Generation of photovoltage in graphene on a femtosecond timescale through efficient carrier heating

K. J. Tielrooij[1]*, L. Piatkowski[1†], M. Massicotte[1†], A. Woessner[1], Q. Ma[2], Y. Lee[3], K. S. Myhro[3], C. N. Lau[3], P. Jarillo-Herrero[2], N. F. van Hulst[1,4]* and F. H. L. Koppens[1]*

Graphene is a promising material for ultrafast and broadband photodetection. Earlier studies have addressed the general operation of graphene-based photothermoelectric devices and the switching speed, which is limited by the charge carrier cooling time, on the order of picoseconds. However, the generation of the photovoltage could occur at a much faster timescale, as it is associated with the carrier heating time. Here, we measure the photovoltage generation time and find it to be faster than 50 fs. As a proof-of-principle application of this ultrafast photodetector, we use graphene to directly measure, electrically, the pulse duration of a sub-50 fs laser pulse. The observation that carrier heating is ultrafast suggests that energy from absorbed photons can be efficiently transferred to carrier heat. To study this, we examine the spectral response and find a constant spectral responsivity of between 500 and 1,500 nm. This is consistent with efficient electron heating. These results are promising for ultrafast femtosecond and broadband photodetector applications.

Photovoltage generation through the photothermoelectric (PTE) effect occurs when light is focused at the interface of monolayer and bilayer graphene, or at the interface between regions of graphene with different Fermi energies E_F (refs 1–6). In such graphene PTE devices—which operate over a large spectral range[7,8] that extends even into the far-infrared[9]—local heating of electrons by absorbed light, in combination with a difference in Seebeck coefficients between the two regions, gives rise to a PTE voltage $V_{PTE} = (S_2 - S_1)(T_{el} - T_0)$. Here, S_1 and S_2 are the Seebeck coefficients of regions 1 and 2, respectively, T_{el} is the hot electron temperature after photoexcitation and electron heating, and T_0 is the temperature of the electrode heat sinks. The performance of PTE graphene devices is intimately connected to the dynamics of the photoexcited electrons and holes, which have mainly been studied in graphene samples through ultrafast optical pump–probe measurements[10–17]. As shown in Fig. 1a, the dynamics start with (i) photoexcitation and electron–hole pair generation, followed by (ii) electron heating through carrier–carrier scattering, in competition with lattice heating, both of which take place on a sub-100 fs timescale, and finally (iii) electron cooling by thermal equilibration with the lattice, which takes place on a picosecond timescale. The effect of the picosecond cooling step (iii) on the switching speed of graphene devices has been studied using time-resolved photovoltage scanning experiments with ~200 fs time resolution[18–20]. These studies showed that the picosecond electron cooling time limits the intrinsic photo-switching rate of these devices to a few hundred gigahertz, because faster switching would reduce the switching contrast, as the system does not have time to return to the ground state. Indeed, gigahertz switching speeds have been demonstrated in graphene-based devices[21–25].

However, the most crucial aspects of the PTE response are captured by the heating dynamics, as electron heating corresponds to photovoltage generation. Additionally, these dynamics determine the ultimate intrinsic carrier heating efficiency and the resulting spectral response. Here, we measure the photovoltage generation time with an unprecedented time resolution and assess its effect on the heating efficiency through spectral responsivity measurements. In an ideal PTE detector, all the absorbed photon energy is transferred rapidly to electron heat (before energy is lost through other channels). In this case, doubling the photon energy would lead to doubling of the photovoltage (Fig. 1b) and would result in a flat spectral responsivity $R_{PC} = I_{PC}/P_{exc}$, where I_{PC} is the generated photocurrent and P_{exc} is the excitation power. In strong contrast, the spectral response of conventional semiconductor-based detectors is not flat at all, because it is determined by the bandgap; photons with an energy below the bandgap are not absorbed, that is, $R_{PC} = 0$, and the excess energy above the bandgap typically does not lead to an additional photoresponse, that is, a decreasing R_{PC} with photon energy[26].

Photovoltage generation on a femtosecond timescale

To capture the timescale of photovoltage generation we performed time-resolved photovoltage measurements with the highest time resolution obtained to date (~30 fs). This was achieved by using a broadband Ti:sapphire 85 MHz oscillator (centre frequency of 800 nm, bandwidth >100 nm) that creates <20 fs pulses, and a pulse shaper that corrects for any dispersion (and thus pulse stretching) that the pulses pick up on their way from the laser to the device (Fig. 1c; see Supplementary Section II for details). The concept of the experiment is illustrated in Fig. 1d. We use a pair of ultrashort laser pulses, where the two pulses are separated by a controllable delay time t and focused onto a single spot on the device. The two pulses contain different spectral components to suppress coherent artefacts. We then record the photocurrent, averaged over a large

[1]ICFO – Institut de Ciències Fotòniques, Mediterranean Technology Park, Castelldefels (Barcelona) 08860, Spain. [2]Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [3]Department of Physics and Astronomy, University of California, Riverside, California 92521, USA. [4]ICREA – Institució Catalana de Recerca i Estudis Avançats, Barcelona 08010, Spain; [†]These authors contributed equally to this work. *e-mail: frank.koppens@icfo.eu; niek.vanhulst@icfo.eu; klaas-jan.tielrooij@icfo.eu
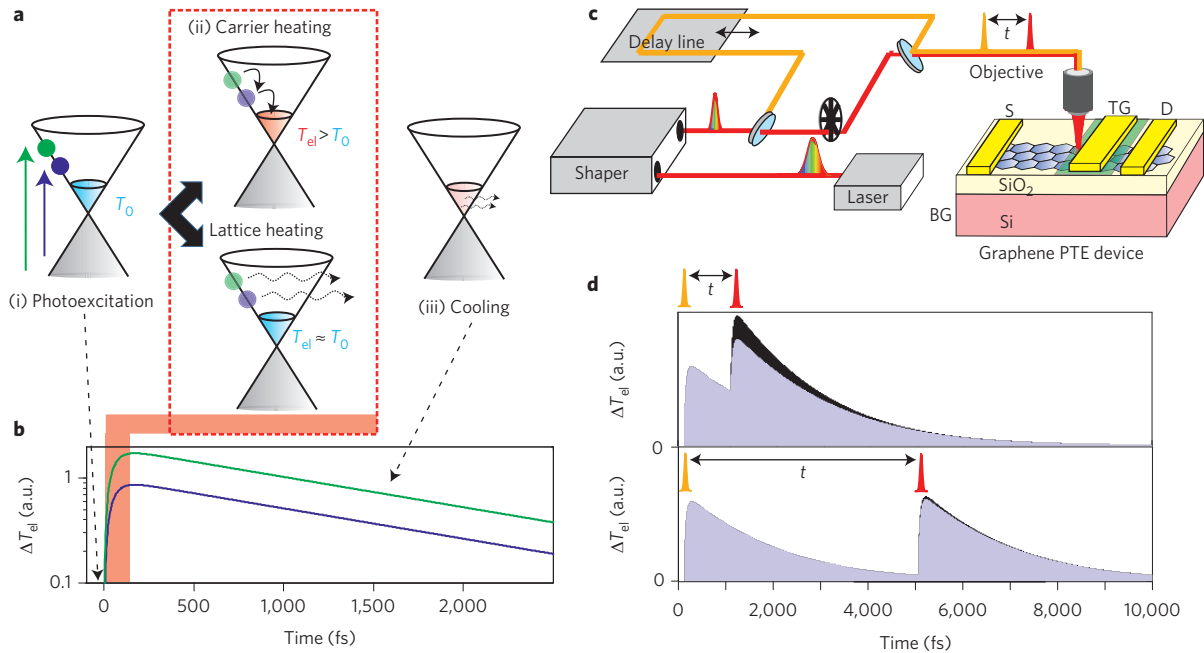
437

Figure 1 | Hot electron dynamics and their experimental extraction. a, Schematic of the electron dynamics in graphene after photoexcitation with two different photon energies (green and blue arrows). After photoexcitation and electron–hole pair generation, electron heating and lattice heating take place. The former occurs through carrier–carrier scattering and leads to hot electrons[11,13,14,16,20,30], which drive a PTE current[2–4]. The latter occurs through phonon emission and leads to the generation of a much smaller photovoltage, because the heat capacity of the phonon bath is orders of magnitude larger than that of the electron bath[3]. **b**, The hot electrons lead to a local photovoltage through the PTE effect, with dynamics governed by the hot electron dynamics. If electron heating is efficient, a higher photon energy leads to a larger photovoltage (green line) than with a lower photon energy (blue line). **c**, Schematic of the ultrafast photovoltage set-up and the 'dual-gated graphene device' (see Supplementary Section I for details). The set-up produces two ultrashort pulses, separated in time by a controllable delay time t. A pulse shaper compensates for pulse stretching in the objective. The compressed pulses are focused onto the device, which contains a back gate (BG) and a top gate (TG) for creation of a p–n junction at the interface between graphene regions of opposite doping. We read out the photocurrent through the source (S) and drain (D) contacts, using a pre-amplifier and a lock-in amplifier, synchronized with the optical chopper. **d**, Development of electron temperature after excitation with an ultrafast pulse pair with $t = 1$ ps for two scenarios: independent heating due to the two pulses (black area) and heating where the heating due to the second pulse depends on the heating due to the first pulse (purple area). The measured photocurrent I_{PC} is proportional to the purple area, and the photocurrent dip ΔI_{PC} is proportional to the difference between the black and purple areas. If $t = 5$ ps, the black and purple areas are very similar; that is, $\Delta I_{PC} \approx 0$. Thus, the photocurrent dip as a function of t reflects the heating dynamics.

number of pulse pairs, as a function of delay t. After absorption of the first pulse, the electron temperature T_{el} rises by ΔT_{el} and then starts cooling to T_0. If the second pulse arrives before T_{el} reaches T_0 (that is, when the delay between the two pulses is shorter than the cooling time), the electron temperature rises again but by less than ΔT_{el}. This is because the electronic heat capacity C_{el} of a degenerate electron gas increases with electron temperature[27] and

$$\Delta T_{el} = \int_{Q_0}^{Q_0 + \Delta Q} dQ / C_{el}(T_{el})$$

where Q_0 is the heat in the system before photoexcitation and ΔQ is the absorbed power from a laser pulse. Thus, the additional amount of photovoltage generated by the second pulse is lower than when the pulses each contribute independently. As a result, the photovoltage as a function of delay time directly reflects the time dynamics of the electron temperature and therefore also the PTE-induced photovoltage.

We applied our femtosecond photovoltage sensing technique to a graphene p–n junction device consisting of a bottom and top dual-gated graphene flake (a 'dual-gated device'; Supplementary Section I). A scanning photocurrent image is shown in Fig. 2a, where the top-gate and bottom-gate voltages are such that a p–n junction is formed at the edge of the top gate. Figure 2b shows the photocurrent with the laser focused at this position, as a function of back-gate and top-gate voltages. The multiple sign reversals indicate that the photovoltage is generated through the PTE effect[3,4]. The photocurrent generated in this device is shown as a function of t

in Fig. 2c, which clearly shows a dip around $t = 0$. Figure 2d,e presents the normalized photocurrent dip ΔI_{PC} for two combinations of gate voltages, both in the p–n regime. The decays on both sides of $t = 0$ reflect the cooling dynamics with a picosecond timescale, as observed in refs 18–20. Around $t = 0$ we notice that the photocurrent dip ΔI_{PC} is remarkably sharp, which is only possible when the time resolution of the complete system (that is, the laser pulses and the graphene photoresponse) is sufficiently high. Any decrease in time resolution, either because of longer pulses or due to a slower generation of the photovoltage in graphene, would lead to a broadening of the apex of the inverted V-shape.

We use two approaches to quantitatively determine the timescale of photovoltage creation. In the first approach, we characterize the sharpness of the photocurrent dip around $t = 0$ by taking the time derivative. In the Supplementary Section III we show that in the case of exponential heating dynamics, taking the derivative directly recovers the heat dynamics, $dI_{PC}/dt(t) \approx (1 - e^{-t/\tau_{heat}})e^{-t/\tau_{cool}}$, where τ_{heat} and t_{cool} are the electron heating and electron cooling timescales, respectively. Figure 2f shows the experimentally obtained $dI_{PC}/dt(t)$ together with the heating dynamics using $\tau_{heat} = 80$ fs, obtained from a best fit. To show that we can indeed resolve timescales shorter than the timescale that was accessible in earlier time-resolved studies[18–20], we also plot the dynamics with $\tau_{heat} = 250$ fs. This slower heating time clearly does not fit the experimental data. In the second approach we develop a model for the photovoltage as a function of t, based on heating dynamics induced by the
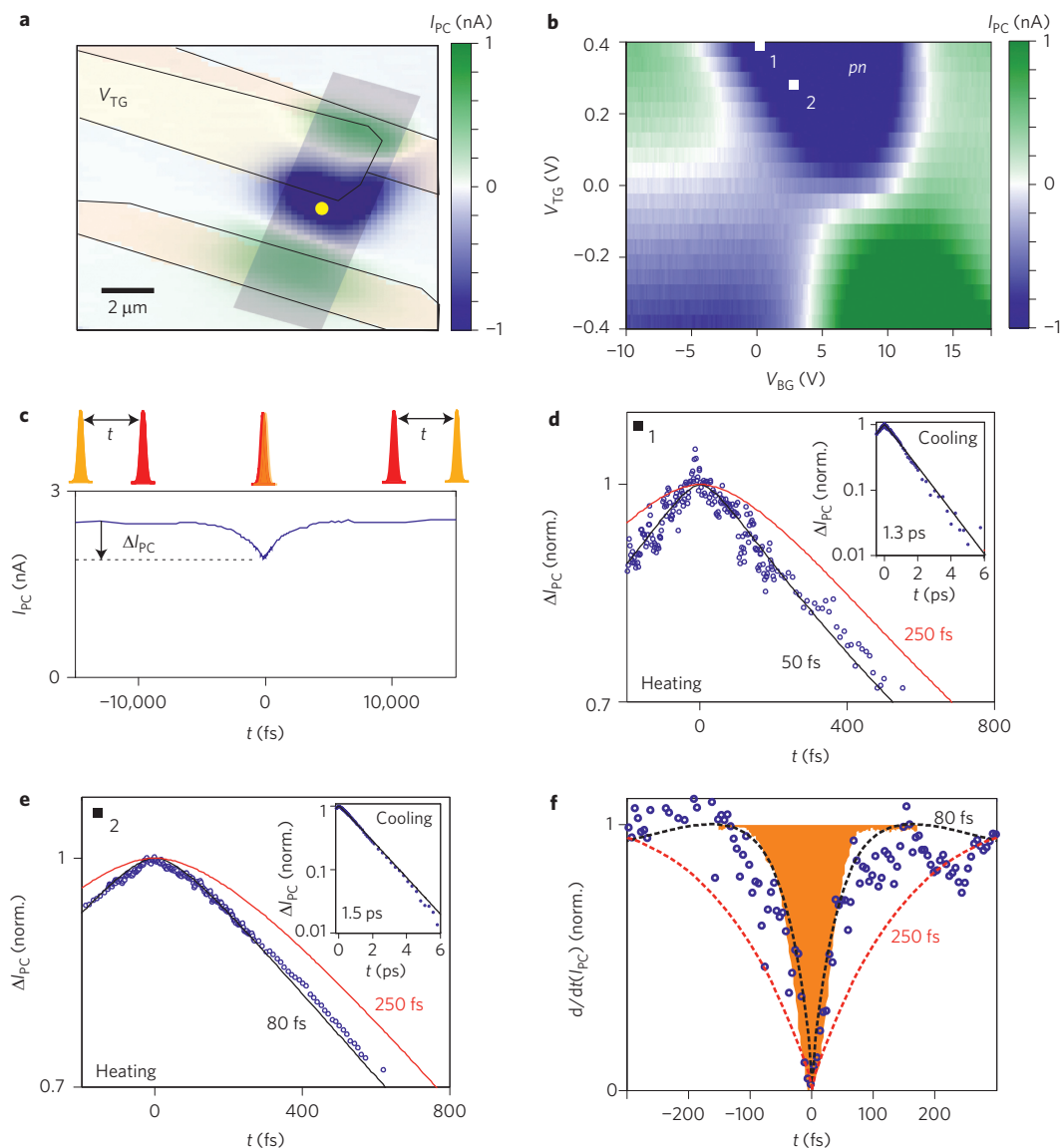
Figure 2 | Femtosecond sensing of hot electrons. a, Scanning photocurrent image (green–blue colour scale) of the 'dual-gated device', with the edges of the two contacts and the top gate indicated, as well as the location of the graphene flake (grey rectangle). At the edge of the top gate there is an interface of two graphene regions whose Fermi energy is separately controlled by the voltages on the back gate and the top gate, here tuned to the p–n junction regime. **b**, Photocurrent as a function of gate voltages at the position of the yellow dot in **a**, showing clear multiple sign reversals that are indicative of PTE current generation[3,4]. **c**, Photovoltage as a function of delay time t, showing a dip ΔI_{PC} when the two pulses overlap and recovery with dynamics that correpond to the hot electron dynamics. **d,e**, Dynamics of the (normalized) time-resolved photocurrent dip ΔI_{PC} around $t = 0$ (blue circles) for gate configurations 1 (**d**) and 2 (**e**), both in the p–n junction regime. Black solid lines describe the model results (see Methods), using a heating time of 50(80) fs and a cooling time of 1.3(1.5) ps for gate configuration 1(2). Red lines show the modelled dip with a slower heating timescale of 250 fs, which is incompatible with the data. Insets: Data and model results over a larger time range. **f**, Time derivative of the photocurrent dip (blue circles) in gate configuration 2, together with the derivative of the modelled photocurrent dip with heating timescales of 80 fs (black dashed line) and 250 fs (red dashed line). The orange area represents the optically measured cross-correlate using a nonlinear crystal.

pulse-pair excitation and including nonlinear heating (see Methods). We show the data for the first (second) gate voltage combination in Fig. 2d,e, together with the modelled photocurrent change $\Delta I_{PC}(t)$. We find excellent agreement between data and model for $\tau_{heat} = 50(80)$ fs and a cooling time of $\tau_{cool} = 1.3(1.5)$ ps. As an illustration we show the model for $\tau_{heat} = 250$ fs, which is clearly in strong disagreement with the data. We thus conclude that for the p–n junction configuration, photovoltage generation occurs within 50 fs.

We now put this capability of ultrafast photovoltage generation into the perspective of an application. The switching speed of

graphene optoelectronic devices using the direct photoresponse is limited by the picosecond cooling time, as shown by earlier reports[18,21], and is limited to a few hundred gigahertz. However, we envision femtosecond photosensing applications by exploiting the nonlinear heating response and combining graphene PTE devices with time-differential operation. Here, we provide one proof-of-principle demonstration, in which we use the graphene photodetector to measure in a direct electrical signal the duration of an ultrashort femtosecond laser pulse. We compare the derivative photocurrent signal $dI_{PC}/dt(t)$ with the optical cross-correlation signal that is measured by overlapping the two laser pulses in a
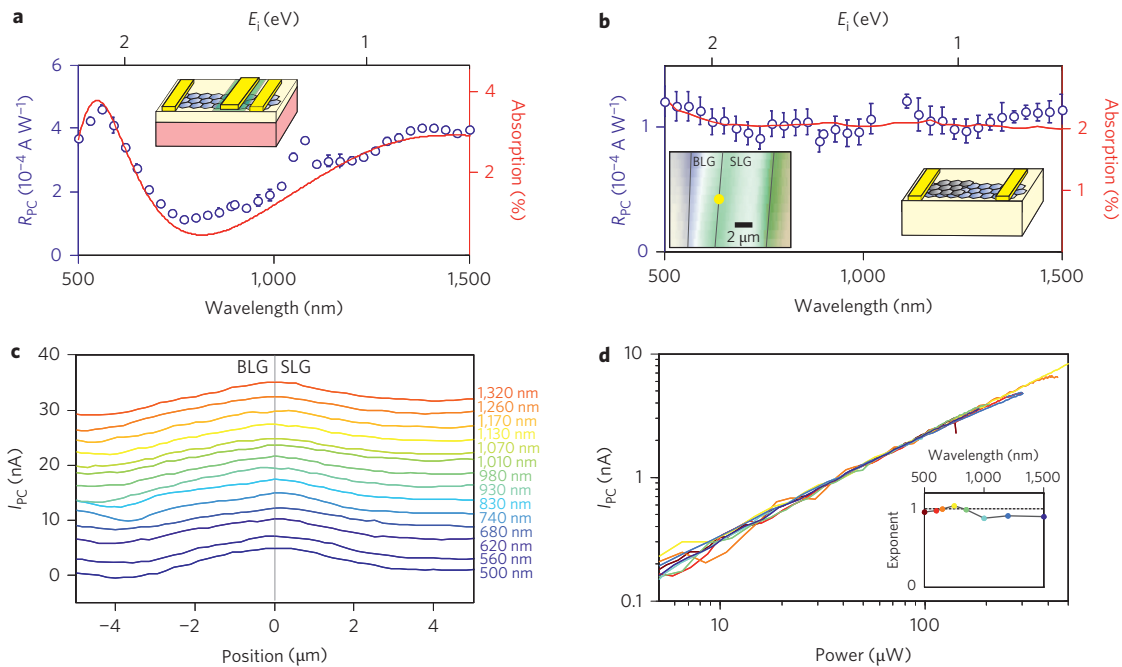
Figure 3 | Spectral response. a, Responsivity as a function of excitation wavelength (blue data points, left axis) and modelled absorption spectrum (red line, right axis) for the 'dual-gated device' in the p–n configuration, measured with a fixed power. Inset: Device layout, with a 300 nm SiO_2 substrate on top of Si. **b**, Responsivity as a function of excitation wavelength (blue data points, left axis) and modelled absorption spectrum (red line, right axis) for the 'transparent substrate device' at the monolayer/bilayer graphene interface (∼50 µW power). Left inset: Scanning photovoltage image with photovoltage generation at the monolayer/bilayer (SLG/BLG) interface. Right inset: Device layout with a 1-mm-thick SiO_2 substrate. The agreement between the responsivity and absorption curve shows that a lower number of absorbed photons (shorter wavelength, higher photon energy) does not lead to a lower responsivity, consistent with PTE current generation[3]. In **a,b** error bars are calculated from independent measurement scans and represent the 68% confidence interval. **c**, Photocurrent as a function of laser spot position for wavelengths between 500 and 1,320 nm (offset for clarity), while scanning the laser through the interface between monolayer and bilayer graphene, showing constant photovoltage amplitudes and spatial extent. The laser focus spot size is 1.5 ± 0.15 µm. **d**, Power dependence of the photovoltage with the laser focused at the SLG/BLG interface (see inset in **b**) for a range of wavelengths (see inset for the wavelength corresponding to each colour), showing linear scaling, which corresponds to the 'weak heating' regime, where $\Delta T_{el} < T_0$. Inset: The fitted power exponent, which is close to one.

second harmonic generation crystal and monitoring the second harmonic signal at 400 nm as a function of delay time (orange area in Fig. 2f). The agreement shows that our graphene PTE device is capable of measuring the pulse width of the laser down to timescales below 50 fs, in a direct electrical signal, and without the use of nonlinear crystals. We note that this technique will work over a much broader spectral range (from the ultraviolet[16] to the terahertz[9]) than techniques based on two-photon absorption in silicon photodiodes or frequency conversion in nonlinear crystals and has a similar sensitivity.

Efficient photoinduced carrier heating

Having established that carrier heating and PTE photovoltage generation occur on a femtosecond timescale, we now address how this step affects the energy conversion efficiency of graphene PTE devices. The main question is whether the carrier heating is fast enough to outcompete energy loss processes, such as optical phonon emission (step (ii) in Fig. 1a). To this end we study the photoresponse for a wide range of photon energies, from 0.8 eV (1,500 nm) to 2.5 eV (500 nm). We use a laser source (quasi-continuous wave (c.w.), because the pulse duration of 20 ps is larger than the cooling time of ∼1 ps) with a controllable wavelength and a constant excitation power P_{exc}. Figure 3a presents the (external) responsivity R_{PC} for the dual-gated p–n junction device. This spectral response is dominated by the strongly wavelength-dependent absorption spectrum of the device, which is the result of reflections at the oxide/silicon interface and the subwavelength oxide thickness (Supplementary Section V). This effect is very similar to

the enhanced reflection for certain combinations of wavelength and oxide thickness, which enhance the contrast of graphene in an optical microscope[28]. Indeed, the calculated graphene absorption (using Lumerical FDTD Solutions software) agrees well with the wavelength-dependent responsivity.

To avoid the strongly wavelength-dependent absorption we used a graphene device that is supported by a transparent substrate consisting of 1-mm-thick quartz ('transparent substrate device'; Supplementary Section I). The flake contains both single- and bilayer graphene, with PTE photovoltage generation at the interface[2]. Figure 3b shows the responsivity spectrum at the monolayer/bilayer interface, together with the measured graphene absorption on a similar device. These data are obtained from spatially resolved measurements (Fig. 3c), which show that the spatial extent of the photoresponse does not change with wavelength. The spectral response (for constant excitation power) of the device is strikingly constant over this broad range of excitation wavelengths. The flat R_{PC} shows that decreasing the number of incident photons (by increasing the photon energy) does not lead to a decrease in photovoltage, so a higher photon energy gives a larger photovoltage, as in Fig. 1b. This is in stark contrast to photovoltaic detectors based on semiconductors, where the photovoltage generally decreases with increasing photon energy, meaning that excess energy is lost[26]. The flat R_{PC} at the monolayer/bilayer interface is therefore consistent with PTE current generation (which also applies to the graphene/metal interface; Supplementary Section V).

To understand the flat, broadband R_{PC} for constant absorbed power we examined what this result means for the electron
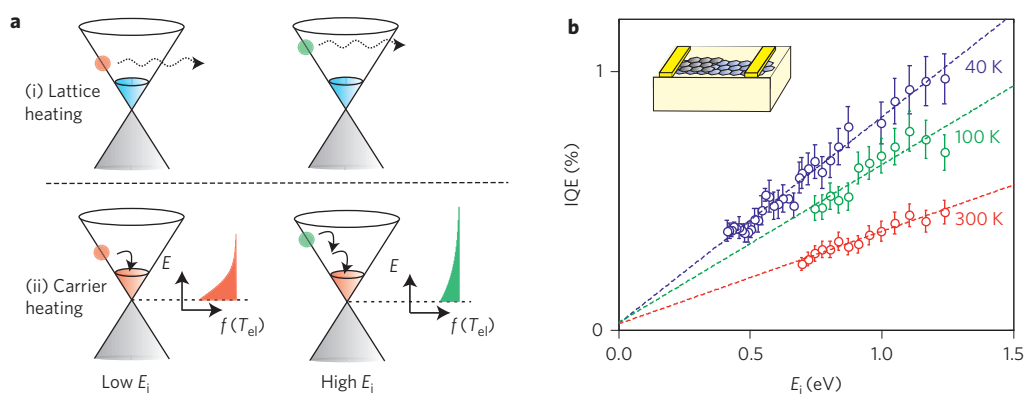
Figure 4 | Electron heating efficiency. a, Schematic representation of ultrafast energy relaxation after photoexcitation through phonon emission, which leads to lattice heating (top), and through carrier–carrier scattering, which leads to carrier heating (bottom). In the case of lattice heating, a larger photon energy leads to a larger phase space to scatter to and therefore more energy is transferred to the phonon bath, predicting sublinear scaling of the photovoltage (normalized by absorbed photon density) with photon energy. In the case of carrier heating, a larger photon energy leads to a hotter electron distribution (see the smeared Fermi–Dirac distributions next to the Dirac cones), predicting linear scaling of photocurrent (normalized by absorbed photon density) with photon energy (in the weak heating regime). **b,** Internal quantum efficiency (IQE), which represents the photovoltage normalized by the absorbed photon density, as a function of initial electron energy E_i after photoexcitation for ambient temperatures $T_0 = 40$, 100 and 300 K. Error bars are calculated from independent measurement scans and represent the 68% confidence interval. The linear scaling through the origin shows that heating dominates the ultrafast energy relaxation and therefore that electron heating is efficient.

heating $\Delta T_{el} = T_{el} - T_0$ as a function of photon energy. First we note that the photovoltage scales linearly with power for all wavelengths (Fig. 3d), which means that we are in the 'weak heating' regime where $\Delta T_{el} < T_0$ and the electronic heat capacity is constant (in contrast to the 'strong heating' regime in the ultrafast experiment, where the scaling is sublinear). The reason for this small amount of heating is that we use more than ten times lower power and quasi-c.w. laser excitation with ~20 ps laser pulses, which is longer than the cooling time of ~1 ps (leading to a peak power that is three orders of magnitude smaller than for ultrafast excitation). In this weak heating regime the cooling rate is constant[20], which means that the conversion efficiency is not affected by the lifetime of the hot electrons. Furthermore, the Seebeck coefficients S_1 and S_2 do not change with excitation wavelength. Therefore, from the flat R_{PC} we conclude that the light-induced increase in electron temperature ΔT_{el} at constant power is the same for all photon energies.

This result enables us to assess the efficiency of the electron heating (in the weak heating regime). We examined two alternative ultrafast energy relaxation pathways for photo-excited electrons and holes: carrier–carrier scattering and optical phonon emission (Fig. 1a). Graphene optical phonons have an energy of ~0.2 eV, so photoexcited carriers above 0.2 eV can in principle relax by emitting a phonon. The faster process of these two competing processes will dominate the ultrafast energy relaxation. We have determined the timescale of carrier–carrier scattering (through photovoltage generation in the 'dual-gated device') to be <50 fs. The timescale of phonon emission is typically <150 fs (refs 11,12), so this does not give a definite answer about which ultrafast relaxation process dominates. However, from the measured spectral responsivity we can extract the heating efficiency.

We illustrate this by considering two contrasting cases (Fig. 4a): (i) dominant coupling of photo-excited electrons to optical phonons, with a small fraction of the absorbed photon energy converted into hot electrons, and (ii) dominant carrier–carrier scattering[29,30], with a large fraction of the energy converted into hot electrons. In case (i), the energy loss rate dE/dt through optical phonon emission increases linearly with initial electron energy E_i, as it is governed by a constant electron–phonon coupling and an energy-momentum scattering phase space that increases linearly with energy[16] (Fig. 4a). Thus in case (i), a larger E_i leads to more

energy loss to phonons. On the other hand, in case (ii), the electron temperature scales linearly with E_i, because the energy of the photoexcited electron is fully transferred to the electron gas. Thus, the role of optical phonon emission can be measured by studying the scaling of ΔT_{el} with E_i. This relationship, extracted from the photovoltage measurements, is shown in Fig. 4b, where we plot the internal quantum efficiency (IQE = $I_{PC}E_i/\Delta Qe$). The IQE represents the generated photovoltage, normalized by the number of absorbed photons. We find (nearly perfect) linear scaling of the IQE with E_i, as the linear fits go nearly through the origin. This means that a higher photon energy corresponds to a larger photovoltage and thus to a larger ΔT_{el}, which is consistent with terahertz photoconductivity measurements[16,17,30], where a terahertz probe provides a measurement of the carrier temperature. We therefore conclude that the generated photovoltage comes from ultrafast, efficient photon-to-electron-heat conversion and an unmeasurably small loss to optical phonons.

To show that carrier heating is indeed efficient, we calculate the expected temperature rise ΔT_{el} and use the measured photovoltage to determine the Seebeck factor ($S_2 - S_1$), which we then compare to its expected value. Details are given in the Methods. To calculate ΔT_{el} we use a simple heat equation and assume fully efficient carrier heating to find ~0.17 K ($P_{exc} = 50$ μW, $T_0 = 300$ K, $E_F = 250$ meV, $\tau_{cool} = 1$ ps). We then use the measured photovoltage of $V_{PTE} = I_{PC} \times R \approx 10$ μV, where R is the device resistance (2 kΩ). We conclude that $S_2 - S_1 \approx 65$ μV K^{-1}. This is very close to the maximum expected value using a charge puddle width of $\Delta = 80$ meV, which gives 90 μV K^{-1}. Having confirmed that carrier heating is efficient, we aim to gain insight into the other factors that determine the overall energy conversion efficiency of the device. One important factor is the hot carrier lifetime, which should be long (low cooling rate) to lead to a larger photovoltage. We verify this by changing the ambient temperature. First we demonstrate that the efficiency of electron heating is independent of lattice temperature, as we obtain the same linear scaling through the origin for a lattice temperature of 40, 100 and 300 K. We then note that the overall photovoltage is larger for lower lattice temperatures. This is caused by the longer cooling time τ_{cool} at low temperatures, due to a lower coupling between electrons and acoustic phonons[3,20,31]. The Seebeck coefficient

decreases with temperature, meaning that the overall generated photovoltage is a factor of about two larger[20].

Conclusion

The unique femtosecond time resolution and the related high carrier heating efficiency are very encouraging results for bias-free (passive) PTE photodetectors. To improve the IQE of ~1% (Fig. 4b), an interesting approach (see Methods for details) would be to use ultraclean, defect-free graphene, such as in ref. 32, which enables detector operation with a higher Seebeck coefficient, because the electron–hole puddle density is lower (see Methods). The small puddle width could increase the Seebeck factor to $S_2 - S_1 \approx 300\ \mu V\ K^{-1}$. Measuring at $E_F = 50$ meV instead of 250 meV would furthermore lead to a larger ΔT_{el}, because this scales with $1/E_F$ (due to the lower heat capacity upon approaching the Dirac point). With 50 μW excitation a ΔT_{el} of almost 1 K is feasible (assuming an unmodified carrier cooling rate), giving $V_{PTE} \approx 300\ \mu V$. A resistance of $R = 1$ kΩ (ref. 32) would then give an IQE of 100% or more. Our results thus show that graphene PTE devices exhibit ultrafast, efficient and broadband photodetection. Future photodetection and light harvesting devices could exploit these graphene properties and combine them with the advantageous properties of other two-dimensional materials. Future work could investigate the effect of the Fermi energy on the heating time, the time-resolved signal generation in devices based on the bolometric effect[33,34] and the effect of the substrate on the heat dynamics.

Methods

Simulation of photocurrent dynamics. We now describe how we extract the hot electron dynamics from the photocurrent dip $\Delta I_{PC}(t)$. This analysis draws on the procedure described in ref. 20. We start with a laser-induced change in electron temperature for single pulse excitation assuming a linear response, $\Delta T_{el,1p}(t') = (1 - e^{-t'/\tau_{heat}})e^{-t'/\tau_{cool}}$, where t' is the 'real' time, τ_{heat} is the electron heating time and τ_{cool} is the electron cooling time. We then introduce a nonlinearity (any type of nonlinearity works) of the form $\Delta T'_{el,1p} = \sqrt{1 + \Delta T_{el,1p}/T_0} - 1$ (where T_0 is the lattice temperature) and calculate the integrated photovoltage generated by two pulses well separated in time (adding up independently): $2 \times V_{1p} = 2 \times \int_0^{\infty} \Delta T'_{el,1p} dt'$. We now follow the same procedure for two-pulse excitation, where the pulses arrive separated by delay time t, and obtain $\Delta T_{el,2p}$ (linear response) and $\Delta T'_{el,2p}$ (with nonlinearity). This temperature change $\Delta T'_{el,2p}$ is shown in Fig. 1d in 'real' time for two different delay times t. We use these electron temperature dynamics to calculate the integrated voltage V_{2p} and repeat this for a range of delay times t. The photocurrent dip as a function of delay time is then given by $\Delta I_{PC}(t) \propto 2 \times V_{1p} - V_{2p}(t)$. This dip grows on approaching $t = 0$, but flattens when $t < \tau_{heat}$. This is because when the two pulses arrive at exactly the same time ($t = 0$) there is at first no heating as a result of either pulse, but they then independently start to produce electron heating and thus contribute to the photovoltage. As a result, the photocurrent dip flattens around $t = 0$. We note that the dip should not disappear, because we measure the time-integrated photocurrent and, once the electrons start heating up, the heating induced by the two pulses is no longer independent. Finally, we point out that all dynamics are determined by local (at the junction) carrier heating and PTE voltage creation. All dynamics that occur after this (propagation of the potential to the contacts, signals travelling through the cables, and so on) do not affect the results of our specific experiment. To measure transit times, one can use two-pulse excitation with two different focus spots[35].

Calculation of ΔT, V_{PTE} and $S_2 - S_1$. We calculate the theoretical and experimental photovoltage V_{PTE} for excitation with light of $P_{exc} = 50$ μW (of which ~2% is absorbed, Fig. 3b), for the 'transparent substrate device', where the photovoltage is created at the interface of the monolayer and bilayer graphene at room temperature. We calculate the time-averaged temperature increase from a simple heat equation (for the case where the cooling length is smaller than the spot size)[6]:

$$\Delta T = T_{el} - T_0 = \frac{\Delta Q}{2d^2 \sqrt{4\pi \ln 2} \times \Gamma_{cool}} \approx 0.17\ K \qquad (1)$$

where the laser focus spot size is $d \approx 1.5$ μm (Fig. 3c) and the cooling rate is given by $\Gamma_{cool} = \alpha T_{el}/\tau_{cool}$, and $\alpha = (2\pi E_F/3\hbar^2 v_F^2)k_B^2$, where $\hbar$, v_F and k_B are the reduced Planck constant, the Fermi velocity and Boltzmann's constant, respectively. We use $E_F = 0.25$ eV and a cooling time of 1 ps, which we obtained from ultrafast photocurrent measurements on the same device. This yields a cooling rate of 0.4 MW m^{-2} K^{-1}, which is close to theoretical estimates (0.5–5 MW m^{-2} K^{-1})[6].

To obtain the photovoltage we use[6]

$$V_{PTE} = I_{PC} \times R \approx 10\ \mu V \qquad (2)$$

where we use a photocurrent of 5 nA for $P_{exc} = 50$ μW and a device resistance of $R = 2$ kΩ. Using $V_{PTE} = \Delta T_{el}(S_2 - S_1)$, we thus obtain $S_2 - S_1 = 65\ \mu V\ K^{-1}$. The maximum Seebeck factor is given by $(S_2 - S_1)_{max} = \pi^2 k_B^2 T/3e\Delta$, which gives 90 μV K^{-1} at room temperature for a charge puddle width of $\Delta = 80$ meV (ref. 4).

References

1. Wei, P., Bao, W., Pu, Y., Lau, C. N. & Shi, J. Anomalous thermoelectric transport of Dirac particles in graphene. *Phys. Rev. Lett.* **102,** 166808 (2009).
2. Xu, X. *et al.* Photo-thermoelectric effect at a graphene interface junction. *Nano Lett.* **10,** 562–566 (2010).
3. Song, J. C. W. *et al.* Hot carrier transport and photocurrent response in graphene. *Nano Lett.* **11,** 4688–4692 (2011).
4. Gabor, N. M. *et al.* Hot carrier-assisted intrinsic photoresponse in graphene. *Science* **334,** 648–652 (2011).
5. Koppens, F. J. L. *et al.* Photodetectors based on graphene, other two-dimensional materials and hybrid systems. *Nature Nanotech.* **9,** 780–793 (2014).
6. Freitag, M., Low, T. & Avouris, P. Increased responsivity of suspended graphene photodetectors. *Nano Lett.* **13,** 1644 (2013).
7. Echtermeyer, T. J. *et al.* Photothermoelectric and photoelectric contributions to light detection in metal–graphene–metal photodetectors. *Nano Lett.* **14,** 3733–3742 (2014).
8. Bonaccorso, F., Sun, Z., Hasan, T. & Ferrari, A. C. Graphene photonics and optoelectronics. *Nature Photon.* **4,** 611–622 (2010).
9. Cai, X. *et al.* Sensitive room-temperature terahertz detection via the photothermoelectric effect in graphene. *Nature Nanotech.* **9,** 814–819 (2014).
10. George, P. A. *et al.* Ultrafast optical-pump terahertz-probe spectroscopy of the carrier relaxation and recombination dynamics in epitaxial graphene. *Nano Lett.* **8,** 4248–4251 (2008).
11. Lui, C. H. *et al.* Ultrafast photoluminscence from graphene. *Phys. Rev. Lett.* **105,** 127404 (2010).
12. Breusing, M. *et al.* Ultrafast nonequilibrium carrier dynamics in a single graphene layer. *Phys. Rev. B* **83,** 153410 (2011).
13. Johannsen, J. C. *et al.* Direct view on the ultrafast carrier dynamics in graphene. *Phys. Rev. Lett.* **11,** 027403 (2013).
14. Gierz, I. *et al.* Snapshots of non-equilibrium Dirac carrier distributions in graphene. *Nature Mater.* **12,** 1119–1124 (2013).
15. Brida, D. *et al.* Ultrafast collinear scattering and carrier multiplication in graphene. *Nature Commun.* **4,** 1987 (2013).
16. Tielrooij, K. J. *et al.* Photoexcitation cascade and multiple hot-carrier generation in graphene. *Nature Phys.* **9,** 248–252 (2013).
17. Jensen, S. *et al.* Competing ultrafast energy relaxation pathways in photoexcited graphene. *Nano Lett.* **14,** 5839–5845 (2014).
18. Urich, A., Unterrainer, K. & Mueller, T. Intrinsic response time of graphene photodetectors. *Nano Lett.* **11,** 2804–2808 (2011).
19. Sun, D. *et al.* Ultrafast hot-carrier-dominated photovoltage in graphene. *Nature Nanotech.* **7,** 114–118 (2012).
20. Graham, M. W., Shi, S-F., Ralph, D. C., Park, J. & McEuen, P. L. Photocurrent measurements of supercollision cooling in graphene. *Nature Phys.* **9,** 103–108 (2013).
21. Xia, F., Mueller, T., Lin, Y., Valdes-Garcia, A. & Avouris, P. Ultrafast graphene photodetector. *Nature Nanotech.* **4,** 839–843 (2009).
22. Mueller, T., Xia, F. & Avouris, P. Graphene photodetectors for high-speed optical communications. *Nature Photon.* **4,** 297–301 (2010).
23. Gan, X. *et al.* Chip-integrated ultrafast graphene photodetector with high responsivity. *Nature Photon.* **7,** 883–887 (2013).
24. Pospischil, A. *et al.* CMOS-compatible graphene photodetector covering all optical communication bands. *Nature Photon.* **7,** 892–896 (2013).
25. Schall, D. *et al.* 50 GBit/s photodetectors based on wafer-scale graphene for integrated silicon photonic communication systems. *ACS Photon.* **1,** 781–784 (2014).
26. Sze, S. M. *Physics of Semiconductor Devices* (Wiley, 1969).
27. Kittel, C. *Introduction to Solid State Physics* (Wiley, 2005).
28. Blake, P. *et al.* Making graphene visible. *Appl. Phys. Lett.* **91,** 063124 (2007).
29. Winzer, T., Knorr, A. & Malić, E. Carrier multiplication in graphene. *Nano Lett.* **10,** 4839–4843 (2010).
30. Song, J. C. W. *et al.* Photoexcited carrier dynamics and impact-excitation cascade in graphene. *Phys. Rev. B* **87,** 155429 (2013).
31. Ma, Q. *et al.* Competing channels for hot-electron cooling in graphene. *Phys. Rev. Lett.* **112,** 247401 (2014).
32. Wang, L. *et al.* One-dimensional electrical contact to a two-dimensional material. *Science* **342,** 614–617 (2013).

33. Freitag, M., Low, T. & Avouris, Ph. Photoconductivity of biased graphene. *Nature Photon.* **7,** 53–59 (2013).
34. Yan, J. *et al.* Dual-gated bilayer graphene hot-electron bolometer. *Nature Nanotech.* **7,** 472–478 (2012).
35. Son, B. H. *et al.* Imaging ultrafast carrier transport in nanoscale field-effect transistors. *ACS Nano* **8,** 11361–11368 (2014).

Author contributions

K.J.T., F.H.L.K., N.v.H. and P.J.H. conceived the experiments. K.J.T., L.P. and M.M. carried out the experiments. K.J.T., M.M., L.P. and F.H.L.K. performed the data analysis. Q.M., M.M., Y.L. and C.N.L. fabricated the samples. K.J.T. and A.W. performed simulations. K.J.T., F.H.L.K., N.v.H. and P.J.H. wrote the manuscript, with the participation of all authors.

Additional information

Supplementary information is available in the online version of the paper. Reprints and permissions information is available online at www.nature.com/reprints. Correspondence and requests for materials should be addressed to K.J.T., N.F.v.H. and F.H.L.K.

Competing financial interests

The authors declare no competing financial interests.