Near-Infrared Quantum Cascade Lasers Designed with van der Waals Materials

Hai-Yao Deng[*]
School of Physics and Astronomy, Cardiff University, 5 The Parade, Cardiff, Wales CF24 3AA, United Kingdom

(Received 2 March 2021; revised 7 September 2021; accepted 30 September 2021; published 21 October 2021)

Quantum cascade lasers (QCLs) constitute a leading source of coherent radiation in the mid-IR region. However, their performance outside this region remains unsatisfactory. Indeed, there are currently no QCLs in the near-IR region. Here, we propose that a superlattice of atomically thin layers held together by van der Waals forces may operate near room temperature as a compact and powerful near-IR QCL emitting at a wavelength of 1.66 μm. It can compress over 100 stages within 0.5 μm. The electric field required for operation is about 3×10^6 V cm^{-1}, while the lasing threshold current density is about 22.4 kA cm^{-2} depending on parameters. Rate-equation analysis reveals that the peak power per unit volume can reach over 0.1 mW μm^{-3} in cw operation. Unlike existing QCLs, our device is p-type working with holes.

DOI: 10.1103/PhysRevApplied.16.044038

I. INTRODUCTION

The idea that a superlattice of quantum wells, when biased by a constant electric field, may amplify light was conceived in 1971 by Kazarinov and Suris [1]. It was realized two decades later by Faist *et al.* [2], who called their device the "quantum cascade laser" (QCL). Nowadays, QCLs constitute the leading source of coherent mid-IR radiation [3]. However, their performance in the far-IR and near-IR regions remains poor [4,5]. Indeed, due to material constraint, no near-IR QCL exists currently [6].

A QCL consists of a number of repeating units called stages forming a one-dimensional superlattice [7]. Each stage contains an *active region* and an *injector* that are separated by a tunneling barrier. The active region is a nanometric quantum well usually designed to support three energy levels that each extends into a conduction sub-band in the first Brillouin zone (1BZ) of the in-plane Bloch wave vector. An electron injected onto the highest level can transit to the middle one emitting a photon, the frequency of which is determined by the energy gap between the levels. Lasing action takes place when population inversion is achieved between them. The lowest level does not take part in optical transitions but facilitates population inversion [2]. One of the biggest advantages of QCLs over other types of lasers is that the energy gap can be widely tuned by varying the well width and barrier height. In principle, by engineering the quantum well, QCLs could emit light at any wavelength. In reality, however, this liberty is severely constrained by the availability of materials. Existing QCLs are based on traditional III-V semiconductor heterostructures, where the energy gap (and hence the phonon frequency) is capped by the conduction-band offset [8,9].

Atomically thin van der Waals (vdW) layers [10] may help us break the bottleneck. Since the isolation of graphene [11], these quasi-two-dimensional (2D) materials have grown into a large family displaying a plethora of properties ranging from semiconducting and insulating to metallic and semimetallic. Reputed members in this family include the semiconducting transition metal dichalcogenides [12] (TMDs, e.g., MoS_2 and $MoSe_2$) and the latest advent of indium monochalcogenides [13] (e.g., InSe), both possessing a hexagonal lattice. Physical and chemical techniques have been developed to organically stack these layers to realize exotic electronic, optical, and mechanical properties [14]. The relevance of vdW layers to QCLs rests with the fact that such a layer can be regarded as perhaps the thinnest (a few angstroms) quantum well [15]. By assembling the layers in the right order, one can in principle create the superlattice and sub-band structure required of QCLs operating in any spectral region.

In this paper, we theoretically propose a near-IR (approximately 1.6 μm) QCL consisting of a vdW heterostructure. The device makes use of few-layer MoS_2 and $MoSe_2$ as the active region and the injector respectively, while InSe layers are employed as tunneling barriers. It is based on holes rather than electrons, hence a p-type QCL, expected to work near room temperature.

In the next section, the detailed spatial and energy structures of the device are described, where the operation conditions (population inversion and threshold current) are also prescribed. In Sec. III, the performance of the device is analyzed using rate equations. We conclude the paper in Sec. IV. Some technical details are provided in Appendices A–D.

[*] DengH4@cardiff.ac.uk

II. DEVICE STRUCTURE AND OPERATION

The device analyzed in this work is schematically shown in Fig. 1. It solely draws on few-layer TMDs and InSe, which have similar lattice structures as sketched in (c). A TMD (InSe) monolayer comprises one (two) layer of metal (In) atoms covalently bonded to two layers of chalcogen atoms on a hexagonal lattice [12,13]. As depicted in Figs. 1(b) and 1(c), the active regions of our QCL are each served by a two-layer MoS_2 and the injectors each by a three-layer $MoSe_2$, while the tunneling barriers are made of InSe layers. The similarity between the crystal lattices of InSe and TMD layers helps to achieve efficient tunneling, which proves useful in attaining the population inversion required for lasing action.

The choice of materials reflects on their band structures. Density-functional theory (DFT) shows that [12] the highest-energy valence band (called the v band) of a TMD monolayer is widely gapped at the Γ point in the 1BZ from both the conduction bands and the rest valence bands. This band is derived from the p_z orbitals of chalcogen atoms and the metal d_{z^2} orbitals. Such orbital character implies that, when a few (say N) TMD layers are brought together, strong coupling exists between the N degenerate individual v bands and gives rise to N sub-bands that differ in energies by as much as a few hundreds of meV [12]. The energies at the Γ point of these sub-bands (labeled 1 and 2) of two-layer MoS_2 and those (labeled $1'$, $2'$, and $3'$) of three-layer $MoSe_2$ are displayed in Fig. 2(a). Sub-band $3'$ is of no use in our device and is disregarded hereafter. The states used for lasing are focused near the Γ point, which means that our device tolerates misorientation between the vdW layers that might occur during the fabrication process.

A uniform electric field E is applied normal to the layers to shift the energy levels in such a way that level 1 of an active region matches level $2'$ of its adjacent downstream (along the direction of E) injector while its level 2 matches level $1'$ of the adjacent upstream injector; see Fig. 2(b). This is achieved at $E \approx 3 \times 10^6$ V cm^{-1} with the active region separated from the downstream injector by a monolayer InSe while from the upstream one by a two-layer InSe, as illustrated in Figs. 1(c) and 2(b). The sub-band structures for this value of E are computed using the $\mathbf{k} \cdot \mathbf{p}$ tight-binding model developed recently [12] and shown in Fig. 2(c). Their dispersions can be well described by a parabolic $\varepsilon_n(k) \approx \varepsilon_n + \alpha_n k^2$, where ε_n and α_n are parameters (see Appendix A). The gap at the Γ point between sub-bands 1 and 2 is about $\hbar\omega_0 = 711$ meV, which converts into a wavelength of about 1.66 μm that falls in the near-IR region. Here $\hbar$ is the reduced Planck constant. The band structure of an InSe monolayer resembles that of a TMD monolayer [13]. It also features a v-type valence band that is originated from the p_z orbitals of In and Se atoms. As indicated in Fig. 2(a), it lies $\Delta \approx 1$ eV below level 1 at the Γ point, giving rise to a tunneling barrier height of Δ for holes, which may be further reduced by the electric field. InSe serves as a good tunneling barrier for TMD layers for three reasons. Firstly, as aforementioned, the lattice constants of InSe and TMD are close with a ratio $r \approx 5 : 4$. Secondly, the v-type sub-bands in both stem from atomic orbitals (p_z and d_{z^2}) that stick out of the layers. These attributes facilitate strong orbital overlap between the InSe layer and TMD layer. The hopping integral t between a TMD monolayer layer and InSe monolayer is essentially that of a bilayer InSe but reduced due to lattice mismatch. Thirdly, the tunneling barrier $\Delta \approx 1$ eV is

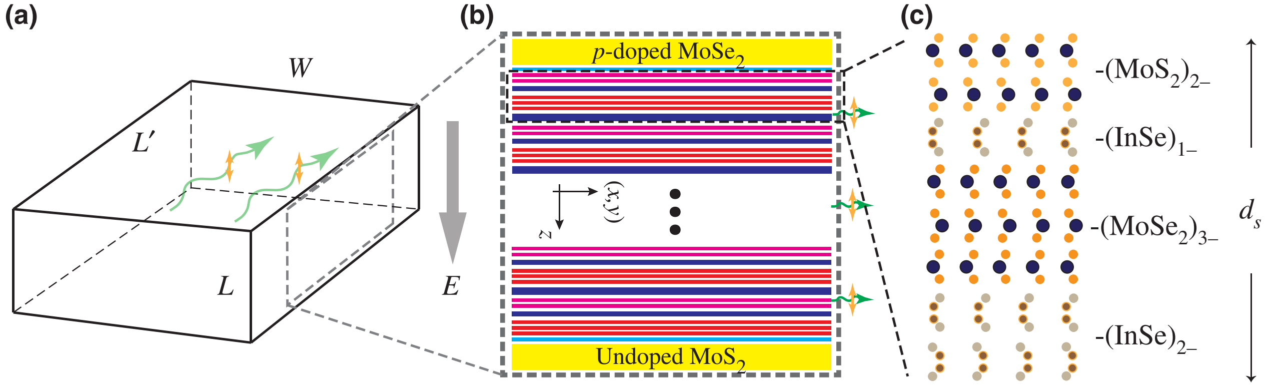

FIG. 1. Schematic structure of the van der Waals quantum cascade laser. (a) Dimensions of the device. Emitted photons (green wiggly lines) are polarized along the thickness direction (z, double arrowed yellow lines) and propagate normal to it. A constant and uniform electric field E is applied along positive z. (b) Side view of the vdW layers stacked into N_s stages between two electrodes (yellow pads), which also use vdW layers and allow holes to be injected into the device. (c) Materials composition of a stage, containing an *active region* made of a two-layer MoS_2 and an *injector* made of a three-layer $MoSe_2$, which are separated by InSe layers that serve as tunneling barriers. An active region is separated from the adjacent downstream (along the direction of E) injector by a monolayer InSe but from the upstream one by a two-layer InSe to achieve level alignment by the field E. The thickness of a stage is $d_s \approx 4.9$ nm. The whole structure is placed in an optical cavity.

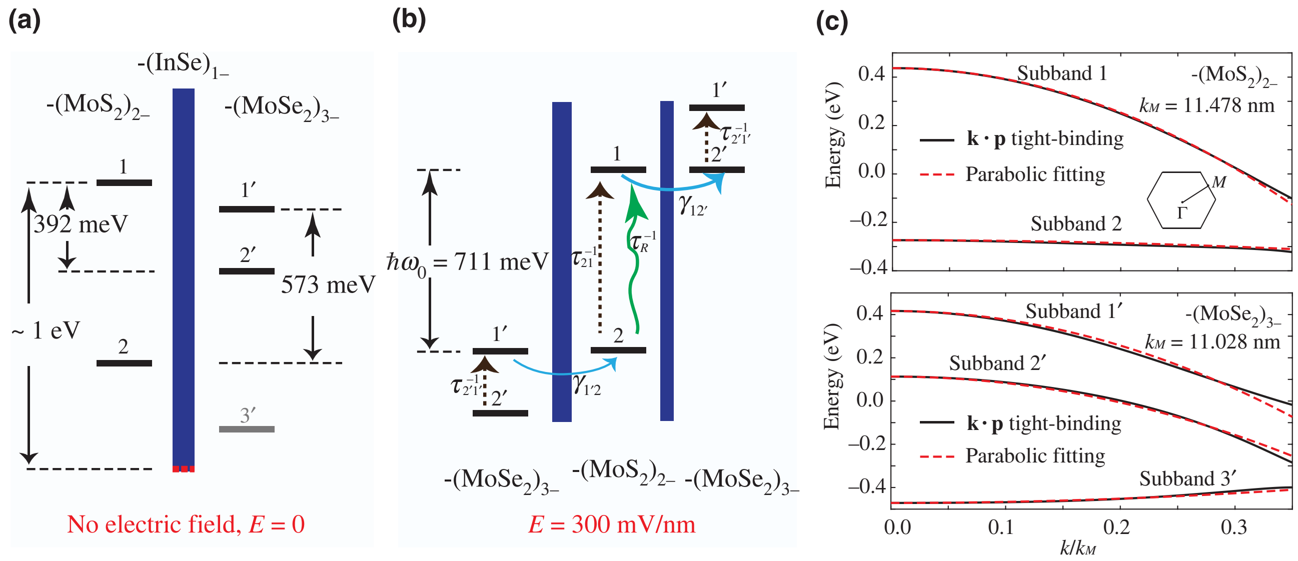

FIG. 2. Relative positions of the energy levels in a stage. (a) Alignment of the levels at the Γ point in the first Brillouin zone (1BZ) for $E = 0$. The active region (two-layer MoS$_2$) has two levels labeled 1 and 2, the injector (three-layer MoSe$_2$) with three levels 1′, 2′, and 3′ while the barrier (monolayer InSe) with one lying below level 1 by $\Delta \approx 1$ eV. Level 3′ is irrelevant for the QCL. (b) Alignment of energy levels and (c) energy sub-band structure (computed using a $\mathbf{k}\cdot\mathbf{p}$ tight-binding model, well fitted by parabolic curves) for $E = 300$ mV/nm ($= 3 \times 10^6$ V cm^{-1}). The frequency of the emitted photons is primarily determined by the gap between 1 and 2, $\hbar\omega_0 \approx 711$ meV. Only a tiny fraction approximately 10^{-6} of the 1BZ about the Γ point participates in the lasing action. The electric field E shifts the levels so that they match in the manner as indicated. As the gap between 1 and 2′ is about a half that between 2 and 1′, the barrier separating the active region from the next downstream injector is thinner than that spacing it from the upstream one. Horizontal arrowed lines: quantum tunneling through barriers ($\gamma_{1'2}$ and $\gamma_{12'}$). Vertical dashed lines: phonon-induced transitions (τ_{21}^{-1} and $\tau_{2'1'}^{-1}$). Vertical solid lines: radiative transitions (τ_R^{-1}). Inset in (c): 1BZ. k_M: wave number at the M point.

moderate [16]. In the case of two-layer InSe, Δ is slightly increased to about 1.3 eV.

The transport of a hole through the device is illustrated in Fig. 2(b). Let us track a hole that is initially in a state of sub-band 2 in some active region. It quickly thermalizes due to intra-sub-band phonon scattering and relaxes toward the Γ point. This fast relaxation is mainly due to the flatness [Fig. 2(c)] of this sub-band and it allows us to ignore states far from the Γ point in subsequent processes. Now this hole can either tunnel into a downstream injector or transit to sub-band 1 radiatively [green wiggly lines in Fig. 2(b)] at rate $1/\tau_R$ or nonradiatively (dashed lines) by phonon emission at rate $1/\tau_{21}$. Once arriving at sub-band 1, the hole, which cannot transit back to level 2 for the lack of phonon, can only tunnel out of the active region at rate $\gamma_{12'}$ into level 2′ of the next downstream injector, thence nonradiatively transiting to level 1′ at rate $1/\tau_{2'1'}$ and further tunneling into the next active region, and it gets recycled and the process is repeated [17]. Population inversion attains for

$$\tau_{21} > 1/\gamma_{12'} + \tau_{2'1'}, \tag{1}$$

as shown below. Physically, this expression says that, to attain population inversion between levels 1 and 2, a hole should stay in 2 as long as possible and tunnel out of 1 into 2′ as fast as possible. Tunneling is possible only if level 2′ is also empty, hence demanding a short lifetime of level 2′.

The various rates introduced above can be evaluated routinely (see Appendices A and B). Phonon-induced transition rates are calculated by the conventional electron-phonon coupling theory that has been detailed in various work [12,18–20]. The impact of vdW forces on the phonon structure of a multilayer TMD sheet is neglected [12,18]. Our calculations take into account the contributions from all six phonon branches that can be supported in a TMD monolayer [21]. With the electronic wave functions computed by the $\mathbf{k}\cdot\mathbf{p}$ model and the material parameters given in Appendix A, we find at room temperature

$$1/\tau_{21} \approx 40 \text{ THz}, \quad 1/\tau_{2'1'} \approx 162 \text{ THz}. \tag{2}$$

That τ_{21} far exceeds $\tau_{2'1'}$ is due to the much stronger electron-phonon coupling in MoSe$_2$ than MoS$_2$. The intra-sub-band relaxation rate is $1/\tau_2 \approx 250$ THz for sub-band 2 and $1/\tau_1 \approx 10$ THz for sub-band 1, the difference arising from the difference in the sub-band flatness, i.e., $\alpha_1 \approx 17\alpha_2$. For 2′, this is about 160 THz. As for the rate $\gamma_{12'}$ of a hole elastically tunneling from level 1 to 2′ through an InSe monolayer, we calculate it using second-order perturbation theory, which gives (see Appendix B)

$$\gamma_{12'} \approx 85 \text{ THz}. \tag{3}$$

Together with Eq. (2), this fulfills condition (1) for our device. The rate of tunneling from level $1'$ to 2 through a bilayer InSe is analogously obtained as $\gamma_{1'2} \approx 3$ THz. Finally, the spontaneous emission rate of a photon of frequency ω due to a hole jumping from sub-band 2 to sub-band 1 near the Γ point is $\tau_R^{-1}(\omega) = \tilde{\tau}_R^{-1}(\omega)/LWL'$, where $\tilde{\tau}_R^{-1}(\omega) = (4\pi^2/\hbar)\Lambda\omega e^2|z_{21}|^2\rho(\omega)$. Here $\Lambda \leq 1$ denotes the mode confinement factor, $z_{21} \approx 0.3$ nm is the transition matrix element [12] and $\rho(\omega)$ is the spectral overlap (see Appendix B). At resonance, i.e., $\omega = \omega_0$, one finds $\tau_R^{-1} \approx 10^{-5}$ THz for $\Lambda = 1$, $WL' = 1\ \mu\text{m}^2$, and $L = N_s d_s = 100 \times 4.9$ nm. This gives a radiative efficiency $\tau_{21}/\tau_R \approx 10^{-6}$, which is comparable to that of conventional QCLs.

III. DEVICE PERFORMANCE

The performance of the device is now analyzed by means of rate equations [7,22] (details in Appendix C). Let us denote by f_1 and f_2 the total population of holes in sub-bands 1 and 2, respectively. The population inversion then reads $\delta f = f_2 - f_1$. The total number of photons of frequency ω in the optical cavity is denoted by n. In cw operation sustained by an injection current J (see Fig. 3), the rate equations require that $f_1 = (J/e)\,(\tau_{2'1'} + 1/\gamma_{12'})$ and

$$\frac{J}{e} - f_2\left(\frac{1}{\tau_{21}} + \frac{1}{\tau_R}\right) - n\nu_{21}\delta f = 0, \quad n = \frac{f_2}{\delta f_{\text{th}} - \delta f}, \tag{4}$$

which can be solved to obtain δf and n. Here $\delta f_{\text{th}} = \tau_R/N_s\tau$ and τ is the cavity lifetime of a photon. As τ_R depends on ω, so do n and δf. For $J < J_{\text{th}}$, where J_{th} is the threshold current for lasing to be determined later, n is small and Eq. (4) implies $f_2 \approx J\tau_{21}/e$ and hence

$$\delta f \approx \frac{J}{e}\left(\tau_{21} - \tau_{2'1'} - \frac{1}{\gamma_{12'}}\right), \tag{5}$$

which dictates Eq. (1) as the condition for population inversion. For $J > J_{\text{th}}$, lasing sets in and n grows drastically while δf approaches δf_{th}, as shown in Fig. 3.

The threshold current J_{th} is directly obtained from Eq. (5) by setting $\delta f = \delta f_{\text{th}}$. Let us write $J_{\text{th}} = WL'\mathcal{J}_{\text{th}}$, where $\mathcal{J}_{\text{th}}$ is the threshold current density. One obtains

$$\mathcal{J}_{\text{th}} = \frac{e\gamma_{12'}d_s}{\tau/\tilde{\tau}_R}\frac{1}{\gamma_{12'}(\tau_{21} - \tau_{2'1'}) - 1}. \tag{6}$$

This expression can be derived in a more physical way [4,23]. To this end, let us note that the flux of photons is $\Phi = cn/L'$, where c is the speed of light in the medium. The number of photons generated per unit time per unit length is $\delta\Phi/\delta y = n\tau_R^{-1}N_s\delta f/L'$. The gain per unit length thus follows as $G = (\delta\Phi/\delta y)/\Phi = g\mathcal{J}$, where $\mathcal{J} = J/WL'$ is the current density and $g = N_s WL'/ec\tau_R\,(\tau_{21} - \tau_{2'1'} - 1/\gamma_{12'})$ is the gain coefficient. At the threshold one expects $Gl = 1$, which leads back to Eq. (6). Here $l = c\tau$ is the photon mean free path. For a Fabry-Perot cavity, one can write $1/l = 1/l_0 + 1/l_m$, where l_0 represents losses due to absorption and $1/l_m = -\ln(R_1R_2)/(2L')$ with $R_{1,2}$ being the reflectivity of the mirrors signifies photons escaping from the mirrors [24].

The threshold current density $\mathcal{J}_{\text{th}}$ is independent of the dimensions of the device. Inserting the rates obtained above and taking $\tau \approx 100$ ps, which may well be achievable with the state-of-the-art cavity [25], we find

$$\mathcal{J}_{\text{th}} \approx 22.4 \text{ kAcm}^{-2}, \quad \text{for } \omega = \omega_0. \tag{7}$$

This is similar to that of conventional QCLs. From this one can show that the states that participate in lasing occupy only a tiny fraction η of the 1BZ. Actually [26], $\eta = \Omega\mathcal{J}/e\gamma_{12'} = 10^{-6}J/J_{\text{th}}$, where $\Omega = 8.65$ Å^2 is the area of the unit cell of the MoS$_2$ monolayer. This justifies our disregarding the states outside the immediate neighborhood of the Γ point, an assumption that is implicit in the rate equations employed here.

The number of photons escaping from the cavity mirrors per unit time is $(n/\tau)\alpha$, where $\alpha = l/l_m$ is called the output

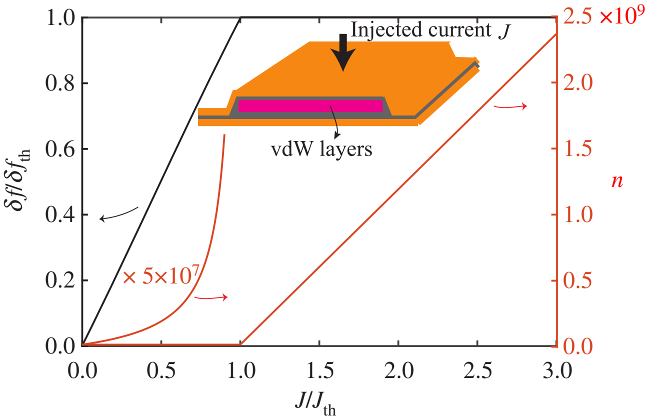

FIG. 3. Lasing action in the device. Left axis: population inversion, $\delta f = f_2 - f_1$, where $f_{1/2}$ is the population of holes in sub-band 1/2. Right axis: number of photons in the cavity, n. J is the injection current while J_{th} is the threshold current, and δf_{th} is the limit of δf at large J. Lasing sets in for $J > J_{\text{th}}$, where n drastically increases as δf approaches δf_{th}. Parameters: photon frequency $\omega = \omega_0$, photon life time $\tau = 100$ ps, and device area $WL' = 305\ \mu\text{m}^2$, for which $J_{\text{th}} \approx 67$ mA. Inset: sketch of the device.

coupling efficiency. The emission intensity is then given by [24]

$$P(\omega) = \frac{\alpha}{\tau} \sum_m \frac{\hbar\omega\, n(\omega_m)\, \delta\omega_m^2}{(\omega - \omega_m)^2 + \delta\omega_m^2}, \tag{8}$$

where $\omega_m = \mathrm{c}m\pi/L'$ are the frequencies of the modes, tagged by $m = 0, 1, \ldots$, supported by a Fabry-Perot optical cavity, and $\delta\omega_m = (1/\tau)\,[1 - \delta f\,(\omega_m)/\delta f_{\mathrm{th}}] \approx 1/n(\omega_m)\tau$ denotes the line width due to the net damping of the photons [27]. We reinstate the dependence of $n(\omega)$ and $\delta f\,(\omega)$ on ω for clarity. Only the modes with $J_{\mathrm{th}}(\omega_m) < J$ can be amplified. In line with the numerical solution displayed in Fig. 3, one gets from Eq. (4) that

$$n(\omega) \approx \begin{cases} \dfrac{\gamma_{12'}\tau_{21}}{\gamma_{12'}(\tau_{21} - \tau_{2'1'}) - 1}\dfrac{J}{J_{\mathrm{th}}(\omega)}, & \text{for } J < J_{\mathrm{th}}(\omega), \\ \dfrac{\tau_R(\omega)}{\tau}\left(\dfrac{J}{J_{\mathrm{th}}(\omega)} - 1\right), & \text{for } J > J_{\mathrm{th}}(\omega). \end{cases} \tag{9}$$

This expression suggests that in the spontaneous emission regime, where $J < J_{\mathrm{th}}$, $n(\omega)$ and hence $P(\omega)$ are independent of the dimensions of the device, whereas in the lasing regime, where $J > J_{\mathrm{th}}$, they increase linearly with the dimensions via $1/\tau_R \propto 1/WLL'$. Actually, $P(\omega) \propto \Lambda\alpha N_s WL'$.

In Fig. 4 is shown $P(\omega)$ for a device with a volume of 150 μm^3, for which $P(\omega)$ of the lasing modes [28] reaches over 100 mW with $\Lambda\alpha = 1$. In practice, $\Lambda\alpha$ is typically about 0.1 rather than unity, which means one has to increase the area by about 10 times to reach this power. That sets the output power per unit volume at approximately 0.1 mW μm^{-3}, much higher than with conventional QCLs.

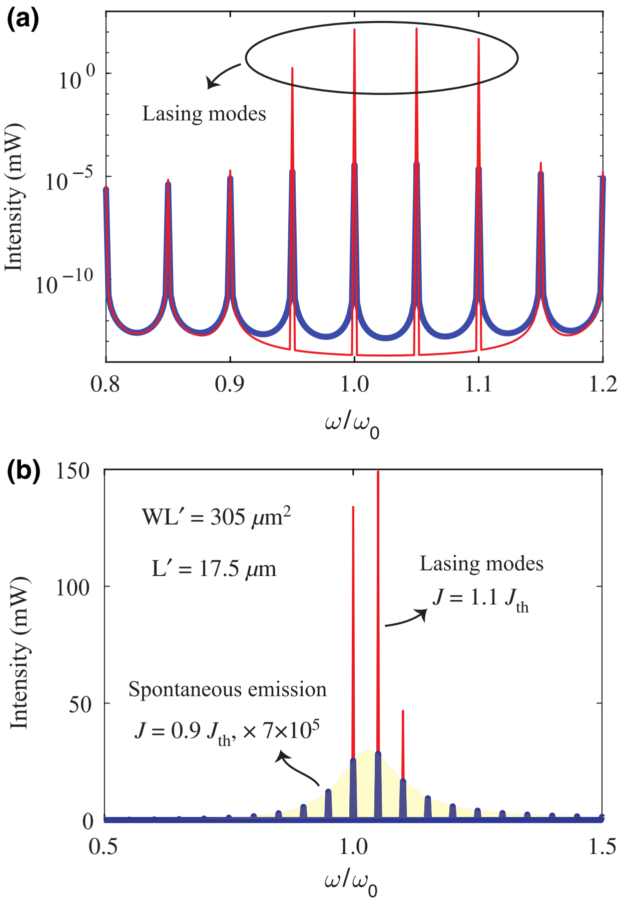

FIG. 4. Computed emission spectrum for a Fabry-Perot cavity. Intensity $P(\omega)$ versus photon frequency ω in (a) logarithmic and (b) linear scale for these parameters: out-coupling efficiency $\alpha = 1$, mode confinement $\Lambda = 1$, photon lifetime $\tau = 100$ ps, number of stages $N_s = 100$, and dimensions $W = L' = 17.5$ μm, for which $J_{\mathrm{th}} \approx 67$ mA. Extreme line narrowing occurs to the four lasing modes for dramatic increase in the number of photons n. The dimension L' is chosen so that ω_0 coincides with one of the eigenfrequencies of the cavity.

IV. CONCLUSION

We thus show that a compact and powerful near-IR QCL can be assembled purely out of vdW materials. In contrast to existing QCLs, our device is p type, with two levels in the active region and three discrete levels (not a quasicontinuum) in the injector. The scheme put forth here may also be used to create terahertz QCLs. Work along this line is in progress.

Our analysis of the device operation is based on rate equations. One may question the validity of these equations, given that the device relies on strong resonant tunneling between level 1 and 2′ to function. Resonant tunneling may lead to significant coherence between these levels that are not captured by the equations. However, such coherence can be easily destroyed by thermal phonon that leads to fast intraband relaxation. It is estimated that due to LO$_2$ alone the dephasing rate is about twice that of the tunneling rate, hence justifying the neglect of coherent transport and the rate equations. Nonetheless, a full quantum-mechanical treatment of the problem may provide complementary insights.

Finally, we make a few remarks concerning the experimental realization of the device. As aforementioned, our device operates with holes sitting within the neighborhood of the Γ point and is hence robust against misorientation between layers. This does away the need to control the twisting angle between the layers and presents a big advantage when it comes to fabricate the device. Misorientation is normally considered detrimental to vertical van der Waals devices [10,14]. That said, the obvious experimental challenge comes in the growth of the heterostructure with a sufficient number of stages possessing sharp and clean interfaces. Interfacial defects (such as adsorbates and lattice defects) could impact the tunneling and the availability of states participating in lasing. As it stands, there seems to be no cheap way to meet this challenge. Molecular beam epitaxy can grow many stages but the layers may not be sharp enough due to interdiffusion. Mechanical assembly generates atomically sharp structure but can hardly grow as many stages as desired. Yet, the rapid progress in manufacturing low-dimensional materials may well resolve the dilemma in the future [29].

ACKNOWLEDGMENTS

Discussions at the start of this work with Professor Vladimir I. Fal'ko (who also suggested the problem studied in this paper) are gratefully acknowledged.

APPENDIX A: PHONON-INDUCED SCATTERING

Let us consider an N-layer TMD, -$(MX_2)_N$-. As we neglect the effects of vdW forces on the phonon structure, the system supports two acoustic phonon branches LA and ZA by the notation of Ref. [21], and four optical branches LO_1, LO_2, ZO_1, and ZO_2. Reflection symmetry dictates that the coupling of electrons to LO_1 and ZA modes vanish and these modes are subsequently ignored.

A hole can transit from a Bloch state $|n\mathbf{k}\rangle$ to another Bloch state $|n'\mathbf{k}-\mathbf{q}\rangle$ by emission of a phonon in mode $|l\nu-\mathbf{q}\rangle$, where n is the sub-band index, $l = 1, \ldots, N$ labels the layers, $-\mathbf{q}$ is the wave vector of the phonon and ν = LA, LO_2, ZO_1, and ZO_2 labels the phonon branches of an isolated layer. Both $\mathbf{k}$ and $\mathbf{q}$ belong to the 1BZ. The rate at which this process occurs is given by

$$\Gamma^{n'\mathbf{k}-\mathbf{q},l\nu-\mathbf{q}}_{n\mathbf{k}} = \frac{2\pi}{\hbar}|\mathcal{M}^{\nu}_{l}(n\mathbf{k},n'\mathbf{k}-\mathbf{q})|^2\,[1+n_\nu(\mathbf{q})] \times \delta\left[\varepsilon_{n'}(\mathbf{k}-\mathbf{q}) - \varepsilon_n(\mathbf{k}) - \hbar\omega_\nu(-\mathbf{q})\right], \tag{A1}$$

where the δ function ensures energy conservation, $\omega_\nu(\mathbf{q})$ is the phonon frequency, $n_\nu(\mathbf{q}) = \left(e^{\hbar\omega_\nu(\mathbf{q})/k_BT} - 1\right)^{-1}$ is the thermal phonon occupation at temperature T, here k_B is the Boltzmann constant and $\mathcal{M}$ denotes the coupling vertex. The opposite process, in which the hole absorbs the phonon and transits backward, occurs at a rate given by

$$\Gamma^{n\mathbf{k}}_{n'\mathbf{k}-\mathbf{q},l\nu-\mathbf{q}} = \frac{n_\nu(-\mathbf{q})}{1+n_\nu(-\mathbf{q})}\Gamma^{n'\mathbf{k}-\mathbf{q},l\nu-\mathbf{q}}_{n\mathbf{k}}. \tag{A2}$$

The interaction vertex depends on the nature of the electron-phonon interaction. For LA and ZO_2 phonons no dipole is carried and the interaction is essentially short range and of the deformation potential type. For LO_2 and ZO_1 phonons, dipole moments are generated and the interaction is electrostatic. With the standard theory of electron-phonon coupling [12,19,20], one can show that the vertex can be written as

$$\mathcal{M}^{\nu}_{l}(n\mathbf{k},n'\mathbf{k}-\mathbf{q}) = D_\nu F^{l\nu}_{nn'\mathbf{k}}(\mathbf{q})\sqrt{\frac{\hbar}{2\mathcal{A}\rho_\nu\omega_\nu(\mathbf{q})}} \tag{A3}$$

up to a constant that could be either 1 or $\pm i$ depending on ν. Here $\mathcal{A} = WL'$ is the surface area and ρ_ν is the areal mass density of a single layer. The parameters D_ν are given by

$$D_{\mathrm{LA}} = qD,\ D_{\mathrm{LO}_2} = \frac{2\pi Ze^2}{\Omega},\ D_{\mathrm{ZO}_1} = \frac{2\pi Z'e^2}{\Omega},$$
$$D_{\mathrm{ZO}_2} = D_v, \tag{A4}$$

where $q = |\mathbf{q}|$, D and D_v are the deformation potential constants for LA and ZO_2 modes, respectively, and Ze and $Z'e$ are the effective Born charges. In addition, we have $F^{l\mathrm{LA}}_{nn'\mathbf{k}} = F^{l\mathrm{ZO}_2}_{nn'\mathbf{k}} = 1/N$ and

$$F^{l\mathrm{LO}_2}_{nn'\mathbf{k}}(\mathbf{q}) = \sum_{\mu l'}\left(C^{\mu l'}_{n'\mathbf{k}-\mathbf{q}}\right)^* C^{\mu l'}_{n\mathbf{k}}\, e^{-q|z_{l'}-z_l|}, \tag{A5}$$

$$F^{l\mathrm{ZO}_1}_{nn'\mathbf{k}}(\mathbf{q}) = \sum_{\mu l'}\left(C^{\mu l'}_{n'\mathbf{k}-\mathbf{q}}\right)^* C^{\mu l'}_{n\mathbf{k}}\, e^{-q|z_{l'}-z_l|}\,\frac{z_{l'}-z_l}{|z_{l'}-z_l|}, \tag{A6}$$

where $z_l = ld$ is the coordinate of the lth layer, and the $C^{\mu l}_{n\mathbf{k}}$ are the coefficients of the Bloch state $|n\mathbf{k}\rangle$, namely

$$|n\mathbf{k}\rangle = \frac{1}{\sqrt{\mathcal{A}/\Omega}}\sum_{\mu,l} e^{i\mathbf{k}\cdot\mathbf{R}} C^{\mu l}_{n\mathbf{k}}|\mathbf{R}l\mu\rangle,$$

with $|\mathbf{R}l\mu\rangle$ denoting the μth atomic orbital in the lth layer of the unit cell $\mathbf{R}$. As said before, for ZA and LO_1, F vanishes due to reflection symmetry.

In evaluating the rates quoted in the main text, we compute $C^{\mu l}_{n\mathbf{k}}$ by the $\mathbf{k}\cdot\mathbf{p}$ tight-binding model [12], and take $\omega_{\mathrm{LA}} = v_s q$, where v_s is the sound speed, and $\varepsilon_n(k) = \varepsilon_n + \alpha_n k^2$, where ε_n and α_n are the parabolic fitting parameters [see Fig. 2(c)] given by $\varepsilon_1 = 0.44$ eV, $\varepsilon_2 = -0.27$ eV, $\varepsilon_{1'} = 0.42$ eV, $\varepsilon_{2'} = 0.11$ eV, and $\alpha_1 = -4.6$ eV$/k_M^2$, $\alpha_2 = -0.3$ eV$/k_M^2$, $\alpha_{1'} = -4$ eV$/k_M^2$, and $\alpha_{2'} = -3$ eV$/k_M^2$. Dispersion of optical phonon frequencies is neglected [12,19]. In addition, we use $D = 3$ eV while the rest parameters in the vertex, i.e., $\rho_\nu, \omega_{\mathrm{LO}_2}, \omega_{\mathrm{ZO}_{1/2}}, Z, Z'$, and D_v, are taken from Ref. [12]. For MoS_2, $v_s = 6.6 \times 10^3$ ms^{-1}. For $MoSe_2$, $v_s = 4.1 \times 10^3$ ms^{-1}. These values of v_s are taken from Ref. [30]. Room temperature $T = 300$ K is used. One may insert a screening factor $1/(1+r^*q)$ in Eq. (B1), which is ignored in the numerical estimate. Including this factor slightly enhances τ_{21} over $\tau_{2'1'}$ and therefore further corroborates the population inversion.

APPENDIX B: TUNNELING THROUGH A BARRIER

Let us suppose that an InSe monolayer, described by Hamiltonian H_{InSe}, is sandwiched between an N_r-layer TMD sheet to the right and an N_l-layer TMD sheet to the left. The coupling to the left TMD is described by a Hamiltonian $V_l^\dagger$ while that to the right by V_r. No direct coupling exists between the left and the right TMD sheets. By second-order perturbation theory, one can show that the rate at which a hole elastically tunnels from a Bloch state $|n\mathbf{k}\rangle$ on the left to a Bloch state $|m\mathbf{q}\rangle$ on the right, where $\mathbf{k}$ and $\mathbf{q}$ belong to the 1BZ of the left- and right-side TMD sheets, respectively, is given by

$$\gamma_{nm}(\mathbf{k},\mathbf{q}) = \frac{2\pi}{\hbar}|\langle m\mathbf{q}|\mathcal{T}[\varepsilon_n(\mathbf{k})]|n\mathbf{k}\rangle|^2\delta\left[\varepsilon_n(\mathbf{k})-\varepsilon_m(\mathbf{q})\right], \tag{B1}$$

where the effective tunneling matrix $\mathcal{T}$ reads

$$\mathcal{T}(\varepsilon) = V_r^\dagger\,(\varepsilon - H_{\mathrm{InSe}} + i0_+)^{-1}\,V_l. \tag{B2}$$

We assume that the tunneling takes place mainly via the v band of the InSe monolayer for two reasons. Firstly, it lies closest to the sub-bands of the TMD layers of interest. Secondly, it is derived from the atomic orbitals (mostly p_z) that has largest overlap (as represented by the elements of $V_{l/r}$) with the atomic orbitals of the TMD sub-bands of interest. Let $|\mathbf{Q}\rangle$ be the states of the v band and $\varepsilon(\mathbf{Q})$ their energies, where $\mathbf{Q}$ belongs to the 1BZ of the InSe layer. Then

$$\langle m\mathbf{q}|\mathcal{T}(\varepsilon)|n\mathbf{k}\rangle \approx \sum_{\mathbf{Q}}\frac{\langle m\mathbf{q}|V_r^\dagger|\mathbf{Q}\rangle\langle\mathbf{Q}|V_l|n\mathbf{k}\rangle}{\varepsilon-\varepsilon(\mathbf{Q})+i0_+}. \tag{B3}$$

As discussed in the main text, the states participating in lasing action are concentrated about the Γ point, for which the translation symmetry along the layers is approximately respected by $V_{l/r}$. As such, we may take $\langle m\mathbf{q}|V_r^\dagger|\mathbf{Q}\rangle \approx \delta_{\mathbf{q},\mathbf{Q}}t^*_{r,m}$, where $t^*_{r,m} = \langle m\mathbf{q}=0|V_r^\dagger|\mathbf{Q}=0\rangle$ is the hopping integral and δ here denotes the Kronecker symbol. Analogously, we take $\langle\mathbf{Q}|V_l n\mathbf{k}\rangle \approx \delta_{\mathbf{k},\mathbf{Q}}t_{l,n}$, with $t_{l,n} = \langle\mathbf{Q}=0|V_l|n\mathbf{k}=0\rangle$. Now we arrive at

$$\langle m\mathbf{q}|\mathcal{T}(\varepsilon)|n\mathbf{k}\rangle \approx \frac{t^*_{r,m}t_{l,n}}{\Delta}, \tag{B4}$$

where $\Delta = \varepsilon_n(0)-\varepsilon(0)$ is the barrier height for tunneling. Needless to say, the effects of the applied electric field E are implicitly included in these parameters: $t_{r,m}$, $t_{l,n}$, and Δ. In particular, Δ is reduced by the field by an amount of eEd_0, where $d_0 \approx d$ is the distance between the InSe layer and the TMD sheets. For $E = 3\times10^6$ V cm^{-1}, this is about 0.2 eV. DFT computation suggests that $\Delta \approx 1.3$ eV for $E = 0$.

To estimate $t_{r,m}$ and $t_{l,n}$, let us consider the case with $N_r = N_l = 1$. In this case, we may take $t_{m,r} \approx t_{n,l} = t$, where we suppress the indices m and n. Given the similarity of the interface between the InSe monolayer and the TMD layer with that between two InSe layers, we may further assume that t is roughly the hopping integral t' for the latter, which can be obtained from the band splitting of the v bands of a bilayer InSe: the splitting is twice t'. Nevertheless, t should be reduced from t' due to lattice mismatch between the TMD layer and the InSe layer. We take $t \approx t'/r^2 = 0.64t'$, where $r \approx 5:4$ is the ratio of the InSe lattice constant to the TMD lattice constant. DFT computation [13] gives $t' \approx 350$ meV and thus $t \approx 225$ meV.

In the general case of $N_{l/r}$, we expect $t_{r,m}$ and $t_{l,n}$ to be further diminished from t due to the fact that the value of the Bloch wave function at the TMD layer neighboring the InSe layer goes like $\propto 1/\sqrt{N_{l/r}}$. Thus, $t^*_{r,m}t_{l,n} \approx |t|^2/\sqrt{N_lN_r}$. With this we immediately arrive at

$$\gamma_{12'} = \frac{2\pi}{\hbar^2}\frac{1}{N_lN_r}\left(\frac{t^2}{\Delta}\right)^2\rho_{12'}(0). \tag{B5}$$

Here we replace the Dirac function in Eq. (B1) by the spectral overlap function $\rho_{nn'}(\omega)$ defined by a convolution of the spectral functions $A_n(\omega)$ of the level n as follows:

$$\rho_{nn'}(\omega) = \int d\omega' A_n\left(\omega'-\frac{\varepsilon_n}{\hbar}\right)A_{n'}\left(\omega'-\omega-\frac{\varepsilon_{n'}}{\hbar}\right)$$

in order to account for the broadening of the levels n and n' due to phonon scattering. We take $A_n(\omega) = \pi^{-1}\Gamma_n/(\omega^2+\Gamma_n^2)$, where Γ_n is the decay rate of level n. We use $\Gamma_1 = 1/\tau_1$, $\Gamma_{2'} \approx 1/\tau_{2'1'}$ and $\Gamma_2 = 1/\tau_2$. In the estimate of the spontaneous emission time τ_R in the main text, we define $\rho(\omega) \equiv \rho_{12}(\omega)$.

The above formalism can be extended to the case with a bilayer InSe as the tunneling barrier, for which there are two v sub-bands via which tunneling can take place. Both sub-bands contribute but destructively to the tunneling matrix elements, i.e., Eq. (B2). It is easy to show that the tunneling rate is reduced by the factor $(t'/\Delta)^2$ if $t'/\Delta \ll 1$, which is about 5% for InSe.

APPENDIX C: RATE EQUATIONS FOR THE DEVICE

The states taking part in lasing action occupies a tiny fraction approximately 10^{-6} of the 1BZ of the active regions. Let us denote the total number of holes populating these states in sub-band 1/2 by $f_{1/2}(s,t)$, where $s = 1,\ldots,N_s$ labels the stages and t is time, and similar quantities are introduced for sub-bands in the injectors. We write the rate equations as follows:

$$\dot{f}_2(s,t) = \frac{J_{\text{in}}(s,t)}{e} - \left(\frac{1}{\tau_{21}} + \nu_{21}\right) f_2(s,t) - n(t)\nu_{21}\delta f(s,t), \tag{C1}$$

$$\dot{f}_1(s,t) = -\frac{J_{\text{out}}(s,t)}{e} + \left(\frac{1}{\tau_{21}} + \nu_{21}\right) f_2(s,t) + n(t)\nu_{21}\delta f(s,t), \tag{C2}$$

$$\dot{n}(t) = -\frac{n(t)}{\tau} + n(t)\nu_{21}\sum_s \delta f\,(s,t) + \nu_{21}\sum_s f_2(s,t). \tag{C3}$$

Here a dot indicates the derivative to time, $\delta f\,(s,t) = f_2(s,t) - f_1(s,t)$, $J_{\text{in}}(s,t)$ denotes the current injected onto sub-band 2 in stage s and $J_{\text{out}}(s,t)$ is the current that flows out of sub-band 1 in stage s. As $\nu_{21} \ll \tau_{21}^{-1}$, one may neglect the spontaneous emission term in the first two equations. In addition, we have

$$\frac{J_{\text{in}}(s,t)}{e} = \gamma_{1'2}\left[f_{1'}(s-1,t) - f_2(s,t)\right], \tag{C4}$$

$$\frac{J_{\text{out}}(s,t)}{e} = \gamma_{12'}\left[f_1(s,t) - f_{2'}(s,t)\right]. \tag{C5}$$

Finally, we have similar rate equations for the sub-bands of the injectors, which read

$$\dot{f}_{1'}(s,t) = -\frac{J_{\text{in}}(s+1,t)}{e} + \frac{f_{2'}(s,t)}{\tau_{2'1'}}, \tag{C6}$$

$$\dot{f}_{2'}(s,t) = \frac{J_{\text{out}}(s,t)}{e} - \frac{f_{2'}(s,t)}{\tau_{2'1'}}. \tag{C7}$$

Now the above equations form a closed set and allow us to analyze the behaviors of the system.

In steady state cw operation, the quantities are independent of t and s. Thus we simply write $f_1(s,t) = f_1$ and similarly for all other populations, as well as $J_{\text{in}}(s,t) = J_{\text{out}}(s,t) = J$. Note that J is an experimental knob and not decided by the equations themselves. With these one is then led to the equations quoted in the main text.

APPENDIX D: EFFECTS OF SPACE CHARGES

Space charges exist in the device, mostly due to holes piled up on sub-band $1'$ in the injector, totaling $-ef_{1'}$. As the electric field E tends to concentrate the wave functions of this sub-band on the layer with the highest electrostatic potential, one may approximate that the space charges are mostly located on that layer. The corresponding areal charge density on that layer is then $\sigma = -ef_{1'}/WL'$, which generates an additional electric field of strength $E' = 2\pi\sigma$. From the rate equations, we find in the lasing regime that

$$f_{1'} \approx \delta f_{\text{th}}\left(\frac{2}{\gamma_{12'}(\tau_{21} - \tau_{2'1'}) - 1}\frac{J}{J_{\text{th}}} + 1\right).$$

Near the threshold, this yields $f_{1'} \approx 4.3\ \delta f_{\text{th}}$ and then $\sigma \approx -4.3 e d_s \tilde{\tau}_R/\tau = -4.3e(\tau_R/\tau)(1/N_s W L')$. It follows that $E' \approx 10^3$ V cm^{-1} for $\tau = 100$ ps and $N_s = 100$, which is negligible in comparison with E.

[1] R. F. Kazarinov and R. A. Suris, Possibility of the amplification of electromagnetic waves in a semiconductor with a superlattice, Sov. Phys. Semicond. **5**, 707 (1971).

[2] J. Faist, F. Capasso, D. L. Sivco, C. Sirtori, A. L. Hutchinson, and A. Y. Cho, Quantum cascade laser, Science **264**, 553 (1994).

[3] J. Faist, *Quantum Cascade Lasers* (Oxford University Press, Oxford, 2013).

[4] C. Sirtori and R. Tessier, *Intersubband Transitions In Quantum Structures* (edited by R. Paiella, McGraw-Hill, 2006).

[5] M. A. Belkin and F. Capasso, New frontiers in quantum cascade lasers: High performance room temperature terahertz sources, Phys. Scr. **90**, 118002 (2015).

[6] M. S. Vitiello, G. Scalari, B. Williams, and P. D. Natale, Quantum cascade lasers: 20 years of challenges, Opt. Express **23**, 5167 (2015).

[7] C. Jirauschek and T. Kubis, Modeling techniques for quantum cascade lasers, Appl. Phys. Rev. **1**, 011307 (2014).

[8] O. Cathabard, R. Teissier, J. Devenson, J. C. Moreno, and A. N. Baranov, Quantum cascade lasers emitting near 2.6 μm, Appl. Phys. Lett. **96**, 141110 (2010).

[9] D. G. Revin, J. W. Cockburn, M. J. Steer, R. J. Airey, M. Hopkinson, and A. B. Krysa, InGaAs/AlAsSb/InP quantum cascade lasers operating at wavelengths close 3 μm, Appl. Phys. Lett. **90**, 021108 (2007).

[10] K. S. Novoselov, A. Mishchenko, A. Carvalho, and A. H. Castro Neto, 2D materials and van der Waals heterostructures, Science **353**, 6298 (2016).

[11] K. S. Novoselov, A. K. Geim, S. V. Morozov, D. Jiang, Y. Zhang, S. V. Dubonos, I. V. Grigorieva, and A. A. Firsov, Electric field effect in atomically thin carbon films, Science **306**, 666 (2004).

[12] D. A. Ruiz-Tijerina, M. Danovich, C. Yelgel, V. Zolyomi, and V. Fal'ko, Hydrid $\mathbf{k}\cdot\mathbf{p}$ tight-binding model for subbands and infrared intersubband optics in few-layer films of transition-metal dichalcogenides: MoS$_2$, MoSe$_2$, WS$_2$, WSe$_2$, Phys. Rev. B **98**, 035411 (2018).

[13] S. J. Magorrian, A. Ceferino, V. Zolyomi, and V. Fal'ko, Hydrid $\mathbf{k}\cdot\mathbf{p}$ tight-binding model for intersubband optics in atomically thin InSe films, Phys. Rev. B **97**, 165304 (2018).

[14] R. Frisenda, E. Navarro-Moratalla, P. Gant, D. Pérez De Lara, P. Jarillo-Herrero, R. V. Gorbachev, and A. Catellanos-Gomez, Recent progress in the assembly of nanodevices and van der waals heterostructures by deterministic placement of 2D materials, Chem. Soc. Rev. **47**, 53 (2018).

[15] P. Schmidt, F. Vialla, S. Latini, M. Massicotte, K. Tielrooij, S. Mastel, G. Navickaite, M. Danovich, D. A. Ruiz-Tijerina, C. Yelgel, V. Fal'ko, K. S. Thygesen, R. Hillenbrand, and F. H. L. Koppens, Nano-imaging of intersubband transitions in van der waals quantum wells, Nat. Nanotechnol. **13**, 1035 (2018).

[16] S. J. Magorrian and V. Zolyomi, By DFT computation Magorrian and Zolyomi found that the neutrality level of graphene sits close to the Γ point of the lowest conduction band of monolayer InSe (private communication), which allows one to infer Δ from the band structures of TMD layers provided in Refs. [12,13].

[17] The hole arrives near point Γ by radiative transitions, but maybe at other points by nonradiative transitions. It also relaxes toward Γ in the latter case but at a slower pace. Nevertheless, whether this relaxation process is fast or slow is not relevant. Actually, a slow process enhances population inversion between states near the Γ point.

[18] M. Danovich, I. L. Aleiner, N. D. Drummond, and V. I. Fal'ko, Fast relaxation of photo-excited carriers in 2-D transition metal dichalcogenides, IEEE J. Sel. Top. Quantum Electron **23**, 6000105 (2017).

[19] T. Sohier, M. Calandra, and F. Mauri, Two-dimensional Fröhlich interaction in transition-metal dichalcogenide monolayers: Theoretical modeling and first-principles calculations, Phys. Rev. B **94**, 085415 (2016).

[20] J. M. Ziman, *Electrons and Phonons: The Theory of Transport Phenomena In Solids* (Oxford University Press, New York, 2007).

[21] X. Zhang, X.-F. Qiao, W. Shi, J.-B. Wu, D.-S. Jiang, and P.-H. Tan, Phonon and raman scattering of two-dimensional transition metal dichalcogenides from monolayer, multilayer to bulk material, Chem. Soc. Rev. **44**, 2757 (2015).

[22] H. Haken, *Laser Theory* (Springer-Verlag Berlin Heidelberg, Germany, 1984).

[23] G. P. Agrawal and N. K. Dutta, *Long-Wavelength Semiconductor Lasers* (Van Nostrand Reinhold, New York, 1986).

[24] O. Svelto, *Principles of lasers* (Translated and edited by Hanna, D. C., 5th edition, Springer, London, 2010).

[25] Y. Takahashi, H. Hagino, Y. Tanaka, B.-S. Song, T. Asano, and S. Noda, High-Q nanocavity with a 2-ns photon lifetime, Opt. Express **15**, 17206 (2007).

[26] One may also estimate that $\eta \sim \delta f_{\rm th}\Omega/WL' \sim 10^-6$ for $J > J_{\rm th}$, in consistency with that given in the main text.

[27] Here we neglect the broadening due to quantum fluctuations, which set the ultimate lower bound, see Ref. [24].

[28] In Fig. 4, there are four modes falling under the lasing gain profile. The space-independent rate equations, as adopted here, would predict that only the mode with the lowest $J_t h$ could actually be amplified. However, if spatial dependence and inhomogeneous line broadening (band dispersion near Γ) are taken into account, all four modes can be amplified, see Ref. [24].

[29] T. Chowdhury, E. C. Sadler, and T. J. Kempa, Progress and prospects in transition-metal dichalcogenide research beyond 2D, Chem. Rev. **120**, 12563 (2020).

[30] S. Shree, M. Semina, C. Robert, B. Han, T. Amand, A. Balocchi, M. Manca, E. Courtade, X. Marie, T. Taniguchi, K. Watanabe, M. M. Glazov, and B. Urbaszek, Observation of exciton-phonon coupling in MoSe$_2$ monolayers, Phys. Rev. B **98**, 035302 (2018).