurua, Nat. Commun. **3**, 825 (2012).

[54] B. B. Dasgupta and R. Fuchs, Phys. Rev. B **24**, 554 (1981).

[55] G. S. Agarwal and S. V. ONeil, Phys. Rev. B **28**, 487 (1983).

[56] N. A. Mortensen, S. Raza, M. Wubs, T. Søndergaard, and S. I. Bozhevolnyi, Nat. Commun. **5**, 3809 (2014).

[57] H. Hapuarachchi, M. Premaratne, Q. Bao, W. Cheng, S. D. Gunapala, and G. P. Agrawal, Phys. Rev. B **95**, 245419 (2017).

[58] H. Hapuarachchi, S. Mallawaarachchi, H. T. Hattori, W. Zhu, and M. Premaratne, J. Phys.: Condens. Matter **30**, 054006 (2018).

[59] C. Tserkezis, N. Stefanou, M. Wubs, and N. A. Mortensen, Nanoscale **8**, 17532 (2016).

[60] P. T. Leung, Phys. Rev. B **42**, 7622 (1990).

[61] S. Raza, M. Wubs, S. I. Bozhevolnyi, and N. A. Mortensen, Opt. Lett. **40**, 839 (2015).

[62] R. Ruppin, Phys. Rev. B **45**, 11209 (1992).

[63] S. Raza, W. Yan, N. Stenger, M. Wubs, and N. A. Mortensen, Opt. Express **21**, 27344 (2013).

[64] S. G. Kosionis, A. F. Terzis, V. Yannopapas, and E. Paspalakis, J. Phys. Chem. C **116**, 23663 (2012).

[65] H. Hapuarachchi, S. D. Gunapala, Q. Bao, M. I. Stockman, and M. Premaratne, Phys. Rev. B **98**, 115430 (2018).

[66] R. Fuchs and F. Claro, Phys. Rev. B **35**, 3722 (1987).

[67] A. D. Boozer, Phys. Rev. A **78**, 053814 (2008).

[68] B. Jones, S. Ghose, J. P. Clemens, P. R. Rice, and L. M. Pedrotti, Phys. Rev. A **60**, 3267 (1999).

[69] V. May, J. Chem. Phys. **140**, 054103 (2014).

[70] R. H. Dicke, Phys. Rev. **93**, 99 (1954).

[71] K. Kolwas and A. Derkachova, Opto-Electron. Rev. **18**, 429 (2010).

[72] M. Sargent, M. O. Scully, and W. E. Lamb, *Laser Physics* (Addison-Wesley, London, 1974).

[73] M. Kalthoff, F. Keim, H. Krull, and G. S. Uhrig, Eur. Phys. J. B **90** (2017).

[74] P. Meystre and M. Sargent, *Elements of Quantum Optics* (Springer-Verlag, Berlin, 2007).

[75] W. E. Hayenga, H. Garcia-Gracia, H. Hodaei, C. Reimer, R. Morandotti, P. LiKamWa, and M. Khajavikhan, Optica **3**, 1187 (2016).

[76] P. B. Johnson and R. W. Christy, Phys. Rev. B **6**, 4370 (1972).

[77] A. Delga, J. Feist, J. Bravo-Abad, and F. J. Garcia-Vidal, Phys. Rev. Lett. **112**, 253601 (2014).