Abstract: Quantum plasmonics is an exciting sub-branch of nanoplasmonics where the laws of quantum theory are used to describe light-matter interactions on the nanoscale. Plasmonic materials allow extreme sub-diffraction confinement of (quantum or classical) light to regions so small that the quantization of both light and matter may be necessary for an accurate description. State of the art experiments now allow us to probe these regimes and push existing theories to the limits which opens up the possibilities of exploring the nature of many-body collective oscillations as well as developing new plasmonic devices, that use the particle quality of light and the wave quality of matter, and have a wealth of potential applications in sensing, lasing and quantum computing. This merging of fundamental condensed matter theory with application-rich electromagnetism (and a splash of quantum optics thrown in) gives rise to a fascinating area of modern physics that is still very much in its infancy. In this review we discuss and compare the keys models and experiments used to explore how the quantum nature of electrons impacts plasmonics in the context of quantum size corrections of localised plasmons and quantum tunnelling between nanoparticle dimers. We also look at some of the remarkable experiments that are revealing the quantum nature of surface plasmon polaritons.

Index Terms: Plasmonics, Quantum Mechanics, Quantum Optics, Nonlocality.

1. Introduction

Plasmonics is the study of how electromagnetic excitations couple to free carrier systems in the bulk and at boundaries between different materials, they can be propagating or localized [9], [93], [113], [116], [163]. In metals an electromagnetic (EM) field can couple to the conduction electrons, which to a good approximation behave as free carriers, and drive them back and forth. In a finite system an induced surface charge will form which causes a restoring force on the electrons and an associated resonance which will depend on the geometry of the system. In the non-retarded limit (where the speed of light can be taken to be infinite) this resonance is known as a plasmon. In the retarded limit the plasmon may couple to the light and a hybrid mode called a surface plasmon polariton is formed. At frequencies close to the system’s plasma frequency this collective oscillation of electrons will dominate the optical response and, via the transfer of energy to electronic degrees of freedom, allows confinement of light to sub-wavelength limits and a concomitant field enhancement. This provides an opportunity to manipulate light on the nanoscale with a huge amount of potential applications including sensing [60], metamaterials/metasurfaces [93], cancer therapy [81], lasing [12], plasmonic waveguiding [18] and integrated nanophotonic circuits [46]. The usefulness of plasmonics stems from its position as a middle ground between diffraction-limited photonics and bandwidth-limited electronics meaning plasmonics can provide the solution to the incompatibility of electronic and photonic circuits [109]. Increasing levels of sophistication of experimental techniques in fabrication and characterization methods, alongside the growing power and ease of use of electromagnetic modelling software, led to a boom in the interest in plasmonics in the early millennium resulting in plasmonics today being a mature field with active researchers from many areas of science. Of course there are still exciting opportunities and challenges not least the overcoming of loss which at the moment seems to limit applications to optical frequencies for individual plasmonic components and metamaterials [59].

In recent years there has been growing interest in exploring when quantum effects are required when
modelling plasmonic systems. As nanoscale fabrication becomes more and more sophisticated, the level of electron confinement is ever increasing. At a length scale on the order of the Fermi wavelength of the valence electrons one would expect the quantum nature of the system to become important. At this point the electrons should be treated as wavelike with a finite extent in space. Simultaneously there have been a number of experiments that explore plasmonics with low intensity quantum sources of light, in this regime one must start to consider the single quantum particle nature of the photon as well as the plasmon polariton. Here we must consider the single quantum particle nature of plasmon polaritons. Thus we will define quantum plasmonics as the study of the quantum nature of the constituent electrons and photon parts of the plasmon. By studying quantum plasmonics we hope to improve our understanding of the plasmonic response in new experimental regimes as well learn more about the fundamental properties of plasmons as quantum objects.

At this point we should address why we think there is a need for another review on quantum plasmonics. Despite being a relatively new field there already exists a couple of excellent recent reviews on the subject [65], [138], but these focus more on the quantum light nature of surface plasmon polaritons. We will also discuss this fascinating area of research, but our emphasis is directed more towards the matter quantization aspects. This topic has both a rich theoretical history and state of the art experimentation is now in a position to test these theories rigorously. The most recent review, that we are aware of, on this topic is the very thorough report by Pitarke et al [113] in 2007. We also take this opportunity to highlight some recent excellent reviews on the topics of nonlocality [117] and on electronic structure methods in plasmonics [105].

The structure of this review is as follows: we shall begin with a brief review of plasmonics and in particular focus on the assumptions made in the classical approximation. We will then discuss when these approximations break down and when one needs to include quantum/nonlocal effects for the matter. In this review we will focus on quantum size effects and dimer systems. We will then look at recent work where surface plasmon polaritons are excited by single photon sources of lights. Lastly we will briefly discuss what we think are the future directions of quantum plasmonics. Unfortunately, we will not have the space to discuss many interesting topics that may deserve to fall under the title of ‘quantum plasmonics’, some notable omissions include quantum emitters near plasmonic structures [106], graphene plasmonics [74], semiconductor plasmonics [90], hot electrons [24] and active quantum plasmonics [97].

2. The Plasmonics Family

Here we present a brief discussion on the family of plasmon oscillations.

2.1. The Volume Plasmon

A volume (or bulk) plasma oscillation is a longitudinal excitation which consists of the coherent motion of an electron gas (all electrons move with the same frequency and wavenumber), the energy quanta is called the volume plasmon. The frequency of the oscillations is the plasma frequency

$$\omega_p = \sqrt{\frac{n_0 e^2}{m \epsilon_0}},$$

which is an intrinsic property of the material, depending only on the electron density \(n_0\), charge \(e\) and mass \(m\): in other words, the bulk plasma frequency is flat in \((\omega, q)\) space (see fig 1a, red dashed line). Typical values for metals are on the order of 10 eV, this large energy (relative to the electron-hole excitation energy) is crucial for the stability of the plasmon and is simply a result of the large number of electrons that make up the plasmon [111]. The first pioneering work on collective modes in a degenerate electron gas was done by Bohm and Pines in the 1950s who, within the random phase approximation (RPA), found that the long range part of the Coulomb interaction manifests as a organized collective oscillation of the whole electron gas [16], [112]. The long range of the Coulomb interaction means each electron experiences a potential from a large number of other electrons, this will result in a small perturbation of its position and momentum. Such ideas were crucial for many body physics as they allowed the division of the many body problem into two (approximately) separate simpler problems of collective excitations and individual electrons interacting via a screened Coulomb potential. This means, via a suitable canonical transformation of the many body Hamiltonian, that the troubles caused by the long range nature of the Coulomb potential can be dealt with in an elegant fashion.

Experimental evidence came from electron energy-loss spectroscopy (EELS) of thin metal foils [48], [98], [115], that shows not only the existence of bulk plasmons but also the quantization in units of \(\hbar \omega\).
2.2. The Surface Plasmon Polariton

Soon after the introduction of the bulk plasmon it was realized by Ritchie [124] that there was a lower energy mode supported by the metal surface (green dashed line in fig 1a), this is the surface plasmon (SP). The surface breaks translational invariance and allows the wavevector of the plasmon to be complex, this corresponds to a mode bound in the direction normal to the surface (no power is radiated perpendicular to the surface) which can acquire a transverse character (see fig 1b) [93]. This means the SP can couple to light (with suitable structuring of the surface to overcome the momentum mismatch [93], [129]) and the resulting non-radiative and propagating hybrid mode is called a surface plasmon polariton (SPP). By solving Maxwell’s equations for two semi-infinite half-spaces, with the usual boundary conditions for a sharp interface, one finds the relation [93], [113]

\[ q(\omega) = \frac{\omega}{c} \sqrt{\frac{\epsilon_1 \epsilon_2}{\epsilon_1 + \epsilon_2}}, \]  

where \( q \) is the in-plane momentum and \( \epsilon_1 \) and \( \epsilon_2 \) denote the dielectric function of the two half-spaces (blue line in fig 1a). If we take an interface where the dielectric function changes sign (say \( \epsilon_2 < 0 \) and \( \epsilon_1 > 0 \)) then it is clear that for a propagating mode we require \(|\epsilon_2| > \epsilon_1\), such a condition is satisfied for a metal-dielectric interface at optical frequencies. Allowing the dielectric functions to be complex does not change the above analysis except in that the SPP will have a finite propagation length; at optical frequencies this loss comes from the metal. At this point one must choose a functional form for the dielectric functions, if for the metal we use the frequency dependent Drude model and for the dielectric a real number then we can calculate the SPP dispersion. Looking at figure 1a one can clearly see two asymptotic limits for the lower SPP curve. In the non-retarded limit, \( q \to \infty \), then the SPP behaves as a plasmon and oscillates at the Ritchie frequency given by \( \epsilon_1(\omega) + \epsilon_2 = 0 \). For a metal-vacuum interface this gives the well known result of \( \omega_p/\sqrt{2} \). In the retarded limit the SPP becomes more light-like and propagates parallel to the surface.

2.3. Localized Surface Plasmons

A localized surface plasmon (LSP) is simply a confined SPP in a small nanoparticle (NP) (see fig 1c). The mode does not propagate and has a flat curve dispersion that allows for direct excitation by light [93]. The theory of the electromagnetic response of a generic spherical particle was developed by Mie over a hundred years ago [101] and can be easily implemented numerically [17]. For nanoparticles smaller than about 20 nm one can assume the electric field phase is constant over the particle and the problem effectively becomes an electrostatic one. This is the quasi-static approximation which simplifies the calculations considerably and allows analytical results [93], [113]. Within this limit, the dipole mode dominates and there is no size dependence of the plasmon resonance. With the approximation of a sharp interface between metal and dielectric half-spaces then the dipole surface plasmon resonances for a sphere is found to be given by the Frölich condition

\[ \Re[\epsilon_2] = -2\epsilon_1, \]  

which for a Drude metal-air interface gives the well known result \( \omega_p/\sqrt{3} \).
2.4. The Classical Derivations: Strength and Shortcomings

We will define classical derivations as referring to the use of basic assumptions like sharp boundary conditions (which correspond to a step-function electron density profile) and a local dispersive bulk dielectric function. Such models have given very accurate results and continue to be used for the huge majority of plasmonics research, they are also simple to use/understand and are easily implemented in numerical simulations. Thus one is well justified in asking: why is there any need to give up the classical model? Most of the time there is no need to, it is only when we push experiments to extreme regions of particle size, or light intensity, that we must consider quantum mechanical corrections. In these regimes the classical predictions may not be just quantitatively, but also qualitatively, wrong!

3. Quantising Matter in Plasmonics

In this section we will focus on two areas where the quantum nature of electrons is required to describe the response of nanoplasmonic systems. The first is the issue of quantum size effects of NPs, for small particles ($R \lesssim 2 \text{ nm}$ which corresponds to a few hundred valence electrons) the electrons will ‘feel’ their confinement and will have a discrete spectrum [55], this can be observed experimentally [71], [165]. Typically energy level discreteness is ignored in nanoplasmonics as the high electron density in metals means very small energy spacing at room temperature [93]; the optical response will only be modified by the confinement of the electron gas for the very smallest NPs. Linked to this is a second size effect which originates from the growing importance of the surface (which in a quantum picture is just the boundary conditions for the electron wavefunctions). The ratio of the surface to the volume scales with $\sim 1/R$ meaning that, as NPs get smaller, the surface region increasingly dominates the optical response. This means the surface region must now be modelled very accurately, in particular for the behaviour of the electron density at the surface.

The theory of this topic has a long history starting with Kubo [79] who was the first to explore in detail how thermodynamic properties will be modified by the discrete energy level in small metal NPs (MNP). This inspired many similar ‘particle in a box’ type models [29], [51], [63], [89], [110], [123], [143], [151]. In particular Rice et al [123] were the first to perform a calculation of the polarizability of small particles that took electron screening into account. An electron in a MNP does not directly experience the applied field, instead it experiences a screened field which is the sum of the external and induced field.

Interest in the quantum size effects of MNPs has grown again in recent years, but with a focus on the excitation of surface plasmons. This is thanks to new experimental techniques that allow the nano-fabrication and detection of individual, and touching, particles and has resulted in some remarkable experiments in the quantum plasmonics size regime [27], [130], [132] (see figure 2). As the particle size is decreased some typical nonlocal effects are seen such as a shift of the resonance peak and broadening of the peak (from increased damping). Predicting the shift of the LSP resonance decreasing particle size is difficult as there can be competing mechanisms between energy level quantisation (which always leads to blue shifts and is most important for sizes below 1 nm) and dynamical surface screening (which can lead to a red or blue shift) [102]. For alkali metal clusters the situation is fairly clear-cut: electron spill out of the valence electrons leads to an effective larger radius (and a reduced density in the interior) than that predicted classically and hence corresponds to a red shift which has been found in experiments [14], [33], [133], [134], [149] (it is worth noting that the redshift with decreasing particle size has been observed to be non-monotonic [121]). Noble metals are not so simple due to the influence of d-shell electrons. The d-electrons form a polarizable background that screens the valence electrons and lowers the energy of collective excitations, they are also strongly localized meaning screening at the surface is reduced. As a noble-metal NPs radius is decreased there is a competition between two opposite trends: red shift from electron spill out and the blue shift from reduction in d-electron screening (as the surface to volume ratio increases), for instance in silver the reduction in screening wins and there is an observed blue shift [25], [56], [118], [120], [132], [142]. The theory behind the role of dynamical surface screening was first derived by Apell and Ljungberg [4]–[6]. In the past it was only possible to create ensembles of MNPs either embedded in a matrix of some sort (e.g. glass) [76] or in a gas phase via high pressure vapor flow [125], [154]; clusters of a specific mass can then be analysed by using a mass spectrometer. Unfortunately, this creates a size distribution that can ‘wash out’ any quantum effects, plus it is hard to control environmental effects which can be difficult to incorporate in theoretical models. For instance, it is known that the surrounding dielectric environment can strongly affect the LSP in small MNPs [83], similar to the classical case. Even if MNPs are isolated there are still challenges with optical detection in the far field as the absorption cross section scales with $R^{-3}$ and the scattering with $R^{-6}$ [93]. Cluster science is a well-established area of research [19], [32], [53] and many of the results of the past 30 years
can aid developments in the plasmonics, in fact the quantum plasmonics of tiny MNPs can be viewed as the merging of cluster physics with plasmonics.

The second area we will focus upon is tunnelling between nearly touching nanostructures. Tunnelling is a quintessential quantum process where a quantum object travels through a barrier that would be classically forbidden. It has no classical counterpart and therefore cannot be described at all by classical models, it is the perfect example of where plasmonics must go quantum! And it is not only an academic question, some of the highest field enhancements in plasmonics (so called hot spots) occur in small gaps where quantum effects become important. The classical theory predicts infinite field enhancement whereas a quantum calculation properly models the surface charge density as being spread over a finite volume with reduced field enhancements [27].

Again, the recent interest in this problem partly stems from increased experimental capabilities. Some recent state-of-the-art experiments have really pushed the limits of sub-nanometre control [27], [130], [131], [139]. It is now possible to control the spacing between nanostructures all the way down to the point of touching and in detail explore how the plasmonic response depends on spacing. The three experiments that drove research in quantum plasmonics are presented in fig 2 [27], [130], [132]. Savage et al observed tunnelling of electrons between two gold metal NPs that were almost touching (∼0.3 nm, see fig 2a) [130]. They measured the electrical and optical properties of two gold nanostructures, varying the separation with sub-nanometre control simultaneously: in this way revealing the quantum limit for plasmonic field enhancements and confining. Similar findings were obtained by Cirací et al where, in this experiment, the gap between a gold sphere and a gold film was controlled via chemical deposited sub-nanometre molecular layers (see fig 2b) [27]. Scholl et al analysed the departure from classical theory probing two closely spaced 10 nm MNPs with EELS by controlling the separation via a transmission electron microscope from 3 nm to a overlap of −8 nm (where the NPs have coalesced) (see fig 2c) [132].

### 3.1. Constructing A Quantum Electron Theory of Plasmonics

In this section will we will assume there are enough photons to treat the light fields classically, thus the light is included via Maxwell’s equations. The key idea of this section is that the optical response of a nanostructure is determined via the electron density: an incoming field will perturb the electron density (or alternatively we can view this as inducing virtual and real transitions amongst the energy levels) which in turn causes an induced field which acts back on the system. Thus our challenge is to self-consistently solve Maxwell’s equations with the many body Schrödinger equation: A daunting prospect! Clearly it is infeasible to solve this problem exactly, the rest of this section is spent exploring various approximations.

---

**Fig. 2:** (a) Optical and electrical schema for measurements of plasmonic modes excited in two MNPs (coated tips) varying the gap distance $d$. The plasmonic resonances are labeled with letters A-C (adapted from ref. [130]). (b) Example of a configuration where nonlocal effects are important, i.e. gold NPs on a gold film separated by a sub-nanometric layer (adapted from ref. [27]). (c) Scanning transmission electron microscopy image of a 20-nm-diameter silver particle and the associated EELS data measured in different positions (adapted from ref. [132]).
3.2. Nonlocality

Nonlocality (or spatial dispersion) can be a somewhat complicated concept when forced to fit in a classical framework. On the other hand, when treated quantum mechanically, one can immediately see it is nothing other than a consequence of the quantum nature of the electrons; they have wave-like characteristics and occupy a finite volume. In the usual local classical electrodynamics, electrons are treated as point-like charges with no physical extent, nonlocality is simply the breakdown of this approximation.

We shall always assume that the external perturbations are small and in this way we can use the machinery of linear response theory. This means we can express out-of-equilibrium properties of an electron gas in terms of its unperturbed properties [50]. The most general linear relation between the polarization of a system and an applied field is

$$P(r, t) = e_0 \int d^3 r' \int d t' \chi(r, r', t, t') E_{e, ext}(r', t'), \quad (4)$$

where we are careful to distinguish between the total (the field actually experienced by a test charge in the system) and bare external field. $\chi$ is a retarded response function (it respects causality), this would not be the case if we were to write a similar equation but for the total electric field. This equation is nonlocal in space and time, the polarization at $(r, t)$ depends on all other positions and (past) times. For systems which are homogeneous in time the response function can only depend on differences in time, not absolute coordinates. We may convert to the familiar frequency representation using the convolution theorem, thus the material is dispersive and responds differently at different frequencies. This is the reason for the complexity in plasmonics: Maxwell’s equations scale very simply with frequency, but the material response will in general scale non-trivially. For the spatial dependence a local approximation is usually made in plasmonics, mathematically this means we take the response function to be $\chi(\omega) \delta(r - r')$: the response of the system at a point $r$ depends only on the field at that point. This is valid if the wavelength is much larger than all characteristic material dimensions (i.e. the lattice spacing and the electron mean free path for bulk materials). If we ignore the microscopic graininess and average over the complicated atomic fields we end up with the usual macroscopic field equations with simple local constitutive relations.

For a system with translational invariance it is not to difficult to generalise too a nonlocal case. We begin by noting that the response function can only depend on the difference $r - r'$ and, by again using the convolution theorem, we obtain a response function that depends on the wavevector $q$:

$$P(q, t) = e_0 \chi(q, \omega) E_{e, ext}(q, t). \quad (5)$$

This equation tells us that if we perturb an infinite system with a perturbation of wavenumber $q$ and frequency $\omega$ then the system will respond at the same $q$ and $\omega$. Approximations to microscopic response functions can be derived using first order perturbation theory and the random phase approximation (RPA), an example being the density-density response function for a homogeneous electron gas which was first derived by Lindhard [88]. When perturbing a bulk system with light it is permissible to take the long wavelength limit and lose the $q$ dependence, this is simply a consequence of $\lambda_{e, \text{electrons}}/\lambda_{\text{light}} \ll 1$. If we are perturbing the system with a swift electron then higher wavenumbers will be accessible and a nonlocal description is crucial [31].

In general, for finite systems with no translational invariance, the full $(r, r')$ (or alternatively $(q, q')$) dependence must be kept. This will be the case for real surfaces where the electron density changes continuously from the bulk value to zero over a range of the Fermi wavelength. We can expect a corresponding electric field change over the same distance meaning there is a spatial variation on a scale comparable to atomic phenomena and a nonlocal description is important [27], [69]. In local theories the polarization of conductors results in an infinitesimal surface charge layer with corresponding discontinuous electric field change, this is clearly unphysical.

For many years there has been much theoretical work on the topic of nonlocality for surfaces, thin films and NPs [4], [26], [69], [100], [117], [119]. Examples of where the local response fails are resonance shifts in small MNPs [14], [25], [33], [56], [108], [118], [120], [121], [132]–[134], [142], [149]. linewidth broadening [7], [78], [84], [160], longitudinal excitations above the plasma frequency (in both thin films [3], [87] and NPs [4], [127], [154]) and very close dimers [23], [27], [54], [66], [68], [75], [130], [131], [139], [166].

We will now discuss some of the most popular models that have been developed to explain these nonlocal effects and comment on their strengths and weaknesses. We will start with the widely used quasi-quantum hydrodynamic model and then move on to full quantum models of varying approximation.
3.3. The Hydrodynamic Drude Model

The hydrodynamic model (HM) is a simple, popular, intuitive and macroscopic model that, via an electron pressure term, includes some quantum (Pauli exclusion principle) and interaction (Coulomb interaction) effects. It is much more tractable than full microscopic models but the price paid is the neglect of quantum effects like tunneling, Friedel oscillations and electron-hole pair excitation. One of its great strengths is the ease of implementing the method with commonly used numerical techniques in plasmonics meaning it can be applied to arbitrary shapes [100], [144], although computational time will increase dramatically as not only sub-wavelength but also sub-Fermi wavelength features must be resolved. It is even possible to derive some analytical expressions for simple geometries.

The model has a long history beginning with Bloch in 1933 [15] and was applied to plasmonic systems in the 1970s [37]. It has recently become very popular in plasmonics [26]–[28], [117], [119] and there has been the development of an improved version which incorporates charge diffusion kinetics called the semiclassical generalized non-local optical response (GNOR) [104]. The electron gas is described in terms of the scalar fields: \( n(r,t) \) (electron density), \( v(r,t) \) (velocity field) and \( p(r,t) \) (pressure field). Within the model it is assumed that these are linked to the macroscopic electric and magnetic fields via the hydrodynamic equation of motion

\[
\frac{n \partial v}{\partial t} + n (v \cdot \nabla) v + \gamma n v = -\frac{ne}{m} (E + v \times B) - \frac{1}{m} \nabla p. \tag{6}
\]

Damping is added in a phenomenological manner. To proceed one needs to link the electron pressure to the pressure term. Normally the Thomas-Fermi theory is used [135] which is popular because of its simplicity but it neglects correlation and exchange effects of the electrons [117]. After working through the calculation the polarization response is found to have an additional term as compared to the local case

\[
\tilde{P} + \gamma \tilde{P} - \beta^2 \nabla \cdot (\nabla \cdot \tilde{P}) = \epsilon_0 \omega_p^2 E. \tag{7}
\]

The nonlocal parameter \( \beta \), which is proportional to the Fermi velocity, comes out at the wrong value when derived within the HM; it should instead be viewed as a parameter to give the best fit to experiment. The reason for this discrepancy is that the hydrodynamic model is a long wavelength approximation whereas plasmons are high frequency excitations: More precisely, in the HM it is tacitly assumed that the electron liquid is in local equilibrium which is not true for plasmon waves because the electron-electron collision rate is much less than that the plasma frequency [50]. The correct parameter can be obtained via the RPA. In the GNOR model \( \beta \) becomes generalised to a complex number [104]. Equation 7 is now solved self-consistently along with Maxwell’s equations, it is known that retardation is important to include and that taking the static (curl free) approximation introduces unphysical modes below the plasma frequency [119]. The HM has been explored in the context of different excitation sources: light, electrons and electric dipole emitters [26].

Within the HM approximation one must distinguish between the longitudinal and transverse dielectric function; this is not required in the local approximation. The transverse response is unchanged from the local approximation and has the usual dispersion relation

\[
q^2 = \epsilon_T(\omega) \frac{\omega^2}{c^2}, \tag{8}
\]

where \( \epsilon_T \) is the transverse dielectric function; this follows from \( \nabla \cdot P_T = 0 \) and equation 7. Nonlocality for the transverse direction is usually associated with electron-hole pair excitations [50] and hence is not included in the hydrodynamic approximation. Recently there has been work using a nonlocal \( \epsilon_T \) to describe diffuse surface scattering [103]. Longitudinal excitations are given by the condition

\[
\epsilon_L(q, \omega) = 0, \tag{9}
\]

where \( \epsilon_L \) depends on the propagation vector \( q \) as well as the frequency. \( \epsilon_L \) can be easily found from equation 7 by Fourier transforming in space and time and by comparing to equation 5 and is given by

\[
\epsilon_L(q, \omega) = 1 - \frac{\omega_p^2}{\omega^2 + i\omega\gamma - \beta^2 q^2}. \tag{10}
\]

The HM is a longitudinal nonlocal response model [153]. Because of the nonlocality the longitudinal mode condition can be met for a continuous set of frequencies and wavenumbers, in other words it has a dispersion. For frequencies above the plasma frequency it is possible to excite propagating transverse and longitudinal modes, below the plasma frequency both modes exponentially decay. The physical explanation for this lies in the inclusion of a more realistic surface density, the quantum pressure term means electrons cannot be...
These nonlocal effects were first taken into account for metallic spheres by Ruppin who generalised Mie theory to include the longitudinal modes [127], [128]. One of the strengths of the hydrodynamic method is the ability to take into account retardation effects which is not possible in density functional theory (DFT, see next section) calculations [117]. Within the non-retarded limit, where the response is dominated by the dipolar mode, it is possible to obtain an approximate analytical expression for the LSP resonance position and the linewidth (if electron diffusion is included). It is found [104], to first order in $1/R$, that

$$Re[\omega_{sp}] = \frac{\omega_p}{\sqrt{3}} + \frac{\sqrt{2}\beta}{2R},$$

(11)

$$Im[\omega_{sp}] = -\frac{\gamma}{2} - \frac{\sqrt{6}D\omega_p}{24\beta R},$$

(12)

where $D$ is a diffusion constant. It is instructive to compare the surface plasmon resonance for spheres with different radius using the local approximation and the Drude model (with and without phenomenological broadening) and the HM results (with and without the diffusion constant $D$).

This is presented in fig 3 [104]. We can clearly see that nonlocal effects are present in both the resonance frequency and the linewidth. Also, we can appreciate that without the diffusive term, $D$, the HM captures the blueshift of the plasmon resonance (see fig 3a). On the other hand, the local Drude approximation does not have any size dependence for the resonance position or broadening. In order to recover the broadening in the local framework a phenomenological size-dependent damping has to be considered (Kreibig broadening, see fig 3b). This is consistently captured in the HM by the diffusive term, until the development of the GNOR approach the inclusion of nonlocal line broadening was implemented by an ad hoc phenomenological model by Kreibig [77], [78] which considers the electrons as having a limited mean-free-path because of the surface. This damping reflects the surface to volume scaling and hence has the form

$$\gamma = A \frac{\nu_F}{R}.$$  

(13)

The inclusion of a simple and intuitive method in explaining the increase of damping in small MNPs is quite a coup for the hydrodynamic model.

There are full quantum theories of size-dependent damping which in the quantum picture is called Landau damping and is explained as the coupling of plasmons with electron-hole pairs, due to the momentum transfer via the surface, which leads to fragmentation of the plasmon peak and is the main decay mechanism in the small size regime [19], [83]. The are various full quantum theories of Landau damping which agree with the $1/R$ dependence of equation 13 but values of the coefficient $A$ depends on the particle shape and approximations made [67], [83], [86], [92], [160], [161].

Of course there are problems with the HM, the usual approximation made is that the equilibrium density has a step profile (it goes to zero discontinuously at the MNP surface) which corresponds to an infinite potential barrier. This is a problem if the spill out is a dominant mechanism for the LSP shift (as for alkali MNPs) as the hydrodynamic model can only predict blueshifts: this is simply because the crude boundary conditions push up the electron density inside the particle giving a larger effective plasma frequency. Recently there have been studies including more realistic density profiles [30], [145], [159]. Including the spill out leads to a new resonance called the multipole surface plasmon or Bennett resonance [10]. A major problem with the HM is that it does not take into account d-band screening. These effects are much more easily accounted for.
in more sophisticated models, for example in [94] the plasmonic response of atomically thin nanodisks are calculated with d-band screening taken into account in a very natural manner along with the RPA equations.

The HM has also been extensively used to model the optical response of close dimers (as well as particles close to surfaces) [27], [28], [104], [144]. It has been found that the HM leads to reduction in plasmonic enhancement, in fact nonlocality can play a larger role than material losses in reducing the plasmonic response in confined geometries [27].

### 3.4. Time Dependent Density Functional Theory

We are keen to stress in this review that the optical response of a nanostructure is governed by the self consistent response of the electron density to the perturbing field. Clearly it is an ideal problem to attack using the method of time dependent density functional theory (TDDFT) [126]. Within TDDFT the electrons are modelled as moving independently within a time dependent effective potential which is a functional of the electron density. This potential is made up of the confining potential (from the ions), the Hartree potential and the exchange-correlation potential which is given by the functional derivative of the exchange-correlation energy functional. The exact form of the energy functional is not known and so an approximation needs to be made; the nature of this approximation determines the accuracy of the model. Setting the exchange-correlation contribution to zero gives a time dependent Hartree theory which is equivalent to the RPA [50].

A common approximation used for modelling MNPs is the local-density approximation (LDA) which was first proposed by Kohn and Sham [72] and involves expressing the exchange energy of an inhomogeneous electron gas in terms of an integral over the exchange energy density of a homogeneous electron gas (which is exactly known) evaluated at the local density [147]. This is a good approximation for MNPs down to a very small size as they are saturating systems and hence have a nearly constant density in the interior, this holds for a system where the radius scales with \( N^{1/3} \) [99]. Only at the surface is there serious deviation from the bulk density and this is why these types of system can often be modelled to a good approximation by bulk models.

The pioneer in applying the formalism of DFT to small metallic clusters was Ekardt who was able to calculate the effective mean potential and electron density within the LDA and Jellium approximation [39] and soon was able to extend his model to the dynamic case (within a frequency space description) and explore excitations of the system [38], [40], [41]. This was inspired by earlier work on atomic systems [137], [162]. Another notable work is by Yana and Bertsch [155] who used real time calculations for small clusters and obtained similar results to the earlier frequency space calculations. Real time TDDFT is advantageous for some applications as scales better with particle size and it does not rely on linearised response equation as is the case for the frequency space versions [146] and may be applied to nonlinear calculations [96].

More recently larger NPs were studied, within a frequency-space representation, by Nordlander’s group who studied nanoshells [114], dimers [167] and nanorods [168]. There has also been much work with the real time representation on linear atomic chains/quantum wires (usually with the explicit atomic structure as the reduced dimensionality mean they are computationally 'cheap' to simulate) [36], [157], [158] and spheres (with the jellium approximation [146] and with the ionic background explicitly included for various metals like sodium [85], [148], gold [20] and silver [156]) where, for the noble metals, d-electrons can be explicitly included. Electronic structure calculations within the jellium approximation for simple geometries, where symmetries can be exploited, have proved tractable enough to study systems of up to 10^5 electrons, allowing comparison to classical results. For the full atomic structure, with the use of pseudopotentials, simulations involving 100−200 electrons can be achieved. This how allowed the study of plasmon resonances from the molecular to the bulk and the intriguing question of exactly when a resonance can be labelled as a plasmon is being tackled (it is not as simple as simply identifying a large oscillator strength with a certain mode) and proposed methods include the scaling method of Bernadotte et al [13] (which is based on the different dependence of the energy of excitations on the electron-electron interaction) and a beautiful new concept known as the plasmonicity index which gives a measure of an excitations ability to enhance external EM radiation [21]. Employing full \textit{ab initio} methods, with the atomic structure included, is important not only for accurate quantitative results (for instance it has been found that there is a significant lower field enhancement, as compared to the jellium model, for \textit{ab initio} studies of dimer systems [164]) but also for capturing the full physics in the increasingly tiny systems studied. As an example we highlight reference [95] where it was shown that atomic reorganisation results in a jump-to-contact instability between two Na_{138O} clusters which leads to a sudden change in the intensity and position of the plasmon resonance.
3.5. A Comparison of Models: Dimer systems

In this section we review work on the plasmonic response of strongly coupled NP dimer systems and in particular focus on comparing the most commonly used models. There has been a lot of recent interests in determining exactly how TDDFT and quasi-quantum models compare [136], [141]. It is clearly advantageous to use the simpler quasi-quantum models if one can get away with it!

One of the big challenges in quantum plasmonics is to model the plasmonic response of two close NPs (or a NP close to a metal surface where it will couple to its image charge). The electrons in the two approaching NPs will electrically couple (like a capacitor) which results in large enhanced fields (the intensity of the plasmonic hotspots can be orders of magnitude larger in intensity than the incident light) and has important applications such as single molecule surface-enhanced Raman scattering (SERS) [93]. In the classical regime (gap size larger than 1 nm) Drude-type models work well and the underlying physics is well explained by the idea of plasmon hybridisation [107]. The longitudinal bonding dipolar plasmon (BDP) dominates the response and is known to redshift with decreasing separation. At gap sizes around 1 nm purely classical models begin to fail and the system is better described by the hydrodynamic model. For sub-nanometre gaps (around 0.5 nm, or larger for gaps bridged by molecules [139]) quantum tunnelling becomes possible and charge transfer between the two particles occurs. A new lower energy mode appears called the charge transfer plasmon (CTP) which corresponds to an oscillating electrical current across the gap [44], [80], [130], [131]. This mode will be highly dependent on the conductance of the junction (and hence the separation). It is an interesting question to ask how nonlocal/quantum mechanical effects weaken the field enhancement as predicted by classical theories (which allow for infinite field enhancement) [27], [166]. It is particularly difficult to use full quantum models for systems such as two nearly touching spheres as the spheres may be quite large but the gap small!

The sub-nanometre regime has been explored experimentally by Baumber and co-workers et al [130] who were the first to observe quantum effects in plasmonic dimer systems. Ingeniously two gold NP atomic force microscope tips were used for the tunnelling system which allowed simultaneous measurement of the optical and electrical response. They were able to observe the capacitive regime (50 nm → 1 nm) and the quantum regime (1 nm → 0 nm). Another relevant experiment was done by Scholl et al [131] using EELS. Both experiments found that tunnelling effects become important on the order of 0.5 nm. This is incredibly close, especially when one remembers that the lattice constant of gold and silver is about 0.4 nm, and is is a consequence of the tunnelling probability exponentially decaying with distance. Above 0.5 nm the hydrodynamic model can be expected to be valid and indeed this was what was found by Ciraci et al [27] who looked at MNPs on gold films separated by very precise spacer layers of molecules. Looking at gap sizes from 0.5 to 2 nm they found excellent agreement with the hydrodynamic model. We also mention experiments done by Yoon's group [23], [66] where the nonlocal quantum regimes were observed for colloidal solutions of molecularly linked dimers. Via the process of photooxidative desorption of the linkers, they were able to control the interparticle distance down to the point of merging using UV light. In another experiment, by Gordon's group, the onset of the quantum tunnelling regime has also been observed by monitoring the third harmonic signal which is very sensitive to the near-field intensity [54].

An interesting idea is to use a strong electric field to enhance the tunnelling rate using Fowler Nordheim tunnelling (emission of electrons induced by an electrostatic field) [152]. Another fascinating modification to the dimer setup is to bridge the gap with suitable molecules that will increase the tunnelling distance. This intriguing combination of molecular electronics with plasmonics was shown by Nijhuis' group who observed, using EELS, a charge transfer plasmon up to a separation of 1.3 nm for silver nanocubes functionalised with a monolayer of self-assembled molecules (SAMs) [139]. They found that the frequency of the CTP was dependent on the type of bridging molecule, for instance the size of the molecules HOMO-LUMO gap will dictate the tunnelling rate (we take this opportunity to point out that there has been some dispute of their interpretation of the experimental results regarding the CTP using the quantum corrected model [70]). This opens up the possibility of new biosensors that use changes in the tunnelling current to detect chemical concentrations and processes, for instance Benz et al have shown that difference of a thiol group (which allows the formation of a covalent bond to gold and hence can allow charge transfer) between biphenyl-4-thiol and biphenyl-4,4’-dithiol results in a shift of the coupled plasmon mode depending on the mole fraction of the two molecules [11]. Other possible applications include electrically controlled plasmonic nano switches and gates which operate at the single atom level [43], [52], [91]. The role of various bridging molecules has been explored theoretically (for example with water layers where the roles of dielectric screening and tunnelling were compared [140]) and experimentally (a remarkable recent experiment has shown charge transfer for DNA tethered NPs up to a gap size of 2.8 nm [82]). It also possible to use molecules as a probe...
of the field enhancements present in dimer systems to reveal quantum effects [166]. In SERS, the Raman scattering from molecules scales to the fourth power with the local field, this means small Raman molecules can provide a very sensitive measurement of plasmonic enhancement. For a recent and in-depth review on the emerging field of molecular electronic plasmonics the reader is directed to reference [150].

At this point we mention an *ad hoc* but useful model called the quantum corrected model (QCM) [44], where quantum mechanical effects are incorporated via a local approximation within the gap region. In this model a fictitious material is placed in the gap to account for tunnelling, the properties of the material are calculated using the electron transmission probability through the potential barriers seen by an electron. This problem can either be solved using *ab initio* calculations or using simple model potentials like square wells. In general one would allow this material to spatially vary. The model has been found to agree well with full quantum mechanical calculations using TDDFT and unlike the HM is valid for quantum tunnelling regimes.

A comparison between different models is done in fig 4 where we show the local field enhancement in the gap of a dimer (two 9.8 Å Na nanowires) varying the gap size [141]. The considered models are the TDDFT, the QCM, the local Drude calculations and the HM. As expected TDDFT and QCM agree for very small gaps below 0.5 Å, i.e. when the tunnelling dominate, while HM and local Drude overestimate the enhancements.

### 4. Quantising Light in Plasmonics

We will now look at what happens if non-classical light is used to excite SPPs. In a full quantum theory the SPP, like photons, come in quanta of energy; individual surface plasmons are indivisible which is remarkable given their many body nature. Quantization of plasmons is in principle (if we neglects plasmon damping) similar to normal photons, each mode can be thought of as an harmonic oscillator, hence they have very similar behaviour, for instance SPPs are bosonic in nature [35]. There have been a number of relevant experiments which have shown this in quantum optics type experiments with single photons. The quantum theory for SPPs was first introduced by Ritchie [42]. The first full microscopic theory for dielectrics was proposed by Huttner and Barnet [64] who used the method of Hopfield [62] to quantise the matter field as well as the light field.

The first experimental evidence for the quantum light nature of SPPs was reported by Altewischer et al in 2002 [2] who found that photons preserve their polarization entanglement if converted into SPPs at a metal (subwavelength) hole array and then converted back into photons again. Two quantum systems, system A in state $|\psi\rangle$ and system B in state $|\chi\rangle$, are entangled if the (tensor) product of the total system cannot be written in a product form $|\psi\rangle_A |\chi\rangle_B$. This will occur if there exists a correlation between the two states which may exist over arbitrary distances in space and time and hence is a demonstration of quantum nonlocality (not to be confused with the nonlocality of earlier). Entanglement has no classical counterpart and it a vital process for possible quantum information and computing processes. The discovery of SPP entanglement seems surprising at first given the large number of electrons that make up a SPP as well as its lossy nature. But, as Barnes [8] points out, we would not be surprised if entanglement survived reflection of a metallic mirror despite the same collective electron behaviour being responsible!

Entanglement in other degrees of freedom is also possible. Fasel *et al* have shown energy-time entangle-
Fig. 5: Experimental setup used for a plasmonic version of the Hong-Ou-Mandel effect. The last figure shows a pronounced coincidence dip, a key signature that SPPs are bosons (adapted from [35]).

5. Conclusions And Future Directions

Both quantization of matter and light is beginning to play a role in plasmonics and we are on the cusp of being able to understand and manipulate plasmonics on the quantum scale. This is thanks to recent developments in nano-fabrication and nano-detection which have ignited interest in quantum plasmonics.

For the electron quantization of plasmonics, quantum size effects can occur for very small MNPs ($\lesssim 2$ nm) as well as tunnelling effects between nearly touching nanoplasmonic components ($\lesssim 0.5$ nm). There are a number of models used in describing these phenomena, and for the design and modelling of future quantum plasmonic devices it will be crucial to understand when certain approximations are valid. These models may also help bridge the behaviour of large classical MNPs and very small clusters, this is still an open question and can help define the boundaries of condensed matter physics.

Whilst the power of TDDFT is apparent for the foreseeable future TDDFT calculations involve considerable computational effort for even the smallest MNPs. This means phenomenological methods like the MH will continue to play a large and crucial role in plasmonic theory, but one should always be wary of pushing these models beyond where they are appropriate.

We have also discussed the quantum light side of plasmonics which has become a hot topic on the back of some impressive experiments. With progress on both fronts moving ahead it is starting to become relevant to ask where might one have both quantization of light and matter to be important. Only with a full quantum model will a complete fundamental description of plasmon polaritons be possible.

Many body quantum theories for the behaviour of the electron gas already exist and from quantum electrodynamics we have a description of the photon, a merging of these theories will provide a full description of quantum plasmonics. Such theories will be necessary for developments in a deeper nanoplasmonics area including hot electrons, molecules interacting with plasmonic nanostructures and quantum information devices as well as aiding our understanding of this fascinating area of physics.


Acknowledgment
This work was supported by EPSRC (grant number EP/L024926/1).

References


