The Fable Computer, Part II: Quantum-Limited Analog Tensor Processing on the Regenerative Graphene-Plasmon Fabric

Interference computing with mesoscopic plasmon pulses, classical on-chip decoding to 1–2 bits, and the temperature budget of the quantum–classical crossover

Ryoji Furui

www.ryoji.info

July 2026 · companion to The Fable Computer [1]

Authorship and Status Disclaimer. This manuscript extends the author's design study [1] ("Part I" throughout). Generation of content involved the use of an AI, Anthropic's Fable 5, at the author's direction, including construction and adversarial cross-checking of the released model package (Appendix QA). The author declares a lack of formal expertise in quantum optics and graphene-plasmonic device physics, and the document has not received independent expert review. All claims are reduced-order model results in the sense of Part I ("in-model": established within the released chain, under its stated assumptions, not measured), and must be treated as a starting point for discussion, not as peer-reviewed engineering guidance.

Abstract

Part I designed a room-temperature terahertz half adder on a regenerative graphene-plasmon fabric and answered its feasibility question in-model. This Part asks what the same fabric computes when its pulses are treated as what they are: quantum fields. Quantizing the Part-I Dyakonov–Shur (DS) cell mode puts one plasmon at a fractional density amplitude of 0.16 %, the 1-dB compression knee at ≈38 plasmons and the logic rail at ≈3.8×10^3 — logic pulses are mesoscopic, and quantum noise is a quantitative design input, not a curiosity. Two results follow. First, an honest no-go: the single-plasmon Kerr nonlinearity of the 576-nm cell falls short of its loss rate by five orders of magnitude ($\chi/\kappa \approx 8.2\times10^{-6}$ at 4 K), so gate-model (photon-blockade) quantum logic is excluded on this fabric at any temperature; the blockade would demand ~3-nm cells, consistent with the quantum-plasmonics literature. Second, what survives is exactly the machine the fabric's own architecture suggests: a quantum-limited analog (continuous-variable) computer, in which comb-locked coherent pulses interfere at gate-programmed junctions — plasmon–plasmon interaction in the analog-computer sense — and the result is quantized on-chip by the Part-I threshold cells themselves, re-programmed as a 1–2-bit flash decoder. The minimal tensor unit, QMAC-1 (two inputs, one interference, one output decoded to 2 bits), is specified to cell-map level; in its digital special case the decoded word is the binary count A+B, so the Part-I half adder re-emerges as the decoder of a single quantum interference. Error rates versus temperature are computed on one axis for the quantum-analog unit and the classical digital fabric: the 2-bit decode is unusable per-shot at 300 K (0.14), reaches 3.1×10^{-3} at 77 K, and saturates at a vacuum-set floor of 1.3×10^{-6} below ≈20 K (N = 400 plasmons per input); the 1-bit decode is already usable warm (3.9×10^{-3} at 300 K) — lower-bit decoding is what the physics supports per shot. $T_Q = \hbar\omega/k_B \approx 48$ K is where the noise budget crosses from thermal- to vacuum-dominated; below ≈15–20 K the error curves saturate at vacuum-set floors that no further cooling improves. That saturation is the quantum limit made visible, and its observation is a pre-registered bench gate. Part I's no-cryogenics lock is explicitly relaxed here and the fork is stated as such.

1. Introduction: from restored bits to quantum-limited amplitudes

Part I [1] built a machine out of one physical object — a resonant DS cell [2] biased below its self-oscillation threshold — and one discipline: every gate restores its logic levels against a rail, so decisions never erode. The price of that discipline is nonlinearity: digital restoration lives at the saturating rail. This Part starts from the opposite end of the same transfer curve. Below the 1-dB knee the fabric is a linear, phase-coherent wave medium: pulses split, merge, and interfere at gate-programmed junctions, and the comb-locked clock of Part I — distributed at femtosecond-class phase — gives every cell a common phase reference. A linear interferometer with programmable weights is an analog computer: it multiplies by transmission amplitudes and adds by superposition. The question Part II asks is how well that analog computer can compute, and the answer is a quantum one, because the pulses in play are small.

Quantizing the Part-I operating point (Section 3) puts the knee of a cell at ≈38 plasmon quanta and the rail at ≈3833. At those occupations three quantum facts dominate the machine design. The 1-THz quantum is small against the room-temperature bath ($\hbar\omega/k_B \approx 48$ K), so at 300 K every mode carries ≈5.8 thermal plasmons of noise; below ≈48 K vacuum fluctuations take over, and by ≈15–20 K the noise floor stops improving altogether. Phase-insensitive gain — the only kind the DS cell provides [2, 3] — adds at least the Caves quantum noise, which is precisely the $F = 2 - 1/G$ floor Part I already carried as its Eq. (7). And amplitude resolution per shot is bounded by shot noise, so the number of bits one pulse can carry is finite and computable. The machine that emerges is not a quantum computer in the gate-model sense — Section 3.3 quantifies why that is excluded on this cell — but a quantum-limited analog tensor engine with a classical digital wrapper: the same fabric, the same five-cell block as Part I, re-programmed so that two of its cells act as comparators of a flash analog-to-digital ladder.

One constraint of Part I is explicitly relaxed: no-cryogenics. The quantum-analog unit works at 300 K only in its 1-bit or averaged modes; its 2-bit per-shot mode wants ≤77 K, and its quantum floor lives below ≈20 K (Section 6). Part II therefore defines temperature classes rather than one operating point, keeps the ≤80 °C cap for the warm class, and states plainly that the cold classes are a fork from Part I's charter, not a refinement of it. Cooling also changes the classical fabric itself (Part I, Section 8.3): the DS threshold tracks temperature, so all cold-side operation inherits Part I's tracked-bias requirement, and below ≈250 K the hydrodynamic tier that certifies the gain numbers expires — every cold-side gain figure here carries that caveat to the Boltzmann–Maxwell tier, exactly as in Part I.

2. Related work and positioning

Quantum plasmonics. That surface plasmons preserve quantum coherence is established: entangled photons survive conversion through plasmonic channels [6], and the field has a developed single-quantum toolbox [5]. For graphene specifically, the extreme mode confinement prompted the prediction that few-nanometre graphene nanostructures could reach single-plasmon Kerr nonlinearities of order unity [4] — the photon-blockade regime. Section 3.3 places the Part-I cell against exactly that scaling and finds it five orders of magnitude away: this design claims the opposite corner of quantum plasmonics, many-quantum Gaussian operation at the standard quantum limit.

Continuous-variable computing. Encoding analog values in field quadratures and processing them with Gaussian operations is the continuous-variable framework [9, 10]; universal quantum computation in that framework is known to require a non-Gaussian resource [9], which is precisely what the Kerr no-go denies this fabric. What Gaussian operations do provide is quantum-limited analog arithmetic: the amplifier bound of Caves [7] (anticipated by Haus and Mullen [8]) sets the noise cost of gain, and homodyne-class detection against vacuum noise sets the per-shot precision.

Photonic analog accelerators. Coherent interferometric multiply-accumulate is the operating principle of photonic neural accelerators, from programmable nanophotonic processors [11] to comb-based convolutional tensor cores [12]. The niche claimed here is different in three ways: the carrier is a THz plasmon confined ≈100× below the free-space wavelength (footprint), the weights and routing are gate-voltage-programmable on one chip, and — the load-bearing difference — the analog result is quantized by the same regenerative cell family that computes, so no electronic ADC enters the signal path (Section 5). The cost is the THz quantum: at 1 THz the thermal bath is hot in quantum units at room temperature, which is why Section 6's temperature budget is the central result of this Part.

3. Quantum scales of the Part-I operating point

3.1 How many quanta is a logic pulse?

The hydrodynamic energy of the acoustic (gate-screened) standing mode [15] of one DS cell (length L ≈ 576 nm, effective area A ≈ 10^{-8} cm^2, density n = 10^{12} cm^{-2}, speed s from Part-I Eq. (1)) with fractional density amplitude ε = δn/n is

$$E(\varepsilon) = \tfrac{1}{4}\, n\, m^* s^2 \varepsilon^2 A, \quad m^* = E\_F/v\_F^2 \qquad \text{(Q1)}$$

(kinetic plus potential of the shallow-water mode, spatially averaged over the standing wave). Setting E = $\hbar\omega_0$ defines the single-plasmon amplitude and the quantum content of any classical swing:

$$N(\varepsilon) = (\varepsilon/\varepsilon_1)^2, \quad \varepsilon_1 = \sqrt{(4\hbar\omega_0 / n\, m^* s^2 A)} \approx 1.62\times10^{-3} \qquad \text{(Q2)}$$

One plasmon is a 0.16 % density swing. The Part-I solver's 1-dB gain-compression knee (ε ≈ 1 %) is therefore N ≈ 38 quanta (0.025 aJ intracavity), and the ≈10 % rail is N ≈ 3833 (2.5 aJ). Two scales should not be confused. Part I's “intrinsic field energy of ≈10 aJ per bit” counts a traveling few-cycle pulse over its ≈3λ_p in-flight extent (an area an order of magnitude larger than one cell); Eqs. (Q1–Q2) count quanta in the quarter-wave cavity mode of one cell, which is the mode the noise physics and the comparator decisions live in. The two are consistent — a traveling pulse at rail-class swing over 3λ_p carries 10–30 aJ — and both sit below the launch-side slot energy before the still-unquantified coupling (Part I, Section 9.5). A logic pulse on this fabric is a mesoscopic object — tens to thousands of quanta — and Figure Q1(A) shows the whole ladder.

3.2 The bath: 48 K in quantum units

The 1-THz quantum is $\hbar\omega_0$ ≈ 4.14 meV, so the crossover temperature is T_Q = $\hbar\omega_0$/k_B ≈ 48 K. Table Q1 gives the Bose–Einstein occupation of the mode across the temperatures used in this Part. At the Part-I band (300–353 K) every mode carries six to seven thermal quanta — the

fabric computes while swimming in a bath an order of magnitude above the vacuum floor; at 77 K one thermal quantum remains; at 20 K the mode is effectively dark. The plasmon lifetime improves in parallel but saturates: $\tau(T)$ follows Part I's $1/T$ phonon scaling down to ≈150 K and is clamped by impurity scattering below (Part I, Section 8.3), so Q saturates at ≈12.6. The saturated τ = 2 ps sits slightly above the best measured cryogenic graphene-plasmon lifetimes (≈1.6 ps [13]) — a premium-material assumption in exactly the sense of Part I's $\tau(300\ K)$ = 1 ps, and flagged as such. Cooling buys a quieter bath much faster than it buys a better resonator.

T (K)	$\bar{n}$ (thermal quanta)	$Q = \omega\tau$	per-gate loss (dB, $\lambda_p/2$)
353	6.87	5.3	2.56
300	5.76	6.3	2.17
150	2.65	12.6	1.09
77	1.16	12.6 (sat.)	1.09
48 (= T_Q)	0.58	12.6 (sat.)	1.09
20	0.100	12.6 (sat.)	1.09
4	6.2×10^−6	12.6 (sat.)	1.09

Table Q1. The 1-THz mode against its bath. Lifetime saturation below ≈150 K follows Part I's stated cold-side caveat; premium (graphite-gated) material is carried as upside, not assumed.

3.3 The honest no-go: no photon blockade on this cell

Gate-model quantum logic needs one quantum to shift the cell's resonance by more than a linewidth — $\chi/\kappa \gtrsim 1$, with χ the single-plasmon Kerr rate and $\kappa = 1/\tau$ the decay rate. The hydrodynamic nonlinearity gives, at reduced order,

$$\chi \approx c_K\, \omega_0\, \varepsilon_1^{\,2}, \quad c_K = O(1)\ \text{band}\ 0.05\text{–}1\ (0.25\ \text{carried}) \qquad \text{(Q3)}$$

i.e. $\chi/2\pi \approx 0.7$ MHz against $\kappa/2\pi \approx 160$ GHz at 300 K: $\chi/\kappa \approx 4.1\times10^{-6}$ at 300 K and 8.2×10^{-6} even at the lifetime-saturated 4 K — five orders of magnitude short, at any temperature this stack can reach. Because $\chi \propto 1/A$ at fixed density and frequency, closing the gap would demand ≈3 nm structures — a bookkeeping scale, not a device: no few-nanometre cell hosts a 1-THz quarter-wave mode at all (its resonance would sit near 200 THz), so blockade is excluded here twice over, by magnitude and by geometry. Both exclusions are consistent with the graphene quantum-plasmonics literature, whose blockade proposals live at few-nanometre scale and correspondingly mid-infrared frequencies [4]. The conclusion is stated plainly: this fabric supports no plasmonic qubit. The mesoscopic cross-Kerr interaction is real but sub-resolution per shot — a 400-quantum control pulse writes ≈6.6 mrad of phase in 2 ps, versus ≈25 mrad single-shot homodyne resolution at 4 K (≈15 shots to resolve): measurable as a systematic with averaging (bench gate QG5), not computational per slot. Everything that follows is therefore Gaussian: coherent states, linear optics, phase-insensitive gain, and homodyne-class decisions — quantum-limited analog computing, with the limits computed rather than assumed.

Figure Q1. Quantum scales of the plasmon fabric (operating point of Part I)

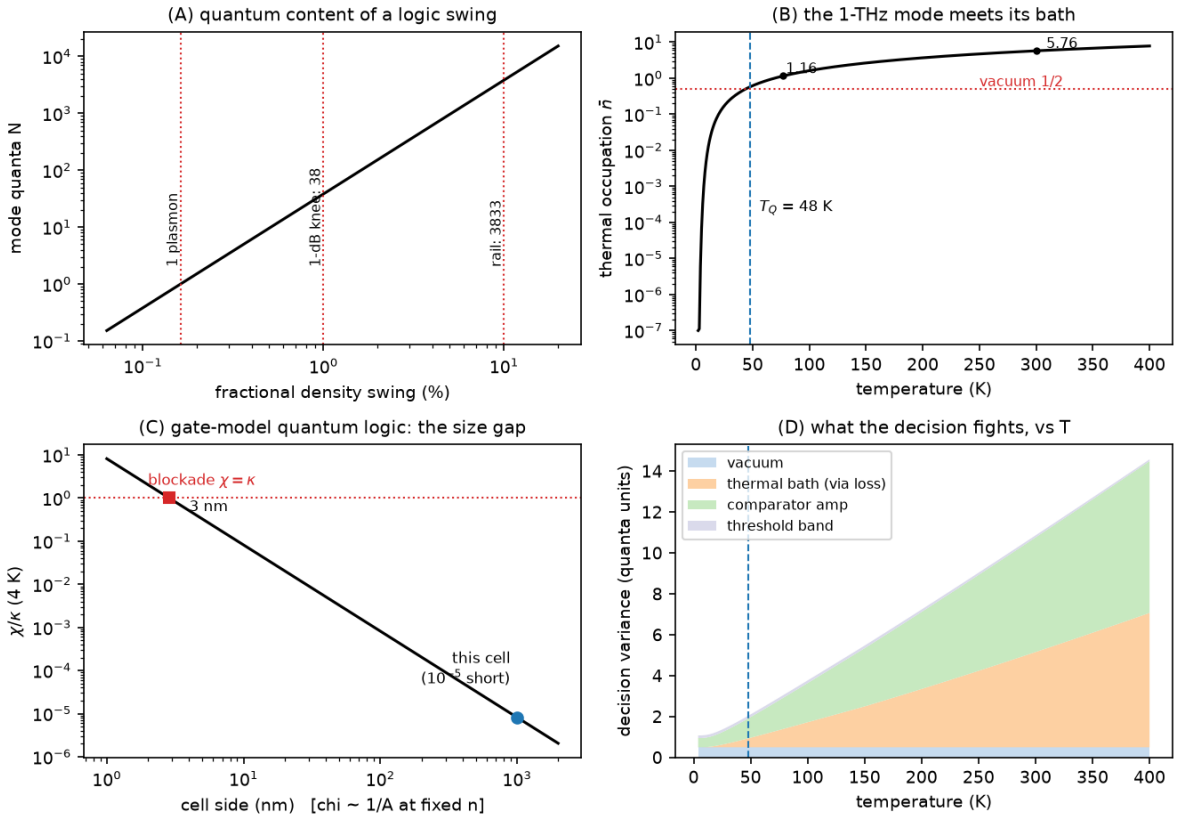

Figure Q1. Quantum scales of the fabric at the Part-I operating point. (A) Quanta in a classical swing, Eq. (Q2): one plasmon, the 38-quantum knee, the 3.8×10^3-quantum rail. (B) Thermal occupation of the 1-THz mode: T_Q ≈ 48 K separates the thermal and vacuum regimes. (C) The blockade gap: χ/κ versus cell size at fixed density; the Part-I cell sits five orders below blockade, which would demand ≈3-nm cells [4]. (D) Decision-noise budget versus temperature: vacuum, thermal bath, comparator amplifier noise, and threshold-band terms (Section 5).

4. The computing model: quantum-limited interference

4.1 Gaussian rules on the fabric

Pulses are comb-locked coherent states; in the measured quadrature x (vacuum variance ½) a pulse of N quanta has mean $\sqrt{(2N)}$. Three operations exhaust the analog layer, each with an exact Gaussian rule: propagation loss η at temperature T mixes in the bath, $V \to \eta V + (1-\eta)(\bar{n}+½)$; a junction combiner with amplitude weights $(\sqrt{w}, \sqrt{(1-w)})$ superposes means and averages variances; and regenerative gain G adds the matched-temperature phase-insensitive amplifier noise, $V \to GV + (G-1)(\bar{n}+½)$. The last rule is Part I's noise model made explicit: referred to a matched input floor it reproduces $F = 2 - 1/G$ (Part-I Eq. (7)) at every temperature, and reduces to the Caves bound [7] as $T \to 0$. All three rules are verified in the released package against exact Fock-space quantum evolution (truncated-basis Lindblad for loss, two-mode squeezing for gain) to better than 0.01 % (Appendix QA). One deliberate conservatism: decisions are modeled as midpoint-threshold homodyne discrimination — what the comparator cells physically implement — rather than the optimal joint quantum measurement, whose

Helstrom bound [16] would improve the error exponent by at most a factor of two for these level sets.

Weights are programmable the way everything on this fabric is programmable: a junction's split ratio is a gate-map choice, and a weight's sign (phase) is a programmed delay — the plasmon speed under a segment is density-tuned through Part-I Eq. (1), so a half-wave of extra path is a voltage, not a mask change. The interference core therefore implements $y = \Sigma\, w_i\, x_i$ with amplitude weights of either sign: a full four-quadrant multiply-accumulate.

4.2 Two computed design rules

The analog core must be passive. The regenerative cell's linear window ends at the ≈38-quantum knee, so any pre-amplifier caps the analog signal exactly where thermal noise is most damaging. Computed head-to-head (Appendix QA): the pre-amplified variant of the tensor unit delivers 0.50 symbol error at 300 K and 0.10 even at 4 K — against 0.14 and 1.3×10^{-6} for the passive core. Phase-insensitive gain before quantization loses at every temperature on this fabric; gain belongs after the decision, inside the decoder cells, which are threshold devices and want the rail anyway. This is the quantum restatement of Part I's own discipline: restore digitally, never amplify analog.

The launch must be quiet. A launch carrying Part I's budgeted 20-dB classical SNR (modeled as amplitude noise at a fixed fraction of the full swing, on both arms) pins the decode error near 3.4×10^{-3}–1.9×10^{-2} at every temperature — classical launch noise does not cool away (Figure Q3, grey curve). Quantum-limited operation requires a near-shot-noise-limited launch, which the photonic chain of Part I does not currently promise. This is Part II's sharpest new hardware requirement and is carried as an open item alongside Part I's launch-coupling efficiency.

5. QMAC-1: the minimal tensor unit, decoded to 2 bits

The smallest complete tensor operation is one weighted sum of two inputs, quantized. QMAC-1 (Figure Q2) is that unit on the fabric: two comb-locked input pulses (N_{op} quanta each at logic 1; $N_{op} = 400$ default, launch-budget-capped) pass gate-programmed weight segments, interfere in one junction combiner, and the output amplitude is split to two comparator cells — Part-I threshold cells with staggered, trimmed thresholds and sharpness $k \geq 16$ — whose thermometer code is mapped to binary by one inverting cell and one AND cell. Five active cells, ≈12×6 lattice sites (≈13×14 μm at the Part-I site pitch, Figure Q5), one MAC per 4-ps slot, wave-pipelined like everything on this fabric.

The decoder is not new hardware. Comparators, inverter, AND: these are the Part-I gate vocabulary, and the block is cell-for-cell the Part-I half adder with its input stage re-trimmed as a two-level flash ladder. In the digital special case (both weights ½, inputs ∈ {0, N_{op}}) the interference levels are 0, $N_{op}/2$ and $2N_{op}$ — equally spaced in amplitude — and the decoded word b_1b_0 is the binary count A+B: sum and carry (Table Q2). The Part-I half adder re-emerges as the decoder of a single quantum interference — the demonstrator of Part I and the decoder of Part II are the same five cells, re-programmed.

A	**B**	**interference level (quanta)**	**$t_2\ t_1$**	**$b_1\ b_0$**	**= A+B**

0	0	0	0 0	0 0	0
0	1	N_op/2	0 1	0 1	1
1	0	N_op/2	0 1	0 1	1
1	1	2 N_op	1 1	1 0	2

Table Q2. QMAC-1 digital truth table: the 2-bit decode of one interference is the half adder of Part I (b_1 = carry, b_0 = sum). Verified digitally in the released package.

In analog mode the same unit computes y = √w·x_A + √(1−w)·x_B and the ladder quantizes y to 2 bits. The per-shot analog precision — effective bits of one MAC against the full noise budget — is 0.6 bits at 300 K, 2.0 at T_Q and 2.5 at 4 K (Figure Q4A): the physics itself says this fabric is a 1–2-bit-per-shot analog machine, which is why the decoder is sized at 2 bits and why lower-bit decoding is not a concession but the design point. A worked weighted example (w = 0.7, 5×5 input grid, 77 K) holds worst-case rms noise at 0.36 LSB of the 2-bit ladder (Appendix QA). Deeper precision is bought with slots: accumulating 16 slots in the electronic charge-memory layer (Part I, Section 6.4) averages the per-slot noise — though not the static threshold band, which is common to all slots — and gains five decades at the Part-I band (Section 6) at 1/16th throughput, still 1.6×10^{10} MAC/s per unit.

Figure Q2. The minimal quantum-analog tensor unit and its 2-bit decode (N_{op} = 400 quanta)

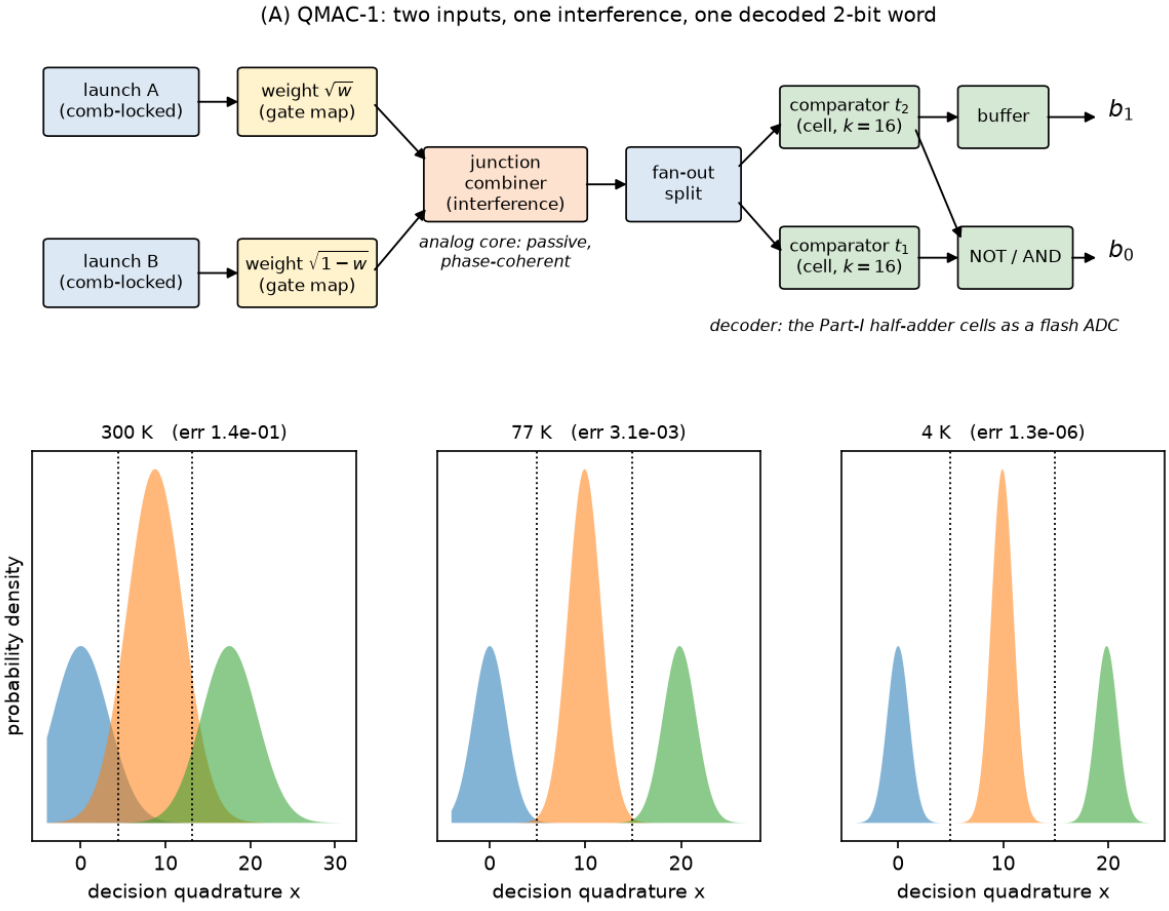

Figure Q2. QMAC-1. (A) Cell-level netlist: comb-locked launches, gate-programmed weights, one junction combiner (the interference that computes), fan-out split, two comparator cells, and the thermometer-to-binary cells; the decoder is the Part-I half-adder block re-programmed. (B) The three interference levels and their Gaussian noise at 300, 77 and 4 K with midpoint thresholds (N_op = 400): the same levels drown at 300 K, separate at 77 K, and sit an order of magnitude apart at the vacuum floor.

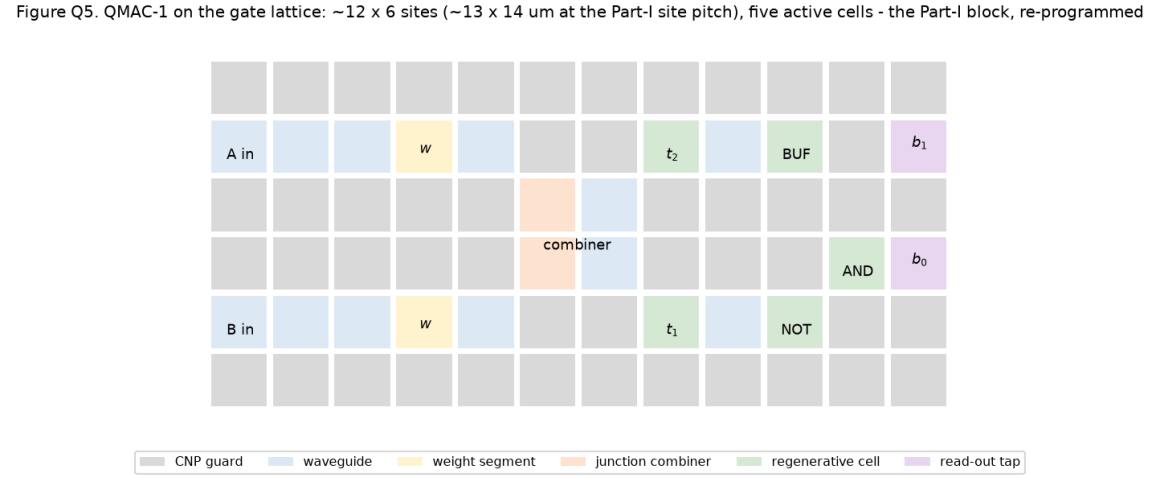

Figure Q5. QMAC-1 on the gate lattice (Part-I Figure 5b convention): ≈12×6 sites including guards, five active cells, weights and combiner as programmed junction segments, read-out taps at the boundary. The map is a re-programming of the Part-I half-adder block, not a new tile.

6. Error rates versus temperature

Figure Q3 and Table Q3 put the quantum-analog unit and the classical digital fabric on one temperature axis, 4–400 K, holding the hardware fixed and letting only $\bar{n}(T)$, $\tau(T)$ and the loss/gain budget move. The classical-digital curve is this package's own construction (Part I never quotes a bit-error rate): a rail-restored decision — swing at the ≈10 % rail (≈3833 quanta), Part I's static noise margin (0.256 of swing), against the matched-temperature amplifier noise of one restoring cell. It is the physics floor of the digital fabric, with technical noise excluded on both sides of the comparison.

T (K)	$\bar{n}$	classical digital BER	quantum-analog 2-bit	quantum-analog 1-bit	2-bit, 16-slot avg	analog bits/shot
353	6.87	9.1×10^−10	1.8×10^−1	9.6×10^−3	6.5×10^−6	0.43
300	5.76	3.6×10^−11	1.4×10^−1	3.9×10^−3	4.7×10^−7	0.62
150	2.65	2.4×10^−20	2.5×10^−2	1.1×10^−5	2.2×10^−13	1.29
77	1.16	5.7×10^−37	3.1×10^−3	4.8×10^−9	2.6×10^−19	1.73
48	0.58	1.8×10^−55	3.9×10^−4	1.9×10^−12	1.6×10^−23	2.00
20	0.10	2.3×10^−98	6.5×10^−6	2.9×10^−19	5.3×10^−29	2.36
4	0.00	<10^−99	1.3×10^−6	6.3×10^−22	1.7×10^−30	2.46

Table Q3. Error per operation versus temperature (N_op = 400 quanta per input; classical digital at the rail). Every column is recomputed by run_all.py in the released package.

Reading the table downward: at the Part-I band the classical digital fabric is essentially error-free in this idealization (3.6×10^−11 at 300 K) while the 2-bit quantum-analog decode is unusable per shot (0.14) — digital restoration wins at room temperature by restoring at ≈10× the analog unit's per-input amplitude and giving up linearity. The quantized noise floor turns that from a choice into a requirement: had the digital fabric restored its levels at the 1-dB knee (≈38 quanta) instead of the rail, its decision would be thermal-noise-limited to $\approx 10^{-1}$ BER at 300 K — a constraint on Part I that only becomes visible once the mode is quantized, and one its 20-dB launch-SNR budget (≥600 quanta over the thermal floor) already implies. The 1-bit decode, spending the full decode range on one boundary, is already serviceable warm (3.9×10^−3 at 300 K, 9.6×10^−3 at the 80 °C cap) — the quantitative content of "lower-bit decoding is acceptable". Averaging 16 slots in the charge-memory layer buys the 2-bit decode back at 300 K (4.7×10^−7). Per-shot 2-bit operation is a cold-class mode: error 10^{-3} needs ≤59 K; the 1-bit decode reaches 10^{-6} by 117 K and 10^{-9} by 69 K. All of these assume the optimistic end (−1 dB) of Part I's −1 to −3 dB junction band; at −3 dB junctions the 77 K 2-bit error degrades to 5.6×10^−2 and the 77 K 1-bit error to 1.8×10^−4 — junction quality is a first-order lever on every cold-class number, and the sensitivity is reported alongside the defaults in the released results.

The temperature structure has two landmarks, and they must not be conflated. $T_Q \approx 48$ K is where the noise density crosses from thermal- to vacuum-dominated ($\bar{n} = 0.58$ there — still comparable to the vacuum ½). The error curves, being erfc tails, keep falling steeply until $\bar{n} \ll ½$ and saturate at their vacuum-set floors below ≈15–20 K: 1.3×10^−6 (2-bit) and 6.3×10^−22 (1-bit) at $N_{op} = 400$, the 2-bit floor falling to 1.3×10^−10 at $N_{op} = 800$ (Figure Q4B). The floor's value is an engineering choice of pulse energy, but its temperature-independence is not: a noise floor that flattens through the T_Q knee and stops improving below ≈15 K is the standard quantum limit made visible on a graphene chip, and Section 8 makes its observation a pre-registered gate. The grey curve is the cautionary tale: with a 20-dB-SNR classical launch, no temperature helps.

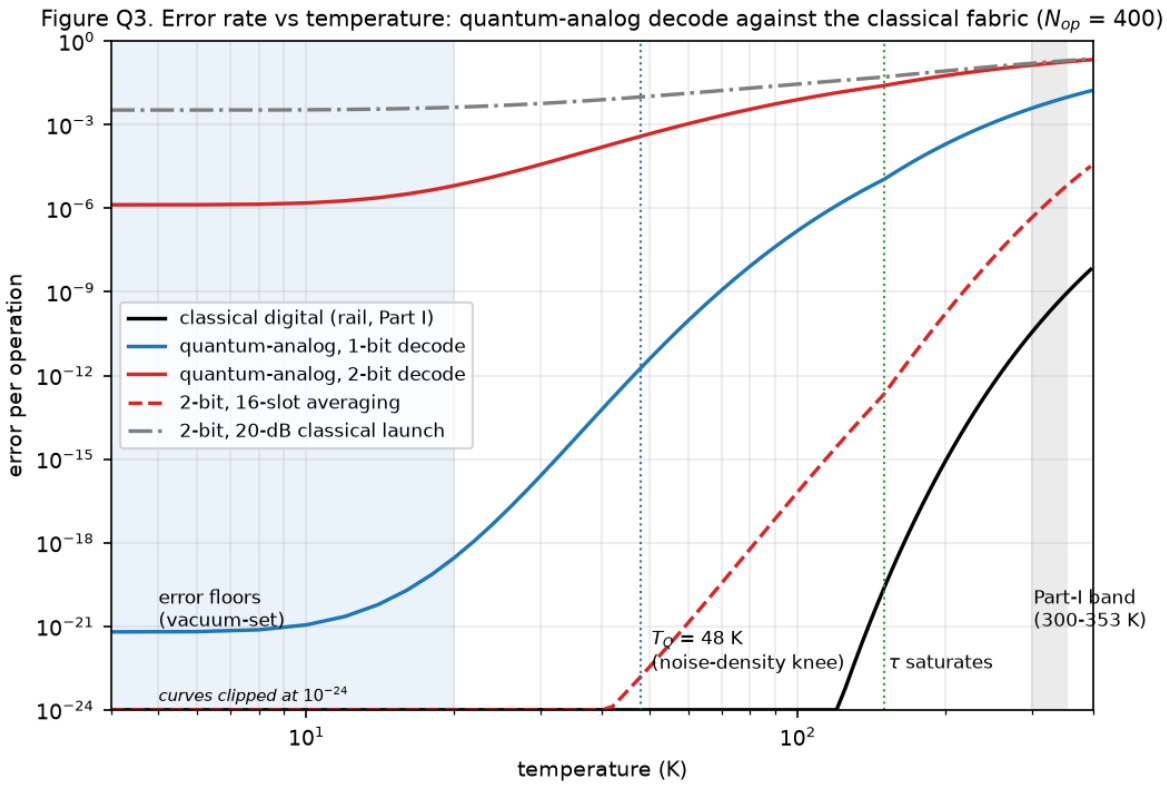

Figure Q3. Error per operation versus temperature (N_op = 400): classical digital fabric at the rail (black), quantum-analog 1-bit (blue) and 2-bit (red) decodes, 2-bit with 16-slot averaging (dashed), and the 20-dB classical-launch wall (grey). Markers: T_Q = 48 K (the noise-density knee), 150 K (lifetime saturation), the Part-I 300–353 K band (shaded), and the sub-20 K region where the quantum-analog curves sit on their vacuum-set floors — the quantum limit as an observable.

Figure Q4. The analog precision budget

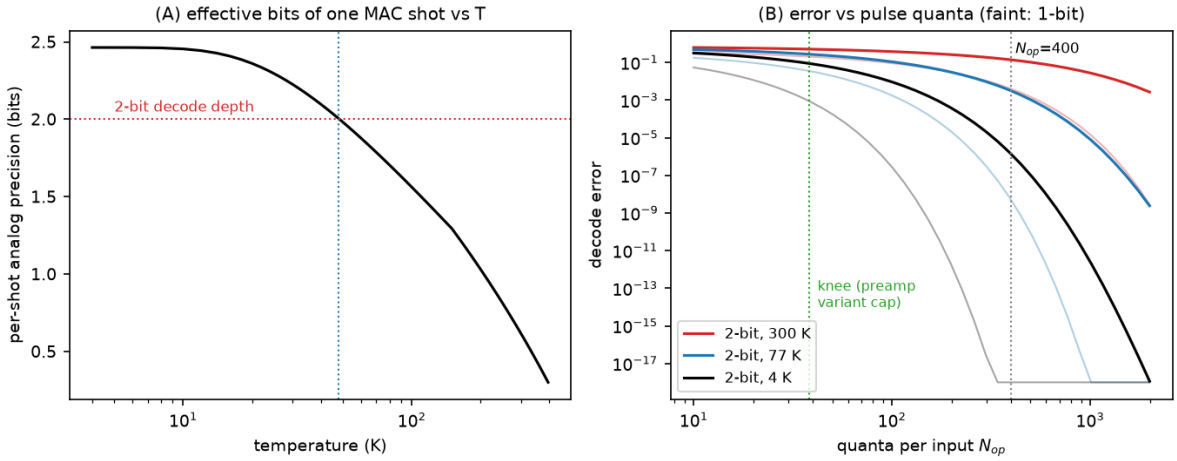

Figure Q4. The precision budget. (A) Per-shot analog precision of one MAC versus temperature: ≈0.6 bits at 300 K, 2 bits at T_Q, 2.5 bits at the floor — the physics that sizes the decoder at 1–2 bits. (B) Decode error versus quanta per input at 300/77/4 K (faint: 1-bit): the knee (38 quanta) caps any pre-amplified variant; the passive core scales until the launch budget, with the 2-bit floor reaching 1.3×10^−10 at N_op = 800.

7. What is, and is not, quantum here

Is: the noise. Below T_Q the decision noise becomes vacuum-dominated — the machine computes against $\hbar\omega/2$ per mode rather than a thermal bath — and from ≈15–20 K down its error floors are set by quantum mechanics (vacuum variance, the Caves cost of the decoder's gain), not by temperature. The per-shot information of one pulse is shot-noise-bounded (Figure Q4A). These are quantitative, falsifiable quantum statements, and gate QG2 tests the sharpest of them.

Is not: the computation. Coherent states, beamsplitters, phase-insensitive amplifiers and homodyne decisions are all Gaussian; a Gaussian machine with coherent-state inputs is efficiently classically simulable, and universal continuous-variable quantum computation requires exactly the non-Gaussian resource [9, 10] that Section 3.3 shows this cell cannot supply ($\chi/\kappa \approx 10^{-5}$). No entanglement is generated or claimed anywhere in QMAC-1. The honest name for this machine is quantum-limited analog computing: classical arithmetic executed at the precision ceiling quantum mechanics assigns to unentangled light.

The upgrade paths are named, with their prices. (i) Blockade cells: ≈3-nm structures at $\chi/\kappa \approx 1$ [4] — a different lithography universe, and lifetime-limited even then. (ii) Squeezed operation: a DS cell driven parametrically (bias modulated at $2f_0$) would be a phase-sensitive amplifier, which evades the $F = 2 - 1/G$ floor and could push decisions below the standard quantum limit; graphene plasmon squeezing has no experimental precedent and is listed as an open item, not an ingredient. (iii) Entanglement distribution across the fabric: plasmonic channels preserve entanglement at optical frequencies [6]; nothing comparable exists at THz. None of these is assumed by any number in this Part.

8. Falsifiability: the bench gates

In Part I's discipline, every claim above reduces to pre-registered pass/fail gates on the same class of device the Part-I protocol already specifies, plus the quantum-specific read-out chain. The five gates are listed in Table Q4; QG2 is the existence proof of the quantum regime and the single most informative outcome, pass or fail.

Gate	Observable	Pass threshold	Model target
QG1	two-pulse interference visibility at a junction combiner (comb-locked launches)	≥ 90 %	≥ 99 % (phase budget 42 fs vs 1.4 fs demonstrated lock)
QG2	read-out noise floor vs temperature, 300 → 4 K	departs from the classical ∝ T line by ≥ 40 % at 20 K and flattens below ≈15 K	floor $\bar{n} + ½ \rightarrow ½$: knee at T_Q ≈ 48 K, within 10 % of vacuum below ≈16 K
QG3	QMAC-1 2-bit symbol error at 77 K, $N_{op} \approx 400$	$\leq 10^{-2}$	3.1×10^−3 at −1 dB junctions; 5.6×10^−2 at −3 dB
QG4	1-bit (full-range) decode error at 300 K	$\leq 10^{-2}$	3.9×10^−3
QG5 (stretch)	cross-Kerr phase from a 400-quantum control pulse, averaged	bound $\chi/2\pi$ within the 0.13–2.6 MHz band	6.6 mrad per 2 ps, ≈15 shots to resolve

Table Q4. Pre-registered gates for the quantum extension. QG1 and QG2 precede the tensor unit: without visibility there is no analog computer, and without the noise-floor saturation there is no quantum regime to claim. QG3 is deliberately a joint test of the decode chain and the −1 dB junction budget — at −3 dB junctions the model itself fails the gate, so the Part-I junction-variant die gates QG3. All gates assume the Part-I calibration sequence (σ, τ(T), s, M_th, coupling) has run first.

Two dependencies are inherited from Part I and stand in front of everything here: the launch-coupling efficiency into the acoustic mode (unquantified there — photoconductive launch into gated graphene plasmons is demonstrated only at 4 K [14] — and Part II additionally requires the launch to be near shot-noise-limited), and the Boltzmann–Maxwell tier, which now also owns the Kerr coefficient c_K, the absolute noise calibration, and every gain number below the ≈250 K hydrodynamic validity line.

9. Conclusion

Part I asked whether room-temperature graphene can gain back what a plasmon loses; this Part asked what the same fabric is worth as a computer when its pulses are counted in quanta. The answers are symmetric. Downward, to the single quantum: nothing — the Kerr gap is five orders of magnitude, and the fabric will host no qubit. Sideways, at tens to thousands of quanta: a genuine machine — a phase-coherent interference core that multiplies and adds in one junction, decoded to the 1–2 bits per shot that shot noise actually affords, by the same five cells Part I used to add. The temperature budget is the honest price list: 1-bit decoding or 16-slot averaging at the Part-I band with no cryogenics, per-shot 2-bit operation from ≈60 K down, and below ≈20 K a floor that cooling cannot improve — which is not a defect but the standard quantum limit, and the most falsifiable statement this Part makes. The model chain is released, the gates are pre-registered, and the design is now exactly as falsifiable as Part I was built to be.

Appendix QA. The quantum model chain: conventions and released code

The released package fable-model-quantum mirrors Part I's fable-model-chain: pure Python, NumPy only (Matplotlib for figures), every module independently runnable and self-checking, one driver (run_all.py) reproducing every number in this manuscript and writing results.json, and figures.py regenerating Figures Q1–Q5. Conventions: x-quadrature with vacuum variance ½ and coherent mean $\sqrt{(2N)}$; thresholds at trimmed midpoints; all losses at the Part-I budgets (−1 dB junctions, −3 dB splits, λ_p/2-convention propagation); decoder comparators at k = 16 with F = 2 − 1/G noise at the working bias. The two elementary Gaussian rules are verified against exact Fock-space evolution (loss+thermal Lindblad; two-mode-squeeze amplifier with vacuum and thermal idlers) to relative error < 10^{-4}, and the assembled symbol-error expression is verified by direct Monte-Carlo sampling; the checks run as part of the package (qlindblad.py, qnoise.py).

Module	Role	Key outputs
qconstants.py	operating point + quantum scales	$\hbar\omega_0$, T_Q, $\bar{n}(T)$, $\tau(T)$ with 150-K saturation; Part-I anchors reproduced
qmode.py	mode quantization, Eqs. (Q1–Q3)	ε_1, knee/rail quanta, χ/κ, blockade size

qnoise.py	Gaussian calculus	loss/amp/combine rules; Eq.-(7) equivalence; symbol error + MC check
qdecode.py	flash decoder	2 comparators + NOT + AND + buffer; truth table b_1b_0 = A+B
qmac.py	QMAC-1	levels, 2-bit/1-bit errors, preamp no-go, weighted demo, bits/shot
qlindblad.py	verification	Fock-space checks of both Gaussian rules, < 10^{-4} relative
qerrors.py	temperature sweeps	Figure Q3/Table Q3 curves, threshold temperatures, floors
run_all.py / figures.py	driver / figures	results.json; Figures Q1–Q5

Table QA1. The released quantum model chain, module by module.

Appendix QB. Symbols

Symbol	Meaning	Value at the operating point
$\hbar\omega_0$	one plasmon at 1 THz	4.14 meV = 0.66×10^{-3} aJ
T_Q	thermal crossover $\hbar\omega_0/k_B$	48.0 K
$\bar{n}(T)$	thermal occupation of the mode	5.8 at 300 K; 1.2 at 77 K
ε_1	single-plasmon density amplitude, Eq. (Q2)	1.62×10^{-3}
N_knee, N_rail	quanta at the 1-dB knee / the rail	38 / 3833
χ, c_K	single-plasmon Kerr rate and its O(1) coefficient	$\chi/2\pi \approx 0.7$ MHz (c_K = 0.25, band 0.05–1)
κ	mode energy decay rate $1/\tau$	10^{12} s^{-1} at 300 K; 5×10^{11} saturated
x, V	measured quadrature and its variance	vacuum V = ½; coherent mean $\sqrt{(2N)}$
N_op	quanta per input at logic 1	400 (default; launch-budget-capped)
k	comparator threshold sharpness	≥ 16 (decoder design rule; Part-I logic uses ≥ 8)
b_1b_0	decoded 2-bit word	= A+B in the digital case

Table QB1. Part-II nomenclature; Part-I symbols (s, L, M_th, Q, τ, F, G) carry over unchanged.

Data and code availability

The quantum model chain is released as the runnable Python package fable-model-quantum alongside Part I's fable-model-chain; run_all.py reproduces every quantitative claim in this manuscript (results.json) and figures.py regenerates Figures Q1–Q5. The package is archived with the manuscript at the same repository as Part I.

References

[1] R. Furui, “The Fable Computer: A Room-Temperature Terahertz Half Adder on a Regenerative Graphene-Plasmon Logic Fabric,” companion manuscript (Part I), June 2026. doi.org/10.5281/zenodo.20674840

[2] M. Dyakonov and M. Shur, “Shallow water analogy for a ballistic field effect transistor: New mechanism of plasma wave generation by dc current,” Phys. Rev. Lett. 71, 2465 (1993). doi:10.1103/PhysRevLett.71.2465

[3] S. Boubanga-Tombet et al., “Room-Temperature Amplification of Terahertz Radiation by Grating-Gate Graphene Structures,” Phys. Rev. X 10, 031004 (2020). doi:10.1103/PhysRevX.10.031004

[4] M. Gullans, D. E. Chang, F. H. L. Koppens, F. J. García de Abajo, and M. D. Lukin, “Single-Photon Nonlinear Optics with Graphene Plasmons,” Phys. Rev. Lett. 111, 247401 (2013). doi:10.1103/PhysRevLett.111.247401

[5] M. S. Tame, K. R. McEnery, Ş. K. Özdemir, J. Lee, S. A. Maier, and M. S. Kim, “Quantum plasmonics,” Nat. Phys. 9, 329–340 (2013). doi:10.1038/nphys2615

[6] E. Altewischer, M. P. van Exter, and J. P. Woerdman, “Plasmon-assisted transmission of entangled photons,” Nature 418, 304–306 (2002). doi:10.1038/nature00869

[7] C. M. Caves, “Quantum limits on noise in linear amplifiers,” Phys. Rev. D 26, 1817 (1982). doi:10.1103/PhysRevD.26.1817

[8] H. A. Haus and J. A. Mullen, “Quantum Noise in Linear Amplifiers,” Phys. Rev. 128, 2407 (1962). doi:10.1103/PhysRev.128.2407

[9] S. Lloyd and S. L. Braunstein, “Quantum Computation over Continuous Variables,” Phys. Rev. Lett. 82, 1784 (1999). doi:10.1103/PhysRevLett.82.1784

[10] C. Weedbrook, S. Pirandola, R. García-Patrón, N. J. Cerf, T. C. Ralph, J. H. Shapiro, and S. Lloyd, “Gaussian quantum information,” Rev. Mod. Phys. 84, 621 (2012). doi:10.1103/RevModPhys.84.621

[11] Y. Shen et al., “Deep learning with coherent nanophotonic circuits,” Nat. Photonics 11, 441–446 (2017). doi:10.1038/nphoton.2017.93

[12] X. Xu et al., “11 TOPS photonic convolutional accelerator for optical neural networks,” Nature 589, 44–51 (2021). doi:10.1038/s41586-020-03063-0

[13] G. X. Ni et al., “Fundamental limits to graphene plasmonics,” Nature 557, 530–533 (2018). doi:10.1038/s41586-018-0136-9

[14] K. Yoshioka et al., “On-chip transfer of ultrashort graphene plasmon wave packets using terahertz electronics,” Nat. Electron. 7, 537–544 (2024). doi:10.1038/s41928-024-01197-x

[15] P. Alonso-González et al., “Acoustic terahertz graphene plasmons revealed by photocurrent nanoscopy,” Nat. Nanotechnol. 12, 31–35 (2017). doi:10.1038/nnano.2016.185

[16] C. W. Helstrom, Quantum Detection and Estimation Theory, Academic Press (1976).