Fast modulation of terahertz quantum cascade lasers using graphene loaded plasmonic antennas

Riccardo Degl'Innocenti*, David S. Jessop1, Christian W. O. Sol1, Long Xiao1-2, Stephen J. Kindness3, Hungyen Lin3,4, J. Axel Zeitler3, Philipp Braeuninger-Weimer2, Stephan Hofmann2, Yuan Ren1, Varun S. Kamboj1, Jonathan P. Griffiths1, Harvey E. Beere1 and David A. Ritchie1

1Cavendish Laboratory, University of Cambridge, J. J. Thomson Avenue, Cambridge CB3 0HE, United Kingdom
2Department of Engineering, University of Cambridge, J. J. Thomson Avenue, Cambridge CB3 0FA, United Kingdom
3Department of Chemical Engineering and Biotechnology, University of Cambridge, Pembroke Street, Cambridge CB2 3RA, United Kingdom
4Department of Engineering, Lancaster University, Lancaster LA1 4YW, United Kingdom

Keywords: Terahertz optical modulator, graphene, plasmonic antennas, quantum cascade lasers

Abstract: We report the fast amplitude modulation of a quantum cascade laser emitting in single mode operation in the terahertz frequency range by employing compact, integrated devices based on the interplay between plasmonic antenna arrays and monolayer graphene. By acting on the carrier concentration of graphene the optical response of these plasmonic resonances was modified. The modulators characteristics have been studied by using both time domain spectroscopic laser systems, yielding the broad frequency response of these resonant arrays, and quantum cascade lasers, providing us with a narrow and stable laser source, a mandatory prerequisite for the determination of the modulation speed of these devices. The measured modulation speed exhibits a cut-off frequency of 5.5 MHz ± 1.1 MHz. These results represent the first step toward the realization of fast integrated circuitry for communications in the terahertz frequency range.

Recently, an impressive advancement has been achieved by terahertz (THz) solid-state sources in terms of performance with the quantum cascade laser (QCL) 1,2 and by time-domain spectroscopic (TDS) systems, in terms of costs, compactness and bandwidth. However, the lack of suitable optoelectronic devices, both passive and active, has so far hindered technology in this spectral region from reaching its full potential in many demanding sectors, such as spectroscopy or communications. The realization of a tuneable external THz amplitude/phase/frequency modulator would allow for fast reconfigurable spectroscopic systems, e.g. to be implemented for amplitude, frequency or phase stabilization.3 Terahertz communications represent a fast evolving research area. Wireless communications always strive for broader bandwidths and the THz range represents a non-allocated frequency range. Further to this, because of the high carrier frequency, THz radiation has the potential to yield high transmission rates, as high as 100 Gbit s⁻¹, as recently demonstrated in the sub-THz frequency range.4 The basic optical building blocks for a complex THz communications system consist in components such as amplitude/phase...
modulators, beam steering devices and phase arrays, capable of fast reconfiguration rates. Research in THz optoelectronics based on graphene is fast gaining popularity because of their fundamental importance at these frequencies where natural occurring materials have a poor electromagnetic response, and because of the promise of the implementation of the unique graphene properties into innovative and more performant devices. Plasmonic and metamaterial devices have already established themselves as a versatile tool for the creation of efficient and compact devices in the mid-infrared/terahertz spectral range for spectroscopic applications, modulation, emission and detection and more fundamental physical studies. These devices, which exploit the extreme subwavelength light concentration of the surface polaritons to boost the light-matter interaction, are normally engineered as metallic features over a positive dielectric constant substrate. The amplitude modulation of THz radiation with hybrid metamaterial-graphene architectures by using ultra-low bias or high magnetic fields was recently investigated. Further, a graphene-based modulator integrated on a QCL was demonstrated with remarkable performance. However, the realization of an external fast modulator capable of efficiently modulating terahertz radiation from a QCL has not been demonstrated yet. Such a device in fact, would uniquely address the need of independently operating source and modulator and provide a fundamental and versatile tool in light of spectroscopic experimental arrangements, on top of being compatible with multiple sources. Fast
external terahertz modulators have been realized by implementing two dimensional electron gas (2DEG) with a modulation speed up to 2 MHz\textsuperscript{29} or incorporating high electron mobility transistors in a metamaterial device\textsuperscript{30} with a modulation cut-off frequency of about 10 MHz. Lately, an increasing number of experiments have been focused on the realization of graphene-based terahertz modulators because of its remarkable properties such as the wide carrier concentration modulation range. The modulation speed has significantly increased from the 20 kHz of the first devices,\textsuperscript{14} to the 12.6 MHz achieved in a complex metamaterial graphene integrated modulator.\textsuperscript{15} In our approach comparable modulation speeds were achieved but in a simpler configuration. Further to this, the system has been tested also with a THz quantum cascade laser thus paving the way to the implementation of this powerful and spectrally narrow source in spectroscopic and communication applications.

In this manuscript we aim to investigate the reconfiguration speed of amplitude modulators operating in the THz frequency range based on the interplay of plasmonic antenna arrays and graphene. These results represent an important milestone toward the realization of a fast, integrated platform for THz communications. At the same time the present work opens the way to new efficient routes aimed at the amplitude/frequency modulation of THz sources, thus finding applications in diverse research areas, e.g. spectroscopy.

![Figure 2](image.png)

Figure 2. The black curves report the typical source/drain resistance plots of the arrays similar to what is shown in Fig. 1, having an $a$ value of 26 $\mu$m and 24 $\mu$m, respectively. These data, as indicated by the black arrow, are reported in the left-bottom layer axis. The red curve represents the gate leakage for the 26 $\mu$m array, with the corresponding curve for the 24 $\mu$m being almost identical. These values, as indicated by the red arrow, are reported in the top-right layer axis.
In order to achieve an active optical device, plasmonic antenna arrays were fabricated on top of a transferred graphene monolayer with a procedure similar to Refs. 25,26. The coupling of these resonant elements with graphene allows the modification of the whole of the device's optical response, by acting on the graphene carrier concentration. In analogy to electronic circuits, the gating of graphene results in a modified load impedance. Following the electrical characterization of the graphene in particular the determination of the position of its Dirac point, the samples were tested with both a broadband THz-TDS system and a narrow frequency QCL. Whilst the first approach yields the broad frequency response of the device for different values of the graphene conductivity, the implementation of a powerful and stable single frequency QCL source, allows the determination of the modulation speed of the device, which is found to be in the MHz range.

Results and Discussion
Graphene growth and transfer is described in the Methods section together with the sample fabrication. Figure 1 a) and b) show scanning electron microscopy (SEM) images of an array of the final device. The final device integrates four different arrays on the same sample, each composed of identically sized plasmonic antennas, as shown in Fig. 1 c), each characterized by a different total length $L$ equal to 42 $\mu$m, 46 $\mu$m, 50 $\mu$m and 54 $\mu$m. The antenna size unit was scaled in the different arrays to have a frequency response approximately covering the range between 2 THz and 3 THz, where both THz–TDS and THz QCLs overlap. The device is then mounted on a chip carrier and wire bonded for electrical testing. The pitch, the distance between neighbouring antennas, has been kept fixed to 1.55 times the total length $L$ along the antenna direction and to 16 $\mu$m along the perpendicular one. In order to excite the plasmonic resonances of these antennas the length $a$, as shown in Fig. 1 b), should correspond to $\lambda_{sp}/2 \cdot n_{eff}$ where $\lambda_{sp}$ is the vacuum wavelength of the resonant radiation, $n_{eff}$ is the effective refractive index of the plasmonic mode supported by the metal feature and $m$ is an odd integer number. A more detailed simulation, however, is required to better investigate and disentangle the optical response of the graphene loaded plasmonic antennas and the response of the graphene alone. The Dirac points of the graphene layers were determined for each sample using two Keithley source/measure units (Model 2400), the first of which provided a constant current of 5 $\mu$A to the source/drain whilst the second provided a variable bias to the back gate contact (-40 V to 120 V). A typical example of the electrical characterization is shown in Fig. 2 for the samples with $a = 24$ $\mu$m and $a = 26$
The leakage current toward the gate was below 10 nA for all samples, as also reported in Fig. 2. The Dirac points of all the samples were found between 90 V and 140 V, which is compatible with p-doped chemical vapour deposition (CVD) grown graphene. The source/drain resistances in all the samples were observed to be between 0.4 kΩ and 7 kΩ. These resistance values include also the contributions of the metallic antennas and contacts. Since the metallic features in the arrays introduce periodic parallel paths with different impedance for the current flow, the resistances measured cannot be attributed solely to the graphene carrier concentration and cannot be correlated directly to the graphene resistivity. However, these measurements provide the voltage characteristics which are needed for the optical measurements and yield a range in the device conductivity which, even though overestimated, can be implemented in the finite element simulations.

The commercial software Comsol Multiphysics(R) v. 5.0 was used to simulate the unit cell of our device. The graphene monolayer was modeled in the finite element based software by expanding its thickness to an artificial value of 15 nm. The dc conductivities were inferred from the electrical characterization of the different arrays. In the framework of the Drude model, these values were used to model the complex conductivity σ of graphene at different gate biases with a procedure similar to that reported in Ref. 24. The conductivity σ is then related to the dc conductivity σ₀ through the relation 1/σ = (1 - i · ωτ)/σ₀ where the scattering time τ was assumed to be 15 fs¹⁵ and ω is the angular frequency. Fig. 3 a) shows, for a xy slice of the unit cell having a
length of $1.55 \cdot L$ and a width of 16 $\mu$m, the normalized electric field at resonance for the array having $a = 26$ $\mu$m in the case of no damping, e.g. zero graphene conductivity. As can be seen the mode intensity has a maximum value at the ends of the antennas, in agreement with the condition $m \cdot \lambda_{sp}/2 \cdot n_{eff}$ with $m=1$. Further simulations are reported in the supplementary information in Fig. 1 SI. The device was tested with a commercially available THz pulsed imaging system Imaga 2000 from TeraView (Cambridge, UK), based on biased photoconductive antennas and yielding the broad-band frequency response of the antennas arrays. The experimental procedure follows closely to that previously reported in Ref. 3. The reflected electric field was recorded at different spatial coordinates with a step size of 100 $\mu$m and keeping the sample in the focus of the system, which has a resolution of $\sim 200$ $\mu$m. The reflected THz radiation from the different arrays was
monitored for different back-gate biases with the polarization fixed along the longitudinal direction of the plasmonic antennas. The peak electric field reflected from the sample at different positions yields thus a THz map, similar to the one shown in Fig. 3 b). From these acquired data it is possible to extrapolate the reflected intensity from these arrays, showing a broad peak in correspondence of the plasmonic resonance. The experimental results for the array with $a = 26 \, \mu m$ are shown in Fig. 3 c) for different gate bias voltages. By increasing the graphene carrier

Figure 6. Experimental scheme implemented for the determination of the modulation speed of these devices. A TTL voltage wave is applied to the QCL with low frequency $f_{QCL}$. The fast modulation $f_{MOD}$ is applied to the device superimposed onto a DC high-voltage bias which determines the working point. The reflection signal coming from the sample is detected by a Golay cell and is then demodulated by a lock-in amplifier having the frequency $f_{QCL}$ as reference. The total signal detected is $P_{1\text{meas}}$ for low frequencies a), is then reduced to $P_{2\text{meas}}$ at high frequencies b) and, finally, when the device cannot follow the modulation c), the reflectivity is given by $P_{3\text{meas}}$. 


concentration/conductivity, the resonant plasmonic mode is more attenuated. As expected, for higher values of graphene film conductivity the resonances are red-shifted. Depending on the voltage bias applied, the resonance peaks at 2.1-2.3 THz. The results are in close agreement with the simulations presented in the supplementary information. In particular, it is possible to observe a distinct resonance shift of about 200 GHz between the resonance for 0 V and 120 V, as predicted by the simplistic model. This is attributed to the combined effect of the increased reflectivity of graphene for high conductivity values and reduced plasmonic resonance strength.

In order to test the modulation speed of the device, a more powerful, stable and narrow frequency THz source is required. A bound to continuum quantum cascade laser was used for all these measurements. The QCL emits at 2.05 THz in single mode around the maximum current density, which is a fundamental feature required in order to test the speed performance of these amplitude modulators. In fact, single mode operation is required in order to avoid laser mode hopping which would translate in a laser amplitude modulation. Correspondingly, this would affect the data interpretation because of the convolution of different modes, each characterized by a distinct frequency and power, with the plasmonic resonances. The emission frequency overlaps significantly with the plasmonic resonances for \( a = 26 \, \mu m \) and partly also for the \( a = 24 \, \mu m \) plasmonic antenna arrays. A general schematic drawing of the experimental set-up is shown in Fig. 4. QCL radiation is collimated with a parabolic mirror and after a beam splitter is partly focused onto the device with another parabolic mirror with a focal length of 2.5 cm and partly
focused on a Golay cell detector in order to monitor the stability of the power source. Another parabolic mirror focuses the light reflected from the sample to a second Golay cell, which was used to measure the reflectivity modulation. The first experiment aimed to demonstrate the modulation of the THz light emitted from the QCL. A schematic of the experimental set-up is shown in Fig. 4. In this configuration the QCL laser was operating in continuous pulse mode, with a 30% duty cycle and 100 kHz repetition rate at maximum power current density. The optical modulator source and drain contacts were grounded and a voltage bias was applied to the gate. The DC bias applied to the gate spanned from -40 V to 120 V which set the working point, and an AC slow modulation (7 Hz frequency), consisting in a 0-10 V square-wave modulation added to this DC value. The reflected power was then focused to the Golay cell and demodulated by a lock-in amplifier having the AC modulation frequency as reference. A pin-hole with a 1 mm diameter was positioned in front of the graphene modulator to ensure that the illuminated area was at most the same as the device array area and that all the light reflected was coming from the array of interest. A second Golay cell was implemented to monitor the stability of the QCL emitted power through the measurements and rule out cases of mode hopping. A typical example of the measurements is reported in Fig. 5 for the $a = 26 \mu m$ and $a = 24 \mu m$ arrays corresponding to red and black symbol points, respectively. The total peak power impinging to the array was calculated to be 65 $\mu W$ for the first array and 157 $\mu W$ for the second one. The modulation depth for the first array increases with increasing gate voltage, peaking around 90 V reaching a maximum of ~ 8.6% and then starts reducing again. This trend was expected since by changing the working point the maximum conductivity modulation is reached around 100 V. The differential resistivity of the device is reported in the inset of Fig. 5 for clarity and is consistent with the modulation depth measurements. Measurements performed with the $a = 24 \mu m$ array revealed a similar trend but with a reduced modulation depth of only 1.6%. Even though the extent of the sheet resistivity modulation of graphene for this array was larger, as shown in Fig. 2, the final QCL modulation depth was lower because of the reduced overlap of the plasmonic resonance with the laser frequency in good agreement with the simulations performed with Comsol Multiphysics and reported in Fig. 1 SI in the supplementary information. Finally, the sample was rotated by $90^\circ$ such that the polarization was not exciting the plasmonic resonances in order to appreciate the graphene stand-alone modulation depth. The modulation depth could not be observed within the noise of the measurements. Another device was realized without any metal plasmonic antennas on top in order to perform a comparison with a pure graphene modulator, ruling out any effect introduced by further fabrication steps. The device exhibited maximum source-drain resistance an order of magnitude higher than in the other samples and Dirac point around 70 V but did not show any appreciable modulation depth above the noise floor. This demonstrates that the modulation of the conductivity in graphene alone does not yield any appreciable power modulation depth in these experiments. Conversely, when boosted by plasmonic resonances, the contribution to the total modulation given by the graphene Drude response is not negligible, coherently with Fig. 5.
Testing the device modulation speed, which is expected to be above hundreds of kHz, with a Golay cell detector, which is usually sensitive only to frequencies lower than 100 Hz, was accomplished with a different detection scheme, shown in Fig. 6. Our approach is similar to what has been developed in parallel and recently reported in Ref. 28. The laser was modulated with a frequency $f_{QCL}$ of 22 Hz in continuous pulse mode, 30% duty cycle and at the current density corresponding to maximum output power. A modulation frequency $f_{MOD}$ was applied to the device with modulation depth of 10 V by using a 30 MHz function generator from Stanford Research system model DS345. A lock-in amplifier, having $f_{QCL}$ as a reference signal, demodulated the reflection coming from the device. For sufficiently low $f_{MOD}$ frequencies, as shown in Fig. 6 a), the power detected by the Golay cell was $P_{1\text{meas}}$, which is the sum of the QCL reflected power $P_{QCL}$ plus half of the superimposed power modulation given by the device, $P_{MOD}$. Since $f_{MOD}$ is too fast to be detected by the Golay cell, the contribution to the total reflectivity will be the average value of the fast modulated reflectivity. At higher frequencies, as shown in Fig. 6 b), the device is unable to follow the modulation and therefore the total power seen by the lock-in was reduced to $P_{2\text{meas}}$ with $P_{1\text{meas}} > P_{2\text{meas}}$ since $P_{MOD}$ is reduced. Finally, Fig. 6 c), the modulation signal was too fast to be followed by the modulator and the total reflected power is $P_{3\text{meas}}$ which is simply given by $P_{QCL}$. The amplitude modulator behaves as a low pass filter being limited by both its RC time constant and the electronics driving it, with the latter considered negligible. The modulation depth measurements, acquired with the device working point set to 100 V to detect the maximum modulation depth, are reported in Fig. 7.
measurements have been acquired at different gate voltages of 70 V and 100 V. In order to rule out the possibility of a drift of the optical system as possible cause of the reduction in the modulation depth observed at high frequencies, the set of measurements shown in Fig. 7 was acquired chronologically from 500 kHz to the highest frequency, 20 MHz, and then from 300 kHz to the lowest one, 100 Hz. Similar acquisitions have been recorded after completely realigning the optical system and are shown in the supplementary information presenting consistent values. The continuous line in Fig. 7 represents the best fit using the standard formula for the capacitive low-pass filter having a transfer function of \(1/(1+(f/f_c)^2)^{1/2}\) where \(f\) is the modulation frequency and \(f_c\) the lowest 3 dB cut-off frequency of the filter. The cut-off frequency \(f_c\) was left as free fitting parameter. The results from the fitting procedure are reported in Table 1. Even though in principle it should be possible to appreciate the differences in modulation depths and in cut-off frequencies at different gate biases, the limitations of these measurements due to their accuracy and reproducibility prevent such comparison. The highest cut-off frequency \(f_c\) for this modulator was found to be 5.5 MHz ± 1.1 MHz, as it is shown in Fig. 7. The procedure schematically reproduced in Fig. 6 represents a simplified picture. In fact, because the Golay cell is sensitive only to slow frequency power differences, the fast reflectivity modulation superimposed on top of the reflectivity given by the sample is only seen as an average value. Since the transfer function between the gate voltage applied to the sample and the reflectivity detected by the Golay cell is nonlinear, the average values seen by the Golay at low and high frequencies differs significantly and allowed the determination of the cut-off frequency. Because the transfer function is unknown and the development of an equivalent circuital model capable of describing the system is beyond the scope of this manuscript, a precise quantification of the modulation depths cannot be inferred from the fast modulation speed measurements. The limit in the modulation depth, shown in Fig. 5, is attributed mainly to the finite AC voltage applied. The modulation depth can be increased by increasing the AC voltage applied or by optimizing the device architecture. Atmospheric adsorbants have been found to p-dope graphene films and cause a shift in the Dirac point. This can be prevented by encapsulating the graphene film, e.g. with Al₂O₃, such that the Dirac point can be achieved with low biases, thus allowing the exploitation of the full range of conductivity available. This device architecture was optimized for high reconfiguration speeds, which is the key result of this work. Therefore, the highest resistance seen by the low pass filter is given by the graphene layer. The choice of having the graphene exposed was made in order to achieve the lowest resistance, thus allowing higher cut-off frequencies, at the expense of the modulation depth. An estimation of the theoretical performance of this device can be obtained in analogy with a low-pass frequency filter by developing a model similar to that reported in Refs. 17, 28, where the parasitic effects produced by the contacts (parallel parasitic capacitance) have been taken into account and the resistance of the Si-substrate considered negligible. The capacitance \(C_{\text{graph}}\) of the graphene array area is given by the standard formula for the parallel plate capacitor \(C_{\text{graph}} = \varepsilon_0 \varepsilon_r A/d\), where \(\varepsilon_0\) is the vacuum permittivity, \(\varepsilon_r\) the relative dielectric constant of SiO₂, \(A\) the total graphene area and \(d\) the thickness of SiO₂ dielectric layer. Assuming a thickness \(d\) of 300 nm and a relative dielectric
constant of 3.8, the capacitance is estimated to be \( \sim 126 \text{ pF} \). From the model, a 3 dB cut-off frequency of 5.5 MHz ± 1.1 MHz should correspond to an effective graphene resistance \( R \) which varies between 520 \( \Omega \) and 920 \( \Omega \), in good agreement with the measured value for the \( a = 26 \mu m \) antenna array source/drain resistance, whose plot at different gate voltages is reported in Fig. 2 and which takes into account also the graphene/ contact resistances. The model is a first approximation of the complex transfer function between the voltage applied to the modulator, and its optical responsivity, but is useful in light of further optimizations.

<table>
<thead>
<tr>
<th>Gate Voltage (V)</th>
<th>Cut-off frequency (MHz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>70</td>
<td>3.9±1.1</td>
</tr>
<tr>
<td>100</td>
<td>5.5±1.1</td>
</tr>
<tr>
<td>100*</td>
<td>4.1±1.3</td>
</tr>
</tbody>
</table>

Table 1. Results of the fitting procedure for three different measurements for DC gate voltages of 70 V and 100 V. The measurements have been performed after a complete realignment of the system each time. In order to rule out possible drifts, the acquisition at 70 V has been performed from the lowest frequency to the highest one. The set of measurements at 100 V, also shown in Fig. 2, was recorded in two acquisition steps: the first one from 500 kHz to the highest frequency and the second from 300 kHz to the lowest one. Finally, the measurement at 100 V* was acquired from the highest frequency to the lowest one. The plot at 70 V and at 100 V* are reported in the supplementary information.

The maximum modulation speed can be increased by reducing the size of the arrays. A reduction of the sample to a size comparable to the wavelength, e.g. \( 200 \times 200 \mu m^2 \), a minimization of the contact pads area in order to reduce the parasitic capacitance, and a further optimization of the other resistances, should yield a higher cut-off frequency, beyond 100 MHz. The modulation depth as well could be improved by either encapsulating the graphene film, or implementing multiple independently tunable monolayer graphene stacks, with a double gate architecture, or denser packaging of the antennas in a sort of “superdipole” arrangement.

In conclusion, we have presented the realization of a compact, integrated optoelectronic device based on plasmonic antennas and graphene for the fast modulation of terahertz frequencies. The device has been modeled with both finite element methods and basic electronic circuitual formulas showing a remarkable agreement with the experimental results acquired with broadband TDS systems and single frequency quantum cascade laser emitting around 2 THz. A novel experimental technique has been developed in order to retrieve the cut-off frequency of these devices even at MHz modulation speed using slow-responding detectors, such as a Golay cell. This represents a significant progress compared to other more conventional techniques which allows a direct determination of the cut-off frequency but typically required cryogenic operating temperature, e.g. superconducting bolometers. The device tested present a cut-off frequency of 5.5 MHz ± 1.1 MHz, reproducible over several measurements. These results pave
the way to the realization of a fast integrated optoelectronic class of devices for Terahertz communications.

**Methods**

**Graphene growth and transfer.** Sample fabrication starts from the transfer of a monolayer graphene to a SiO$_2$/Si substrate (300 nm/525 µm thick) grown by chemical vapour deposition (CVD). The Si substrate was Boron-doped (~ 20 Ω· cm) to allow gating of the device. The CVD graphene growth was performed on Cu foils (25 µm thick, Alfa Aesar purity 99.98%) using CH$_4$ as the precursor and PMMA (poly-methyl-methacrylate) as a support for the transfer followed by FeCl$_3$ chemical etching to remove the Cu.

**Sample fabrication.** Four large graphene areas were defined with size of 1.2×1.2 mm$^2$ each, by using optical lithography and oxygen plasma etching. A second step of optical lithography, Ti/Au (10/120 nm) thermal evaporation and lift-off was used to fabricate the metallic contacts, thus realizing source and drain pads. The definition of the antennas was achieved by double layer electron beam lithography. The graphene underneath the exposed patterns was removed with oxygen dry etching. The final definition of the antennas has been achieved by a second step of metallic (Ti/Au, 10/80 nm) thermal evaporation and lift off. The sample is then cleaved, mounted on a chip carrier and wire-bonded in order to perform the electrical characterization and the biasing.

**Author Information**

Corresponding author: *Email: rd448@cam.ac.uk.*

**Acknowledgements**

RD, YR and HEB acknowledge financial support from the Engineering and Physical Sciences Research Council (Grant No. EP/J017671/1, Coherent Terahertz Systems). HL and JAZ would like to acknowledge the financial support from UK EPSRC Research Grant EP/L019922/1.

**Additional Data:** Additional data sets related to this publication are available from the Cambridge University data repository at XXXX.

**Supporting Information Available:** Simulation performed with the finite element commercial software Comsol Multiphysics; simulated terahertz reflection response of the device arrays and their frequency/amplitude response at different conductivity level of the graphene film. Frequency response of the modulator at different gate bias working points.

**References**

Fast modulation of terahertz quantum cascade lasers using graphene loaded plasmonic antennas