nature materials
SUPPLEMENTARY INFORMATION
DOI: 10.1038/NMAT4169

Highly confined low-loss plasmons in graphene–boron nitride heterostructures

CONTENTS

I. Optical signal model i
A. Signal in the bulk (local and launching response) i
B. Edge-reflected fringes ($\lambda_p/2$-period contribution) iii
C. Edge-launched fringes (λ_p-period contribution) iv

II. Fringe fitting (parameter extraction) iv
A. Detection of graphene edge location v
B. Background subtraction v
C. Complex Hankel Transform vi

III. Mode calculations vii
A. Nonlocal conductance vii
B. Dispersion relation comparison viii

IV. h-BN permittivity model ix
A. Parameters of h-BN permittivity and remarks x
B. h-BN losses in thin films xi

References xi

I. OPTICAL SIGNAL MODEL

Ideally, the tip interacts only with the local graphene underneath its apex, responding to the electric susceptibility of the graphene and acting as a localized "point source" for exciting the plasma wave. The plasma wave spreads out as a circular wave (2D radial wave), reflects off the nearby edge of the graphene and returns to the tip. Even in the ideal lossless case, only a small part of this returning wave couples to the tip, due to geometrical decay. In practice, there are further interaction pathways: the light path does not only interact with the tip but also directly with the sample, and moreover the tip does not solely interact with the graphene under its apex.

This section describes the expected optical signal for a reflected circular wave, that has $\lambda_p/2$-period fringes as well as the origin of the fringes with λ_p-period and their expected optical signal.

A. Signal in the bulk (local and launching response)

In Fig. 2b of the main text, we demonstrate that the change in optical signal approximately follows the ac conductivity of the graphene. It is instructive to consider why this is, and why the correspondence might not be perfect.

To exactly calculate the optical signal measured in the s-SNOM is a complicated matter, however to first approximation, the incoming and outgoing light are only coupled to the charge oscillations in the metallized tip. In the near-field limit, these charge oscillations are electrically (capacitively) coupled to the device under study. In this picture, the optical signal is essentially related to part of the tip's self-capacitance that depends on tip-sample distance.

In the limit where the tip-sample system is non-resonant, the tip response can be calculated by some linear convolution of the surface's physical optical response. In Fourier space:[S1]

$$s(\omega) \approx \int w(k) r(\omega, k)\, dk, \tag{S1}$$

where $w(k) \sim k^2 \exp(-2kR)$ is a bell-shaped weighting function with a peak at $k \approx 10/R$, where R is the tip radius. The surface optical response is embedded in $r(\omega, k)$, the evanescent reflection coefficient for transverse magnetic waves having in-plane wavevector k and angular frequency ω.

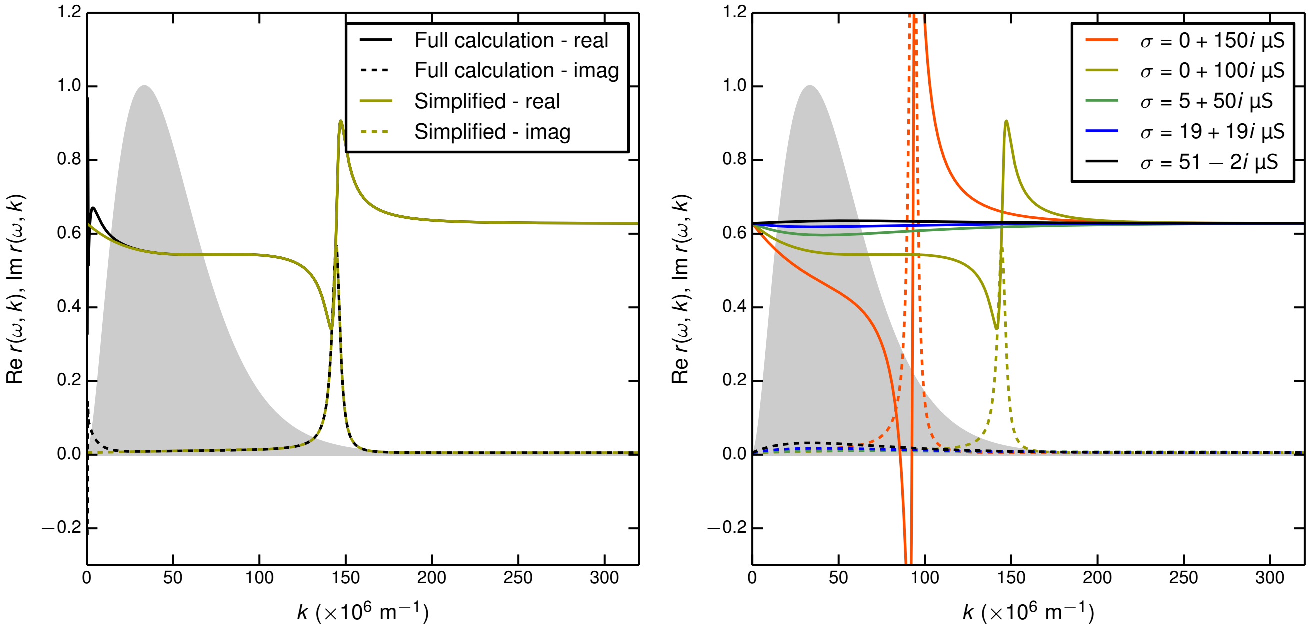

FIG. S1. Dependence of the reflection coefficient on wave vector k. a) Comparison of the simple quasi-electrostatic result (S2) for the BN-Gr-BN system, against the full electromagnetic calculation (Sec. III) that also includes the underlying SiO2. The real part is shown as a solid line and the imaginary part as a dashed line. Here the frequency is a typical $\frac{\omega}{2\pi} = 30$ THz, and for simplicity we have taken the conductivity of the graphene to be $\sigma = 10^{-4} i$ S, which corresponds to a carrier density of $n_s \approx 2.6 \times 10^{16}$ m^{-2} The filled-in curve shows the weighting function $w(x)$. b) Influence of varying the conductivity on (S2). The conductivity variation here ranges from a higher carrier density ($n_s \approx 5.0 \times 10^{16}$ m^{-2} for orange curve) to zero carrier density (black curve). For small k_x values it is apparent that the shift in r is proportional to $-\sigma/i$.

In Sec. III we describe the general procedure to calculate $r(\omega, k)$ numerically, for an arbitrary stack. In the quasi-electrostatic limit ($k \gg c/\omega$), we can write a simple expression for a dielectric-conductor-dielectric stack.

$$r = \frac{\varepsilon - \varepsilon_0 - (\varepsilon + \varepsilon_0)\frac{\alpha}{1-\alpha} e^{-2\eta k_x t}}{\varepsilon + \varepsilon_0 - (\varepsilon - \varepsilon_0)\frac{\alpha}{1-\alpha} e^{-2\eta k_x t}}, \tag{S2}$$

Here, the upper dielectric is taken to have thickness t and the lower to be infinite thickness, and the 2D conductor to be of zero thickness. As the dielectric (h-BN) is anisotropic with in-plane permittivity ε_{xx} much different from out-of-plane permittivity ε_{zz}, we have defined the effective permittivity,

$$\varepsilon \equiv \sqrt{\varepsilon_{xx}\varepsilon_{zz}}, \tag{S3}$$

and the effective field confinement factor,

$$\eta \equiv \sqrt{\frac{\varepsilon_{xx}}{\varepsilon_{zz}}}, \tag{S4}$$

which in our experimental frequency range are $\varepsilon/\varepsilon_0 \approx 4.0$–$4.7$ and $\eta \approx 2$. The effect of the graphene is captured in the parameter α, defined as:

$$\alpha \equiv \frac{\sigma}{2\varepsilon\omega i} k_x. \tag{S5}$$

Fig. S1 plots the reflection coefficient for a typical frequency in our experiment. As k increases, the reflection coefficient probes the optical response closer and closer to the surface. At very high $k \gtrsim 1/(\eta t)$ we only see the response of the top dielectric, which has the form:

$$r(k \to \infty) = r_\infty = \frac{\varepsilon - \varepsilon_0}{\varepsilon + \varepsilon_0}. \tag{S6}$$

In our case $r_\infty \approx 0.6$. As k is lowered, we meet at some point the condition $\alpha \approx 1$, leading to a resonance due to the denominators $1-\alpha$ in Eq. (S2). This is essentially the location of the plasmon; in fact the precise condition is a pole in r, which occurs slightly away from $\alpha = 1$ due to the finite t effects. At the plasmon resonance, r shows a strong peak in its imaginary part, indicating energy transfer to the plasmon. This high-k limit and the plasmon resonance are however both weakly coupled to the tip, being in the tail of the function $w(k)$.

The dominant contribution in our case comes from small $k \approx 30 \times 10^6$ m^{-1}, which is generally below the plasmon resonance. Being below the plasmon resonance, we can make the approximation $\alpha \ll 1$. Expanding (S2) to first order in α, we find:

$$r(k \ll q_{\mathrm{p}}) \approx r_\infty - (1 - {r_\infty}^2) e^{-2\eta k_x t}\alpha. \tag{S7}$$

From this expression it is apparent why changes in graphene conductivity appear proportionally in our optical signal measurements. For fixed frequency, the permittivity parameters ε, η, r_∞ are fixed. In essence, the tip can only couple well to small k_x ($\alpha \ll 1$), and so *regardless of further details of the tip coupling*, the presence of the graphene causes a small perturbation that is proportional to its local conductivity.

Beyond the simple argument presented above, a number of further influences should be considered, and so we do not expect exact correspondence. First, it is only to first order in α that the signal should be proportional to conductivity. Higher order terms certainly do contribute, e.g., our imaging of plasmons requires the plasmon pole to contribute to the optical signal. The plasmon draws energy from the tip and carries it away, and this energy loss appears similarly to dissipation (i.e., like $\mathrm{Re}\,\sigma$ or $\mathrm{Im}\,\varepsilon$). As carrier density increases and the plasmon couples more efficiently, this energy loss becomes stronger. Second, graphene can screen the influence of the dielectric layers underneath, in particular the SiO_2. This screening effect also changes with carrier density, and so the influence of the SiO_2 is variable.

B. Edge-reflected fringes ($\lambda_{\mathrm{p}}/2$-period contribution)

It is well known that in the far field, the amplitude of a lossless circular wave decays as $\sim 1/\sqrt{r}$, where r is distance from source. This ensures energy conservation on the wavefront, which has circumference $2\pi r$. Mathematically, this appears in the 2D Helmholtz equation with point source at $\vec{r}_{\mathrm{s}}$,

$$\nabla^2 E(\vec{r}) + q^2 E(\vec{r}) = -\delta(\vec{r} - \vec{r}_{\mathrm{s}})$$

which has the solution

$$E(\vec{r}) = \tfrac{i}{4} H_0^{(1)}(q|\vec{r} - \vec{r}_{\mathrm{s}}|) \tag{S8}$$

where $H_0^{(1)}(z)$ is the first Hankel function of order zero. In the case of plasmons, the wave field E may represent charge density or out-of-plane electric field. Equation (S8) remains a solution also when q is complex, and describes a decaying wave for $\mathrm{Im}(q) > 0$. As expected, the asymptotic decay of the Hankel function is $H_0^{(1)}(z) \approx \sqrt{\frac{2}{i\pi z}} e^{iz}$.

Now, consider the case where the tip is near a straight edge – the circular wave will reflect off this edge. Assuming that the reflection coefficient is independent of the wave angle, then the reflected wave can be described using the mirror-image method. Let the straight edge be defined by the line $r_x = 0$, and let the tip be at location $\vec{r}_{\mathrm{tip}} = (x, 0)$. Its mirror image is at $(-x, 0) = -\vec{r}_{\mathrm{tip}}$. The resulting total wave will be:

$$E(\vec{r}) = E_{\mathrm{launch}} \tfrac{i}{4} H_0^{(1)}(q|\vec{r} - \vec{r}_{\mathrm{tip}}|) + E_{\mathrm{refl}} \tfrac{i}{4} H_0^{(1)}(q|\vec{r} + \vec{r}_{\mathrm{tip}}|)$$

The reflection coefficient $E_{\mathrm{refl}}/E_{\mathrm{launch}}$ is not necessarily unity. It is expected to be phase shifted[S2] and also its magnitude will be smaller than unity due to energy loss from light emission and scattering at the edge.

We have used the s-SNOM in interferometric mode and so the measured signal is proportional to this complex field.[S3] Ideally, the out-scattered light depends only on the local coupling to $E(\vec{r}_{\mathrm{tip}})$, and so $s \propto E(\vec{r}_{\mathrm{tip}})$. The field from the first term (launched wave) forms part of the bulk signal. The second term adds to the bulk signal and generates the interference fringes. We thus expect:

$$\xi(x) = \xi_{\mathrm{bulk}} + A H_0^{(1)}(2qx),$$

where ξ_{bulk} collects together all contributions that would already occur away from the edge – local response, plasmon launching, etc., and, the complex coefficient A collects together factors of reflection, in-coupling, out-coupling, etc.

C. Edge-launched fringes ($\lambda_{\rm p}$-period contribution)

Broken translational symmetry at the edge provides for matching the small photon wavevector with the large plasmon wavevector. As a simple model, one can think of the wave $E(x)$ being launched by an oscillating electric field at the edge.[S4] This produces a plane wave plasmon without additional geometrical decay:

$$E(\vec{r}) \sim E_{\rm edge} e^{iq_{\rm p} r_x}, \tag{S9}$$

so that a contribution proportional to $E_{\rm edge} e^{iq_{\rm p} x}$ is added to $\xi_{\rm opt}(x)$. This is the case for plasmons being launched directly by the illuminating laser spot which is effectively a plane wave on these nanometer length scales.

There are however other possibilities that lead to plasmons that travel only once the tip-edge distance x. One possibility is the reverse of the above, that the plasmons launched at the tip are scattered to light at the graphene edge. In this process the geometrical decay is less obvious: the plasma wave decays geometrically from the tip so that the field at the edge decays as $1/\sqrt{x}$, yet also the wave arrives in-phase over a larger section of the edge, tending to cancel this decay.

Another possibility is that the near-field tail of the tip interacts with the edge and launches a plasmon there, a plasmon which is then received at the tip after travelling x. This is similar to the far-field case, except the tip acts as a field-enhancing mediator between light and edge. Here additional geometrical decay is expected because the electric field of the near-field tail depends on the tip-edge distance. It is not clear what distance dependence this near-field profile should take—monopolar, dipolar, or somewhere in-between. This profile would also be modified by lateral field focussing by the h-BN. Again, the reverse process (launching at tip, then the long-ranged tail of the edge plasmon field interacts with the tip) is also possible.

To allow for these various mechanisms we include a variable geometrical decay in this contribution to $\xi_{\rm opt}$:

$$\xi_{\rm edge}(x) \propto \frac{e^{iq_{\rm p} x}}{x^a + R^a},$$

where R is the tip apex radius, included to limit the divergence in this expression. In the picture of plasmon plane wave launching at the edge, this would correspond to taking a distance-dependent edge field in Eq. (S9)

$$E_{\rm edge}(x) \propto \frac{1}{x^a + R^a}.$$

The optical signal then shows the period of an edge-launched plane wave, but with additional geometrical decay whose origin is unclear.

II. FRINGE FITTING (PARAMETER EXTRACTION)

In order to extract parameters, such as propagation length, from the fringe signal, we need an accurate model of the expected signal for a given amount of damping. Based on the previous section, we have a decent model for the decay of fringes away from the edge:

$$\xi_{\rm opt}(x) = \xi_{\rm bulk}(x) + AH_0^{(1)}(2q_{\rm p}x) + B\frac{e^{iq_{\rm p} x}}{x^a + R^a} \tag{S10}$$

where the fitting parameters are complex A, B and $q_{\rm p}$, and real a. The tip radius is fixed to $R = 25$ nm.

There are some complications that prevent us from direct fitting of the raw data:

- The location of the edge, $x = 0$, needs to be detected in some way.
- The background part of the signal, $\xi_{\rm bulk}(x)$, is not known a priori and we see clear signs of spatial variations. Fortunately, these variations (due to carrier density gradients) appear to be gradual.
- The model in Eq. (S10) does not necessarily hold for small values of x. For the first fringe, the tip coupling mechanism may become very different than when the tip is over the bulk. Direct fitting of the data with equal residuals weighting is not suitable in this case.

The edge we detect from the topographic data of the s-SNOM apparatus, taking into account tip convolution effects. To avoid biases from the unknown $\xi_{\rm bulk}$ and the unknown first-fringe behaviour, we subtract a smooth background from the signal/model, and then perform fits in a transformed version of the signal/model. In the following we describe this procedure in great detail.

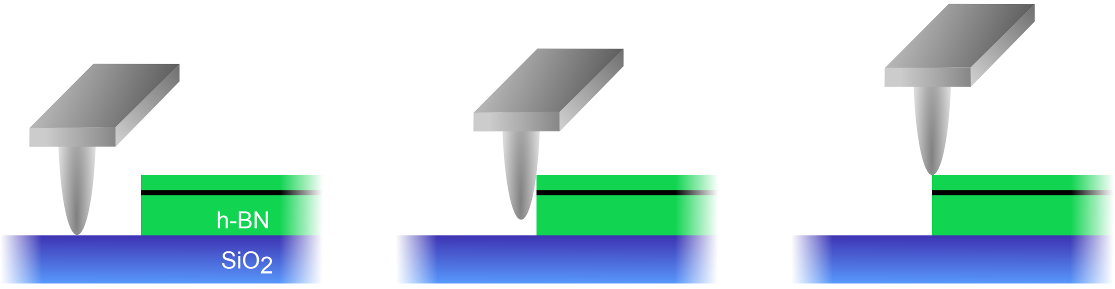

FIG. S2. Tip convolution effects make the topography edge appear away from the graphene edge. We define $x = 0$ to occur when the tip is centered directly above the graphene edge—this is the situation depicted in the third panel.

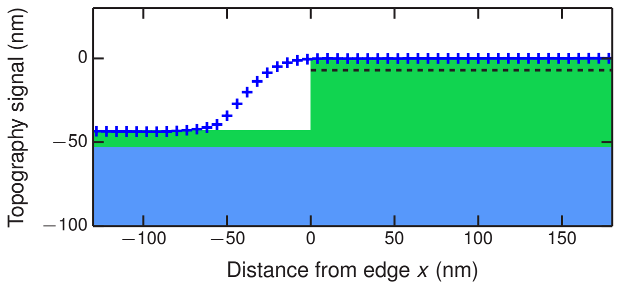

FIG. S3. Topographic signal from measurement (blue crosses) with the interpreted actual topography underneath. The graphene (dashed line) is assumed to terminate at the end of the round feature, since this round feature is interpreted as a tip convolution effect.

A. Detection of graphene edge location

In separating out the contributions from geometrical decay from exponential decay, it is important that the location of $x = 0$ (the graphene edge) has been determined with accuracy. An error in this determination leads to error in the extracted damping.

We have chosen to use our topographic data to determine this edge location. It is well known that tip convolution artifacts result in modified appearances of sharp edges in scanning probe microscopy. We assume that our physical etched edge is sharply vertical as illustrated in Fig. S2, such that the rounding and sloping apparent in the topographic signal is purely due to the AFM tip convolution (Fig. S2). As a result, the edge is located directly beneath the point where the rounding convolution ends, illustrated in Fig. S3. With the chosen edge-detection algorithm it is only possible that the graphene edge is actually further on the left in Fig. S3 which would lead to an underestimation of our extracted inverse damping ratios. Note that even if the edge were not strictly vertical, the error in x would be on the order of a few nanometers since the graphene lies only 7 nm under the surface.

B. Background subtraction

Since $\xi_{\mathrm{bulk}}(x)$ is not known a priori, we can only estimate it from the dataset itself. After discarding the data for $x < 0$, we estimate $\xi_{\mathrm{bulk}}(x)$ by smoothing the measured $\xi_{\mathrm{opt}}(x)$. The difference,

$$\delta\xi_{\mathrm{opt}}(x) = \xi_{\mathrm{opt}}(x) - \xi_{\mathrm{smooth}}(x) \tag{S11}$$

should then be free of influence from the unknown ξ_{bulk}.

Background subtraction always results in removal of some of the desired signal, and is a well known source of statistical bias. In this case, background subtraction leaves transient artifacts near $x = 0$ due to the abrupt termination of the signal, and also selectively removes part of the fringes depending on their period (i.e., affecting more the λ_{p}-period fringes than $\lambda_{\mathrm{p}}/2$-period fringes). In order to give a fair comparison, we apply the same background subtraction procedure to the models used in the fit.

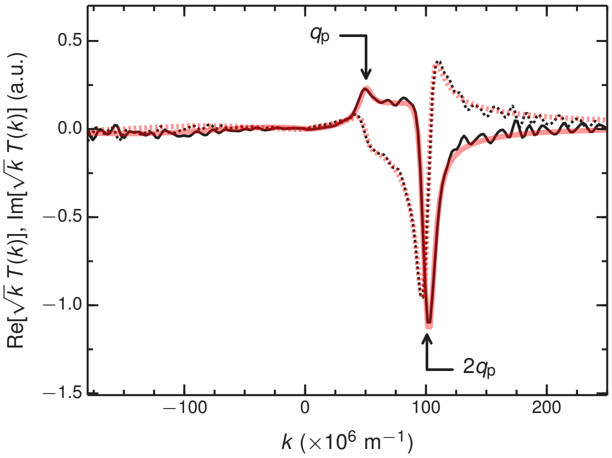

FIG. S4. Complex hankel transform the data in Fig. 3(a) in the main text. The real part are the solid curves and the imaginary part the dotted curves. Black shows the measured data and red the fit.

C. Complex Hankel Transform

Our goal is to access the asymptotic decay away from the edge, where we suppose tip coupling details are captured in the position-independent parameters A, B, a. Close to the edge, such as with the first fringe, the tip coupling details are not necessarily this simple. To this end, we perform fitting not in flat $\delta\xi_{\rm opt}(x)$ space but rather in a transformed $\delta\xi_{\rm opt}(x)$, in effect de-emphasizing the weight of the signal near the edge. One possible approach here would be to take a Hankel transform of the $\delta\xi_{\rm opt}(x)$; the Hankel transform is analogous to a Fourier transform, but more appropriate for circular symmetry, and it naturally gives stronger weight to a larger distance from origin. The Hankel transform is however only a real transform, and does not mix together the real and imaginary parts of $\delta\xi_{\rm opt}(x)$ in a convenient way.

We therefore use a "complex Hankel transform" of the following form:[S5]

$$T(k) = \frac{1}{2}\int_0^\infty x[H_0^{(1)}(kx)]^* \delta\xi_{\rm opt}(x) u(x)\, dx. \tag{S12}$$

This transform has the desirable property that e^{iqx}-type wave will transform to a peak near $+q$, and a e^{-iqx} wave will transform to a peak near $-q$. Note that unlike the proper Hankel transform, this transform is not simply invertible,[S6] however it is linear and successfully distinguishes $+q$ and $-q$ waves. The function $u(x)$ in Eq. (S12) is a "window" function, used to select an appropriate range including sufficient fringes but without too much influence from noise. We use the window $u(x) = 1 - \sin^2(\frac{\pi}{2}x/L)$ which produces a smooth cutoff as x approaches L, with $L = 1\,\mu$m.

After applying Eq. (S12) to our background-subtracted and windowed fringes, we observe a function with two strong peaks (Fig. S4), one peak at $q_{\rm p}$ corresponding to processes where the plasmon travels x, and the other peak at $2q_{\rm p}$ when the plasmon travels $2x$. Note the absence of peaks at negative k, which confirms that we retrieve the phase of the light signal correctly and that the plasmons have positive group velocity. The peak widths in Fig. S4 are related to the decay, though also affected by our choice of background subtraction and windowing procedures. To obtain a fair comparison we can calibrate these peaks against a model with known decay. What we do is to perform the same background subtraction, same windowing, and same transform on the model described in Eq. (S10). We then fit the transformed model onto the transformed data, with equal weighting of the residuals $\sqrt{k}(T_{\rm data}(k) - T_{\rm model}(k))$ for equally-spaced k values, over a specified k range around the peaks.

While the above procedure may seem to overcomplicate matters, we stress that we have only performed a linear transformation on the data and model, and so we are effectively performing non-linear least squares on the source data but with modified residual weights. The ultimate proof of this technique is the quality of fits (e.g., Fig. 3 of the main text, which is very good for the range of parameters presented in the manuscript.). Besides being a reliable way to extract damping information, this technique also allows us to measure accurately the wavelength of fringes that are nearly invisible in the raw data. An additional benefit is that we can directly visualize (Fig. S4) that there are not additional components in the data as might correspond to $3q_{\rm p}$, $-q_{\rm p}$, etc., thereby confirming that the interferometric detection technique has precisely measured the light phasor.

III. MODE CALCULATIONS

We use the AC Maxwell equation for fields oscillating as $\exp(-i\omega t)$ in time,

$$\vec{\nabla} \times (\vec{\nabla} \times \vec{E}) = \frac{\omega^2}{c^2}(\vec{E} + \tfrac{1}{-i\omega\varepsilon_0}\vec{J}), \tag{S13}$$

with current given by

$$\vec{J}_{\rm diel} = -i\omega(\varepsilon - \varepsilon_0)\vec{E}$$

in the dielectrics (note ε is a rank-2 tensor in h-BN), and by the nonlocal 2D conductivity relation

$$\vec{J}_{\rm gr}(\omega, x, y, z) = \delta(z - z_{\rm gr}) \iint dx'\, dy'\, \vec{E}(\omega, x', y', z_{\rm gr})\sigma_{\rm NL}(\omega, x - x', y - y')$$

in the graphene, where $z_{\rm gr}$ is the height of the graphene and $\sigma_{\rm NL}(\omega, x, y)$ is its nonlocal 2D conductivity function.

We neglect magnetic susceptibilities, whose bound currents would take the form $\vec{J}_{\rm mag} = \frac{1}{i\omega}\vec{\nabla} \times [(\mu_0^{-1} - \mu^{-1})\vec{\nabla} \times \vec{E}]$, i.e., we take the materials to be non-magnetic with permeability $\mu = \mu_0$. In fact, even if the materials were slightly magnetic this would not influence the quasi-electrostatic limit described below.

We consider solutions that are plane waves along x and constant along y, i.e., varying as $\exp(ik_x x + 0y)$. This reduces the system to a one dimensional problem in z, which we solve using the transfer matrix method. There are two possible polarizations here: transverse magnetic ($E_y = 0$, $B_x = 0$, $B_z = 0$) and transverse electric ($B_y = 0$, $E_x = 0$, $E_z = 0$). The tip couples essentially only to the transverse magnetic polarization, and plasmons only appear in this polarization. We define the reflection coefficient $r(\omega, k_x)$ for transverse magnetic waves as the ratio of E_z components of the up-decaying wave (positive $\operatorname{Im} k_z$) to the down-decaying wave (negative $\operatorname{Im} k_z$) at the top surface, with the condition that the wave is purely down-decaying at the bottom surface.

Bound propagating modes, such as the plasmon, appear in $r(\omega, k_x)$ as a simple pole in the complex k_x plane, with a residue that is primarily real-valued. Considering damped modes with $\operatorname{Re} k_x > 0$, then for ordinary dispersion (positive group velocity) this pole appears with $\operatorname{Im} k_x > 0$ and positive residue; for anomalous dispersion (negative group velocity), the pole has $\operatorname{Im} k_x < 0$ and negative residue.

For obtaining compact analytic equations such as (S2) it is helpful to take the quasi-electrostatic approximation, where the effects of electromagnetic induction are neglected. First we observe the relation $\vec{\nabla} \cdot (\vec{E} + \frac{1}{-i\omega\varepsilon_0}\vec{J}) = 0$, a consequence of taking the divergence of both sides of (S13). We then take the limit $c \to \infty$ (that is, $\mu, \mu_0 \to 0$, keeping ε intact) which implies $\vec{\nabla} \times (\vec{\nabla} \times \vec{E}) = 0$. This approximation is highly accurate when examining the near field waves (at very high k_x values past the light line, i.e., where $k_x \gg 1/\sqrt{\varepsilon\mu}$).

A. Nonlocal conductance

The effects of 2D nonlocality are easy to include for plane waves, since in this case the convolution is converted into a k_x-dependent conductivity, $\sigma(\omega, k_x) = \iint dx\, dy\, e^{ik_x x}\sigma_{\rm NL}(\omega, x, y)$. The quantity $\sigma(\omega, k_x)$ is known analytically at zero temperature, in the random phase approximation, allowing fast numerical evaluation.[S7] Although early works emphasized the influence of nonlocality,[S8,S9] we note a subtle point which is that for suspended graphene (in a vacuum dielectric) the primary nonlocal effect is the interband nonlocality, whereas for graphene in a dielectric and at low frequencies the dominant nonlocal effect is *intraband*.

We note that this intraband nonlocal effect is not particularly quantum nor special to graphene, but appears in all plasma physics. For example, in the classical quasi-electrostatic plasma (Langmuir wave), microscopic thermal effects give a similar nonlocal conductivity. For small k the classical nonlocal conductivity takes the form

$$\operatorname{Im} \sigma_{\rm classical}(\omega, k) \approx \omega^{-1} e^2 \frac{n}{m}[1 + 3k^2 v_{\rm th}^2/\omega^2],$$

where $v_{\rm th} = \sqrt{kT/m}$ is the thermal speed. The nonlocal effect can be seen as coming from pressurization effects, in the fluid plasma model. From the point of view of kinetic theory (e.g., Vlasov equation) it is a consequence of near-resonant particles that are travelling close to the wave phase speed ω/k, and is closely related to the Landau damping described below. In a bulk plasma, the plasma condition is $\omega_{\rm p} = \sigma/(i\varepsilon)$, where σ is the free-electron conductivity and ε is the background permittivity from vacuum and bound electrons ($\varepsilon = \varepsilon_0$ in a gas plasma). The nonlocal energy transport increases the imaginary part of conductivity and therefore increases the frequency of the plasma, as seen in the resulting Bohm-Gross dispersion, $\omega_{\rm p}^2 \approx e^2 \frac{n}{m\varepsilon_0} + 3k^2 v_{\rm th}^2$.

DOI: 10.1038/NMAT4169

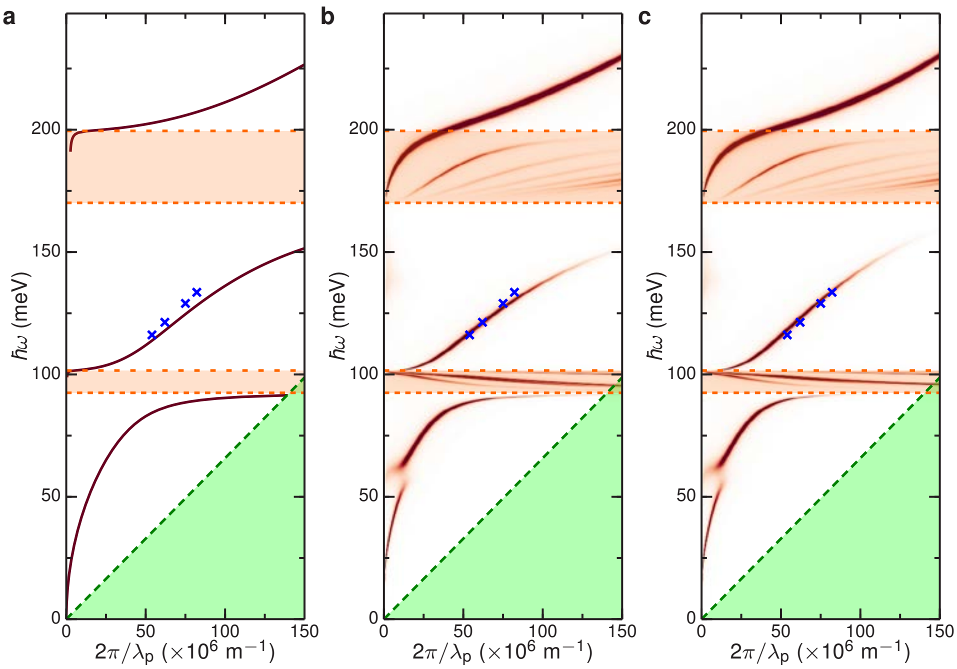

FIG. S5. (a) Drude model for graphene conductivity and simple effective permittivity for h-BN surrounding the graphene. (b) Drude model for graphene conductivity and thin film effects for h-BN surrounding the graphene. (c) Non local RPA for graphene conductivity and thin film effects for h-BN surrounding the graphene. Scattering time $\tau = 500$ fs, $n_s = 7.37 \times 10^{12}$ cm^{-2}.

For the degenerate graphene electron gas we have for small k and for $\omega \ll k_F v_F$,

$$\operatorname{Im} \sigma_{\text{graphene}}(\omega, k) \approx \omega^{-1} e^2 (2 v_F k_F/\hbar)[1 + \tfrac{3}{4} k^2 v_F^2/\omega^2],$$

where $k_F = \sqrt{\pi n_s}$ is the Fermi wavevector. Again, this nonlocality can be interpreted as a near-resonant effect of electrons whose speed (v_F) and direction are close to the wave phase speed ω/k. In a diagrammatic perturbation theory picture this corresponds to virtual intraband excitations. The nonlocality occurs regardless of whether k is comparable to k_F. The corresponding 2D plasma condition is $\omega_p = \frac{1}{2} q_p \sigma/(i\varepsilon)$, and so here too the nonlocal energy transport increases the imaginary part of conductivity and therefore increases the frequency of the plasma, or for fixed frequency it lowers q_p. In fact the full expression for σ_{graphene} contains a diverging conductivity as k approaches ω/v_F. This prevents the plasmon from having a lower phase velocity than v_F.

The striking difference between the classical thermal plasma and the graphene plasma is the effect of Landau damping. In the thermal plasma, the thermal distribution implies that some electrons have a velocity as high as the plasma phase velocity. They ‘surf’ the wave, accelerating to higher speeds and drawing energy out from the plasma. Thus, nonlocal effects in a thermal plasma are rarely observed because Landau damping turns on at the same time.

For plasmas in degenerate electron systems such as metals, low-temperature doped semiconductors, or doped graphene, it is possible to see the nonlocality without Landau damping, since the Landau damping only turns on after the plasmon wavevector q_p passes above ω/v_F. This is because the electrons have a sharp cutoff in their speed distribution, with few electrons travelling faster than v_F. Still, the nonlocality can be difficult to access: in semiconductors for example the Fermi velocity is very low, and so the required q_p to observe nonlocality are quite high. In graphene, the high Fermi velocity allows easier observation of nonlocality in plasmonics. This is most apparent with a high permittivity environment around the graphene, since for frequencies below $k_F v_F$ this drives the plasmon into the intraband nonlocal regime before it is affected by the interband absorption.

B. Dispersion relation comparison

In the simple Drude model the local response conductivity is given by the following expression:[S7]

$$\sigma(\omega, \tau, n_s) = \frac{2e^2 v_F}{h} \frac{\sqrt{\pi n_s}}{1/\tau - i\omega} \tag{S14}$$

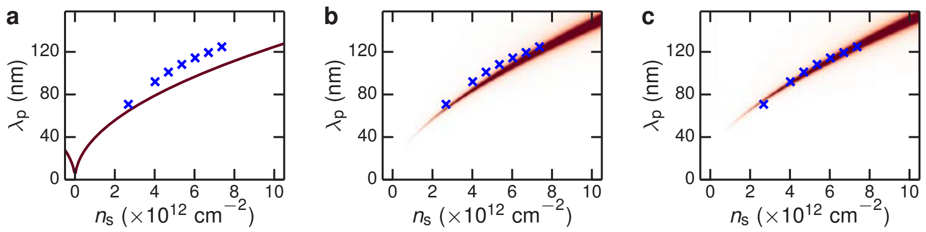

FIG. S6. (a) Drude model for graphene conductivity and simple effective permittivity for h-BN surrounding the graphene. (b) Drude model for graphene conductivity and thin film effects for h-BN surrounding the graphene. (c) Non local RPA for graphene conductivity and thin film effects for h-BN surrounding the graphene. Scattering time $\tau = 500$ fs, $n_s = 7.37 \times 10^{12}$ cm^{-2}.

In Fig. S5 and Fig. S6 we compare different dispersion models. In Fig. S5,S6a we show the result for a simple Drude conductivity for the graphene as in (S14) and the simple graphene plasmon relation[S10] $q_\mathrm{p} \approx 2\omega\epsilon(\omega)i/\sigma(\omega)$, where ω is the angular frequency of the excitation light, ϵ is the effective permittivity of the dielectric environment – see (S3) above – and σ is the local conductivity as defined in (S14). Both top and bottom h-BN layer are considered to be semi-infinite. Note that here no propagating phonon polariton modes exist inside the reststrahlen bands marked in orange.[S11]

A significantly improved fit is achieved by using the Drude model but including thin film effects of the 46 nm bottom and 7 nm top h-BN in Fig. S5,S6b. Due to the thin film effects propagating phonon polariton modes exist in the reststrahlen bands. These modes will be discussed elsewhere.

Including the nonlocal conductivity in Fig. S5,S6c we achieve an even better fit, especially for higher wavevectors where nonlocal effects start playing a more significant role. Also in the carrier density dependence in Fig. S6c a significantly improved fit is achieved as compared to not including nonlocal effects in Fig. S6b.

IV. H-BN PERMITTIVITY MODEL

Hexagonal boron nitride is an anisotropic material and so its permittivity ε is a tensor. Choosing x, y to be the in-plane directions and z to be the out-of-plane direction ("c-axis"), by symmetry the permittivity must be diagonal in a perfect h-BN crystal:

$$\varepsilon = \begin{pmatrix} \varepsilon_x & 0 & 0 \\ 0 & \varepsilon_y & 0 \\ 0 & 0 & \varepsilon_z \end{pmatrix}$$

with components $\varepsilon_x = \varepsilon_y \neq \varepsilon_z$.

As with many dielectric materials, the permittivity of h-BN is frequency dependent with resonances due to internal polar degrees of freedom,

$$\varepsilon_l(\omega) = \varepsilon_l(\infty) + s_{\mathrm{v},l}\frac{\omega_{\mathrm{v},l}^2}{\omega_{\mathrm{v},l}^2 - i\gamma_{\mathrm{v},l}\omega - \omega^2}, \qquad l = x, y, z. \tag{S15}$$

This degree of freedom is a polar lattice vibration, and its permittivity contribution involves real-valued constants $s_{\mathrm{v},l}$ (dimensionless coupling factor), $\omega_{\mathrm{v},l}$ (normal frequency of vibration), and $\gamma_{\mathrm{v},l}$ (amplitude decay rate). Observe that $s_{\mathrm{v},l}$ gives the DC permittivity contribution of the polar lattice distortion, so that in the case of a single vibrational mode as in (S15), one has $s_{\mathrm{v},l} = \varepsilon_{\mathrm{v},l}(0) - \varepsilon_{\mathrm{v},l}(\infty)$. We neglect nonlocal effects in the permittivity of h-BN as they should only appear once k is comparable to the reciprocal lattice vectors, a regime that is two orders of magnitude away from the experimental case.

In h-BN, it is theoretically expected that there are only three polar vibrational modes, one each for x, y, z.[S12] The out-of-plane vibration ($l = z$) has significantly different values of $s_{\mathrm{v},l}, \omega_{\mathrm{v},l}, \gamma_{\mathrm{v},l}$ compared to the in-plane modes ($l = x, y$). In practice, it is sometimes useful to include additional modes to fit the measured permittivity in disordered crystals,[S12] however here we consider ideal h-BN with one mode along each direction.

The permittivity (S15) completely characterizes the h-BN for optical studies at frequencies up to and including the mid-infrared. For bulk h-BN this permittivity is known to produce interesting behaviour of electromagnetic modes since $\mathrm{Re}\,\varepsilon_z \leq 0$ for one frequency band, and $\mathrm{Re}\,\varepsilon_x, \mathrm{Re}\,\varepsilon_y \leq 0$ in another frequency band. Both frequency bands contain:

Model	l	$\varepsilon_l(\infty)$	$s_{\mathrm{v},l}$	$\hbar\omega_{\mathrm{v},l}/\mathrm{meV}$	$\hbar\gamma_{\mathrm{v},l}/\mathrm{meV}$
Geick[S12]	x,y	4.95	1.868	169.5	3.6
		+	0.209	95.1	3.4
	z	4.10	0.530	97.1	1.0
		+	0.456	187.2	9.9
Cai[S15]	x,y	4.87	1.83	170.1	—
	z	2.95	0.61	92.5	—
Caldwell[S13]	x,y	4.9	2.001	168.6	0.87
	z	2.95	0.5262	94.2	0.25
Cai ”clean”	x,y	4.87	1.83	170.1	0.87
	z	2.95	0.61	92.5	0.25
Cai ”damaged”	x,y	4.87	1.83	170.1	6.5
	z	2.95	0.61	92.5	1.9

TABLE S1. Different permittivity models for hexagonal boron nitride. Note that the model of Geick et al. includes two vibrational modes for each direction.

- Transverse phonon polaritons near $\omega_{\mathrm{v},l}$. Near this frequency, the permittivity along l diverges to very large values ($\mathrm{Re}\,\varepsilon_l \sim \pm 100$). For light polarized along direction l, a strong peak in reflectivity (near 100%) is observed at this frequency.[S12,S13]
- Longitudinal phonon polaritons near $\omega_{\mathrm{L},l} = \omega_{\mathrm{v},l}\sqrt{\varepsilon_l(0)/\varepsilon_l(\infty)}$. At this frequency, ε_l passes close to 0. This allows a purely electric oscillation that is longitudinal, i.e., electric field parallel with the phase velocity.[S12]
- Hyperbolic phonon polaritons for $\omega_{\mathrm{v},l} < \omega < \omega_{\mathrm{L},l}$. In this frequency range, $\mathrm{Re}\,\varepsilon_l < 0$ in direction l, yet $\mathrm{Re}\,\varepsilon$ is positive along another direction. This results in a hyperboloidal constant-frequency surface of propagating modes in k-space, rather than the usual ellipsoid that appears for most frequencies. This hyperboloid extends to very high k (short wavelength) allowing propagating modes of very short wavelength.[S13] The group velocities of these confined modes are correspondingly low and are nearly perpendicular to their phase velocities.

These special frequency intervals are marked in Fig. 2 of the main text as orange bands: In the lower frequency band, $\mathrm{Re}\,\varepsilon_z < 0$ whereas $\mathrm{Re}\,\varepsilon_{x,y} < 0$ in the higher frequency band. For thin h-BN films, the effects of the transverse and longitudinal modes are somewhat diminished, yet the hyperbolic modes start to exhibit waveguiding[S14] and have been exploited to produce subwavelength resonant structures.[S13]

For the frequencies investigated in this study, $\mathrm{Re}\,\varepsilon$ is strictly positive and the most important aspect of (S15) is its anisotropy and its dielectric loss. The overall permittivity, including its high anisotropy (with $\varepsilon_x \approx 9$ and $\varepsilon_z \approx 2$ in the studied frequency range), is important for matching the measured plasmon wavelengths. Understanding the dielectric loss is a crucial part of understanding our plasmon damping. The following subsections describe the parameter sets we have considered and how dielectric loss may be modified in thin films of h-BN.

A. Parameters of h-BN permittivity and remarks

For reference, we list five permittivity parameter sets in Table S1. The first three models are from the existing literature[S12,S13,S15] and the last two are hybrids that we have constructed. In future precision studies it may be necessary to take into account (or exploit) the isotope effect of boron which could allow tuning of the resonance frequencies by 3%, or to remove the dielectric loss that originates from the isotope inhomogeneity of natural boron.

The Geick et al. study was performed on a large h-BN sample, and the authours found it necessary to include an additional vibrational mode for each direction, in order to fit their reflectance data. They attributed this necessity to the large degree of axis misalignment among the crystallites, which would mix together the x and z permittivities.[S12]

The Cai et al. model is a theoretical calculation for perfect h-BN,[S15] and results a plasmon dispersion that matches closely to the experiment. The values of Cai et al. were used successfully in modelling the propagating phonon polaritons in Ref. S14 (see the supplement of that paper). This study does not address the expected dielectric losses.

Caldwell et al. present their values in the supplementary material of Ref. S13. This permittivity was inferred from reflectance measurements on thin h-BN exfoliated films, originating from the same source as the h-BN films in our study. The parameters obtained here were very similar to the Cai et al. values.

Cai “clean” and Cai “damaged” in Table S1 take the theoretical modes of Ref. S15 and incorporates empirical losses based on Refs. S12 and S13. Cai “clean” uses the losses for pristine thick films of h-BN as measured in Ref. S13. In Cai “damaged” we amplify these losses to appear similar to those observed for thin (<200 nm) h-BN films in the same work. As no data were available on the losses of the z vibrational mode for thin films, we assume that they increase in proportion with the x,y losses, as described in the next section.

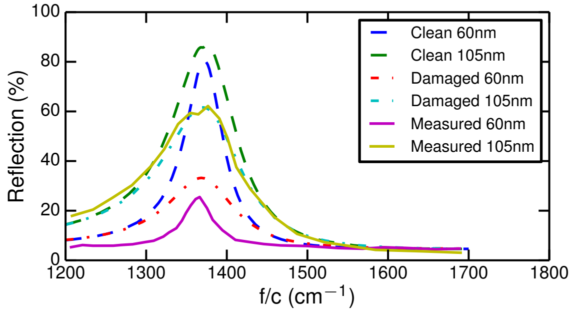

FIG. S7. Comparison between measured reflection from the supplement of Ref. S13 and simulated reflection for thin h-BN flakes. The clean h-BN uses an in-plane phonon linewidth of 0.87 meV from Ref. S13. The simulations of the damaged ones have an increased in-plane phonon linewidth of 3.7 meV in the case of the 105 nm thick h-BN and 6.5 meV for the 60 nm h-BN.

The model Cai "clean" was used to produced dispersion plots where we have matched the measured plasmon wavelength; its low level of dielectric loss aids the visibility of the modes. This last model, Cai "damaged", was used in our calculations of plasmon damping.

B. h-BN losses in thin films

Caldwell et al. have noticed in measuring thin h-BN films in Ref. S13 that the effective $\gamma_{v,l}$ seems to be larger than bulk. As a result our h-BN, especially the thin upper layer, may have higher losses. In Fig. S7 a comparison between clean and damaged h-BN from Ref. S13 is made. The simulations were done using the transfer matrix method taking into account the thickness of the flakes and the substrate and for the BaF$_2$ substrate permittivity values from Ref. S16 are used. It is clear that the resonance linewidth reported for thicker (>200 nm) h-BN flakes becomes broadened for thinner flakes and strongly depends on the thickness. Therefore the dielectric losses of graphene plasmons due to h-BN heavily depend on sample geometry and surrounding flake thickness. The values used in Fig. 4 in the main text are shown as Cai "damaged" in Table S1. The out-of plane phonon width was estimated by using a ratio of 3.5 between in-plane and out-of-plane width as reported in Refs. S13 and S12 for both very clean mono crystalline h-BN as well as for polycrystalline h-BN. Considering the strong thickness dependence of the phonon linewidth as seen in Fig. S7 these values are realistic.

[S1] Z. Fei, G. O. Andreev, W. Bao, L. M. Zhang, A. S McLeod, C. Wang, M. K. Stewart, Z. Zhao, G. Dominguez, M. Thiemens, M. M. Fogler, M. J. Tauber, A. H. Castro-Neto, C. N. Lau, F. Keilmann, and D. N. Basov, Nano Lett. **11**, 4701 (2011).
[S2] A. Y. Nikitin, T. Low, and L. Martin-Moreno, Phys. Rev. B **90**, 041407 (2014).
[S3] N. Ocelic, A. Huber, and R. Hillenbrand, Appl. Phys. Lett. **89**, 101124 (2006).
[S4] L. Zhang, X. Fu, and J. Yang, Commun. Theor. Phys. **61**, 751 (2014).
[S5] C. Craeye, P. Sobieski, L. Bliven, and A. Guissard, IEEE J. Ocean. Eng **24**, 323 (1999).
[S6] M. S. Wengrovitz, A. V. Oppenheim, and G. V. Frisk, J. Opt. Soc. Am. A **4**, 247 (1987).
[S7] F. H. L. Koppens, D. E. Chang, and F. J. García de Abajo, Nano Lett. **11**, 3370 (2011).
[S8] B. Wunsch, T. Stauber, F. Sols, and F. Guinea, New J. Phys. **8**, 318 (2006).
[S9] E. H. Hwang and S. Das Sarma, Phys. Rev. B **75**, 205418 (2007).
[S10] M. Jablan, H. Buljan, and M. Soljačić, Phys. Rev. B **80**, 245435 (2009).
[S11] A. Principi, M. Carrega, M. B. Lundeberg, A. Woessner, F. H. L. Koppens, G. Vignale, and M. Polini, Phys. Rev. B **90**, 165408 (2014).
[S12] R. Geick, C. Perry, and G. Rupprecht, Phys. Rev. **146**, 543 (1966).
[S13] J. D. Caldwell, A. V. Kretinin, Y. Chen, V. Giannini, M. M. Fogler, Y. Francescato, C. T. Ellis, J. G. Tischler, C. R. Woods, A. J. Giles, M. Hong, K. Watanabe, T. Taniguchi, S. A. Maier, and K. S. Novoselov, Nat. Commun. **5**, 5221 (2014).
[S14] S. Dai, Z. Fei, Q. Ma, A. S. Rodin, M. Wagner, A. S. McLeod, M. K. Liu, W. Gannett, W. Regan, K. Watanabe, T. Taniguchi, M. Thiemens, G. Dominguez, A. H. Castro Neto, A. Zettl, F. Keilmann, P. Jarillo-Herrero, M. M. Fogler, and D. N. Basov, Science **343**, 1125 (2014).
[S15] Y. Cai, L. Zhang, Q. Zeng, L. Cheng, and Y. Xu, Solid State Commun. **141**, 262 (2007).
[S16] E. D. Palik, *Handbook of Optical Constants of Solids* (Elsevier, New York, 1997).