Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Physics-Based Compact Modeling and Design–Technology Co-Optimization of MoS2-Based Optoelectronic Synapses for Neuromorphic Vision

Muhammet Arucu

marucu@bandirma.edu.tr

Bandırma Onyedi Eylül University

Research Article

Keywords: advanced semiconductor devices, optoelectronic devices and photonics, MoS2 phototransistor, optoelectronic synapse, charge trapping, compact model, neuromorphic vision, in-sensor computing, design–technology co-optimization, (DTCO)

Posted Date: July 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10407366/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Physics-Based Compact Modeling and Design–Technology Co-Optimization of MoS_2-Based Optoelectronic Synapses for Neuromorphic Vision

Muhammet Arucu[a,*]

[a]Department of Computer Technologies, Bandırma Onyedi Eylül University, Gönen, 10200, Balikesir, Turkey

ARTICLE INFO

Keywords:
advanced semiconductor devices
optoelectronic devices and photonics
MoS_2 phototransistor
optoelectronic synapse
charge trapping
compact model
neuromorphic vision
in-sensor computing
design–technology co-optimization (DTCO).

ABSTRACT

Charge-trapping optoelectronic synapses based on two-dimensional semiconductors are promising building blocks for in-sensor neuromorphic vision, yet their evaluation still relies on empirical double-exponential fits and lookup-table network simulations that cannot capture the underlying trap dynamics. Here, a physics-based compact model of MoS_2 phototransistor synapses is presented, which couples a multi-level trap-kinetics ladder, solved exactly in closed form for arbitrary optical pulse trains, to a smooth field-effect-transistor transducer. Calibrated against digitized measurements from two independently published devices–a monolayer MoS_2/SiO_2 transistor and a CdSe-sensitized MoS_2 heterostructure–a single trap-ladder framework (with the number of levels selected by information criteria) reproduces paired-pulse facilitation (PPF) and multi-decade retention, and independently recovers the empirically reported relaxation constants. The model yields two analytical insights: apparent PPF time constants are emergent quantities limited by the optical pulse width rather than fundamental trap parameters, and linear charge accumulation bounds the PPF index at 200%, so that larger measured values necessarily encode transducer superlinearity, supplied in saturation by the square law and, below threshold, by the exponential regime. Exercising the calibrated model in a photo-programmed synaptic array reveals a design–technology co-optimization knob: reducing the programming irradiance reshapes the conductance ladder through the per-pulse trap filling, lowering training energy 6.7-fold at equal-or slightly-better recognition accuracy (92.6% to 93.4%, five-seed means). Finally, because every synapse carries explicit trap occupancies, the framework enables retention-aware analysis inaccessible to lookup-table approaches, quantifying inference-time weight decay, forgetting during slow training, and an optical-refresh design space sustaining a 94% accuracy floor at sub-microwatt standby power for the demonstrated 640-synapse array.

1. Introduction

Machine-vision pipelines conventionally separate image acquisition from pattern recognition, forcing every frame through an analog-to-digital conversion and a memory–processor data bus that together dominate latency and energy consumption [1, 2]. In-sensor neuromorphic computing addresses this bottleneck by merging sensing, memory, and synaptic weighting in a single optoelectronic device [3, 4, 5]. Among candidate device platforms, phototransistors based on two-dimensional (2D) semiconductors–and monolayer MoS_2 in particular–have attracted sustained attention because their large surface-to-volume ratio maximizes the influence of interface charge trapping, which converts a transient optical stimulus into a persistent conductance change [6, 7, 8, 9]. Devices exploiting this photogating mechanism have demonstrated the full canon of synaptic behaviors, including excitatory post-synaptic currents (EPSC), paired-pulse facilitation (PPF), the transition from short- to long-term memory (STM/LTM) and conductance retention exceeding 10^4 s [10, 11, 12, 13, 14].

Despite this experimental maturity, the modeling methodology of the field has remained largely empirical. The measured PPF and retention transients are routinely summarized by double exposure fits whose time constants are then interpreted as intrinsic device parameters [9, 12, 15], and system-level projections are obtained by feeding the measured potentiation/depression curves into lookup-table (LUT) neural-network simulators [12, 16, 17]. Both practices discard the underlying trap kinetics. Empirical fits cannot predict the response to stimulation protocols other than the one measured; LUT simulations freeze the synaptic weight between updates and are therefore structurally

*Corresponding author
marucu@bandirma.edu.tr (M. Arucu)
ORCID(s): 0000-0001-7620-9044 (M. Arucu)

blind to the volatility that is the defining property of charge-trapping synapses. A complementary research line deliberately exploits this volatility, using short-term dynamics of memristive and synaptic devices such as the fading memory of physical reservoir computers [18, 19]; that paradigm, however, treats the device as a fixed dynamical kernel rather than a trainable weight and likewise lacks a calibrated forward model connecting trap parameters to task-level metrics. A predictive, physics-based compact model that provides this connection–in the spirit of design–technology co-optimization (DTCO)–has been missing for trained optoelectronic synaptic arrays.

This work closes that gap with four contributions. First, a compact model is formulated that couples a trap ladder–a set of independent trap populations ordered by release time and governed by linear capture–release rate equations–to a smooth single-piece field-effect-transistor (FET) transducer valid from subthreshold to strong inversion (Section 2). Because the rate equations are linear with piecewise-constant coefficients, the response to an arbitrary optical pulse train is obtained exactly in closed form, making the model fast enough for array-level simulation. Second, the model is calibrated against measurements digitized from two independently published devices of different material composition and stimulation protocol, and it is shown that a single trap-ladder framework–with the trap count selected per device by information criteria (see the Appendix)–reproduces both PPF and multi-decade retention while independently recovering the empirically reported relaxation constants (Section 3). Third, two analytical consequences of the model are derived: the apparent PPF time constants extracted by double-exponential fitting are emergent quantities bounded by the optical pulse width, and the PPF index of any linearly accumulating charge-trapping synapse cannot exceed 200%, so that larger measured values necessarily encode transducer superlinearity (Section 4). Fourth, the calibrated model is exercised at system level: a photo-programmed synaptic array is trained in situ for image recognition, exposing programming irradiance as a DTCO knob that cuts programming energy several-fold at no cost in accuracy (Section 5), and the explicit trap occupancies carried by every synapse enable a retention-aware analysis of weight decay, slow-training forgetting, and optical-refresh scheduling that is inaccessible to LUT approaches (Section 6).

2. Physics-Based Compact Model

2.1. Multi-Level Trap Kinetics

Multi-exponential trap kinetics is a classical description of persistent photoconductivity, established decades ago for DX centers in III–V compounds [20, 21]; the contribution of this section is not the rate-equation form itself but its exact closed-form propagation under arbitrary pulse protocols and its end-to-end coupling to a transistor transducer and, in Sections 5 and 6, to array-level learning metrics. The model considers an n-type MoS_2 phototransistor in which photogenerated holes are captured by K independent trap populations located at the channel/dielectric interface, as established for MoS_2/SiO_2 MoS_2/SiO_2 [7, 9, 22], for engineered charge-trapping layers [14], and for photosensitized heterointerfaces [12], as sketched in Fig. 1. Each population $i \in \{1, \dots, K\}$ is characterized by an areal density N_i (cm^{-2}), an effective capture cross-section σ_i (cm^2), and a release time constant τ_i (s). Under an incident photon flux F ($cm^{-2}s^{-1}$), the occupancy $n_i(t)$ obeys the linear rate equation

$$\frac{\mathrm{d}n_i}{\mathrm{d}t} \;=\; \sigma_i F \left(N_i - n_i\right) \;-\; \frac{n_i}{\tau_i}. \tag{1}$$

The photon flux follows from the optical power density P as $F = \eta P / E_{\mathrm{ph}}$, with $E_{\mathrm{ph}} = hc/\lambda$ the photon energy and η an absorption/feeding efficiency absorbed into σ_i during calibration.

For any illumination protocol composed of constant-irradiance segments (light on or off), the coefficients of Equation (1) are piecewise constant and the equation admits the exact solution of the segment.

$$\begin{aligned} n_i(t) &= n_i^{\infty} + \left(n_i^{0} - n_i^{\infty}\right) e^{-k_i t}, \\ k_i &= \sigma_i F + \frac{1}{\tau_i}, \\ n_i^{\infty} &= \frac{\sigma_i F N_i}{k_i}, \end{aligned} \tag{2}$$

where n_i^0 is the occupancy at the beginning of the segment. The propagation of Equation (2) segment by segment produces the response to an arbitrary optical pulse train in closed form, without numerical integration. This is both exact and orders of magnitude faster than solving Equation (1) with a stiff ordinary-differential-equation solver, which in practice fails at the pulse-edge discontinuities; the closed-form propagation is what makes the array-level and training-time simulation tractable (Sections 5 and 6).

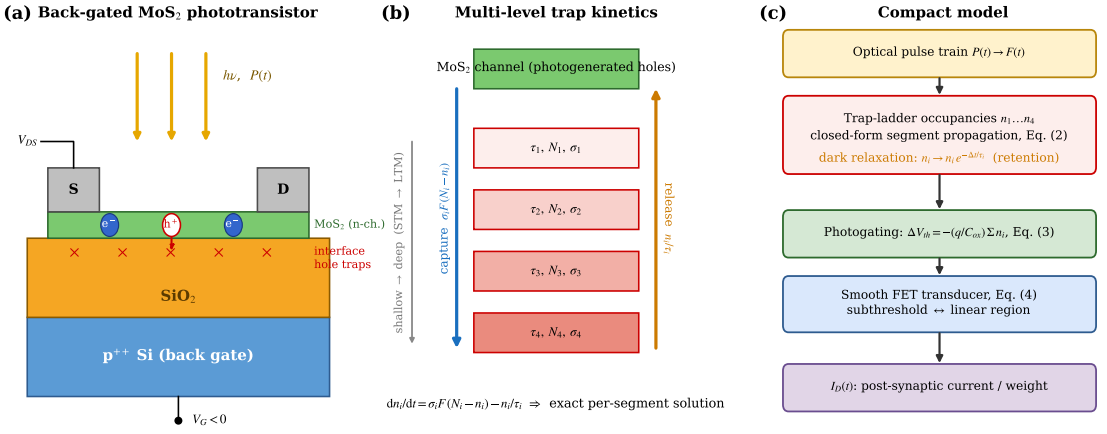


Figure 1: Device and model overview. (**a**) Back-gated MoS_2 phototransistor: photogenerated holes drift toward the channel/dielectric interface under negative gate bias and are captured by interface traps. (**b**) Multi-level trap kinetics: K independent populations with capture rate $\sigma_i F(N_i - n_i)$ and release rate n_i/τ_i; shallow levels govern short-term, deep levels long-term memory. (**c**) Compact-model signal flow from the optical pulse train to the post-synaptic current.

2.2. Transducer: From Trapped Charge to Drain Current

Trapped holes screen the gate field and shift the threshold voltage by

$$\Delta V_{\mathrm{th}}(t) \;=\; -\frac{q}{C_{\mathrm{ox}}} \sum_{i=1}^{K} n_i(t), \tag{3}$$

with C_{ox} the gate-oxide capacitance per unit area. The drain current is evaluated with a smooth single-piece expression valid from subthreshold to strong inversion,

$$\begin{aligned} I_D &= \mu C_{\mathrm{ox}} \frac{W}{L} \Big(V_{\mathrm{ov}} - \tfrac{1}{2} V_{DS}^{\mathrm{eff}} \Big) V_{DS}^{\mathrm{eff}}, \\ V_{DS}^{\mathrm{eff}} &= \frac{V_{DS}}{\big[1 + (V_{DS}/V_{\mathrm{ov}})^{\beta}\big]^{1/\beta}}, \\ V_{\mathrm{ov}} &= 2 n_{\mathrm{ss}} V_T \ln\!\Big[1 + \exp\!\Big(\frac{V_{GS} - V_{\mathrm{th0}} - \Delta V_{\mathrm{th}}}{2 n_{\mathrm{ss}} V_T} \Big)\Big], \end{aligned} \tag{4}$$

where μ is the effective mobility, W/L the aspect ratio, n_{ss} the ideality of the subthreshold, $V_T = k_B T/q$, and $\beta = 4$ sets the sharpness of the smooth triode–saturation transition through the effective drain voltage $V_{DS}^{\mathrm{eff}} \to \min(V_{DS}, V_{\mathrm{ov}})$. Equation (4) therefore continuously recovers all four relevant regimes: the exponential subthreshold law for $V_{GS} \ll V_{\mathrm{th}}$ (where V_{ov}^2 carries the correct slope), the square-law saturation current $\frac{1}{2}\mu C_{\mathrm{ox}}(W/L)V_{\mathrm{ov}}^2$ for $V_{\mathrm{ov}} < V_{DS}$ and the linear region for $V_{\mathrm{ov}} \gg V_{DS}$. Saturation is the operative regime here: both calibration experiments were performed at $V_{DS} = 1.0$ V with near-threshold operating points ($V_{\mathrm{ov}} \ll V_{DS}$), and the Appendix shows that the saturation-aware form fits the data decisively better than a V_{DS}-linear (triode) transducer. The continuous transition between regimes is essential for the facilitation analysis of Section 4. The post-synaptic current is $\Delta I_D = I_D - I_{\mathrm{dark}}$.

Figure 2 demonstrates that the model with two trap populations already reproduces the canonical synaptic phenomenology: single-pulse EPSC with STM decay, PPF that decreases with pulse interval, pulse-number-dependent potentiation, and the STM→LTM transition in which repeated stimulation extends retention–the fingerprint of shallow trap populations (fast-release) and deep (slow-release) acting in parallel [9, 12].

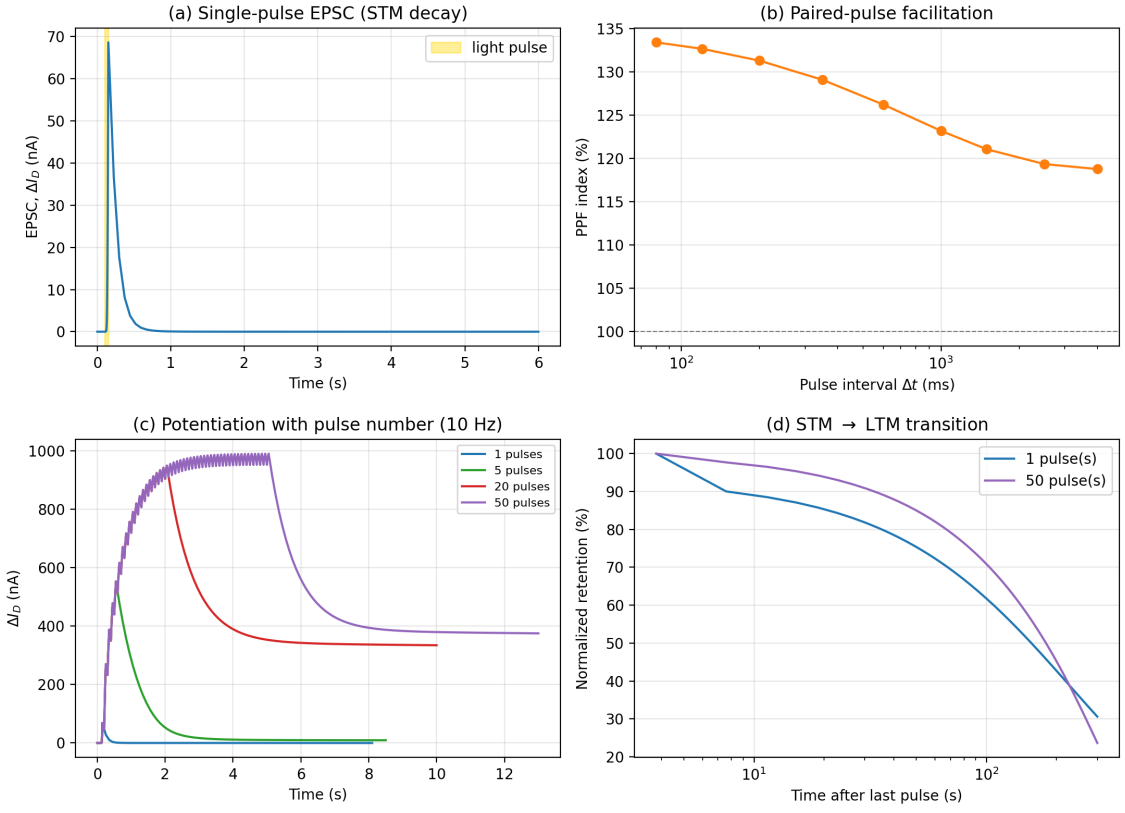


Figure 2: Baseline synaptic phenomenology of the compact model (two trap populations; illustrative, pre-calibration parameters–the calibrated devices are treated in Section 3). (**a**) Single-pulse excitatory post-synaptic current and short-term-memory decay. (**b**) Paired-pulse facilitation index versus pulse interval. (**c**) Potentiation under 10 Hz pulse trains of increasing length. (**d**) Normalized retention after 1 versus 50 pulses, showing the short-term to long-term memory transition.

3. Calibration Against Two Published Devices

3.1. Data Extraction and Fitting Procedure

To test whether one framework generalizes across material systems, the model was calibrated against two independently published charge-trapping optoelectronic synapses: Device A a monolayer MoS_2/SiO_2 back-gated FET stimulated with 5 s, 10 mW/cm^2, 450 nm pulses [9], and Device B a CdSe-nanobelt-photosensitized MoS_2 transistor stimulated with 1 s, 10.84 mW/cm^2 pulses [12]. The PPF and retention curves of both devices (Figures 5b and 2b of Ref. [9]; Figures 4c and 4a of Ref. [12], both published under CC BY licenses) were digitized programmatically: plot frames and tick marks were located from pixel-intensity profiles, curves were extracted through color masking, and scatter markers through connected-component analysis. Each extraction was validated against the anchor values quoted in the source texts; for example, the maximum extracted PPF index of Device A is 203.9% in $\Delta t = 5$ s versus the reported 203.5% [9]. One provenance detail is disclosed explicitly: at $\Delta t = 40$ and 50 s the measured markers of Ref. [9] are fully occluded by the authors' fitted curve, so these two points were recovered from that curve rather than from marker centroids; they are drawn as open symbols in Fig. 3a. The digitization scripts are available from the author (Data Availability Statement) so that the entire calibration is reproducible from the published figures.

For each device, the set of parameters $\{\sigma_i, N_i, \tau_i\}_{i=1}^{K}$ plus the operating-point offset $V_{GS} - V_{\mathrm{th0}}$ ($3K+1$ parameters) was fitted in the logarithmic space using the least squares of the multi-start trust-region, minimizing the joint residual of the relative PPF error and the normalized retention shape. The trap count K was not imposed, but was selected per device by refitting for $K = 2$–5 and comparing Akaike and Bayesian information criteria; both criteria select $K = 4$ for Device A and $K = 2$ for Device B, and parameter confidence intervals and correlations are reported in

Table 1
Fitted model parameters for the two calibrated devices. The trap count K is selected per device by information criteria (see the Appendix). Because the absorption/feeding efficiency η is absorbed into σ_i and the geometry and mobility are nominal, σ_i and N_i are effective fit constants rather than microscopic cross-sections and densities.

Parameter	Device A (MoS_2/SiO_2 [9])	Device B (CdSe/MoS_2 [12])
Pulse protocol	5 s, 10 mW/cm^2, 450 nm	1 s, 10.84 mW/cm^2
Selected trap count K	4	2
$\tau_1, \ldots, \tau_K$ (s)	7.3, 49, 548, 8460	0.8, 1030
$\sigma_1, \ldots, \sigma_K$ (10^{-18} cm^2)	0.20, 9.7, 43.1, 0.39	9.6, 30.5
$N_1, \ldots, N_K$ (10^{11} cm^{-2})	46.2, 1.2, 1.7, 6.4	7.8, 1.3
$V_{GS} - V_{\mathrm{th0}}$ (V)	−0.27	−0.21
Empirical constants of source paper (s)	24.8 (PPF); 348, 3070 (ret.)	— ; > 1500 (ret.)

the Appendix. PPF was simulated exactly as measured – two pulses of the protocol-specific width, amplitudes A_1, A_2 taken from the dark baseline –and retention as the normalized decay after the protocol-specific pulse train: five pulses for Device A, and–since the programming history of the long-retention trace of Ref. [12] is not fully specified – single-pulse programming for Device B, an assumption whose influence is quantified in the Appendix.

3.2. Results: One Framework, Two Devices

Figure 3 compares the fitted model with the digitized data, and Table 1 lists the extracted parameters. For Device A, the PPF curve is reproduced within 1.0% (mean 0.4%) over the full measured range $\Delta t = 5$–200 s–including the maximum, 203.4% predicted versus 203.5% reported [9]–and the retention shape over four decades of time with a mean absolute error of 0.003. The fitted release times $\tau_3 = 548$ s and $\tau_4 \approx 8.5 \times 10^3$ s agree, within their confidence intervals (see the Appendix), with the empirical constants reported in Ref. [9] from separate double-exponential fits of retention ($\tau = 348$ s and 3070 s), and $\tau_2 = 49$ s is of the same order as the slow PPF constant ($\tau = 24.78$ s)–even though none of these values was supplied to the optimizer. A single self-consistent parameter set thus explains two independent measurement sets that the empirical approach must fit separately.

Equally instructive is what the empirical description misses: the measured PPF of Device A saturates on a floor of ≈134–138% for $\Delta t \gtrsim 100$ s rather than decaying to 100%. This floor emerges naturally in the trap model– it is the residue of the slow (retention) populations that have not released between the paired pulses, whereas a double-exponential decay toward unity, as commonly assumed, cannot represent it. Device B, with a different photosensitization mechanism and a five-times shorter pulse, is described within the same framework by only two statistically supported levels (see the Appendix): a slow level with $\tau_2 \approx 10^3$ s that accounts for the retention of > 1500 s highlighted by the authors [12] (mean absolute retention error 0.05), and a fast level that reproduces the sharp sub-second facilitation decay (all PPF points within 2%). However, the nominal value of the fast constant, $\tau_1 = 0.8$ s, must be read with care: its 95% confidence factor exceeds ×20, because a 1-s optical pulse cannot resolve sub-second kinetics, the statistical expression of the resolvability limit derived in Section 4. The data determine that a fast process exists and how much amplitude it carries, but not its timescale or microscopic origin (interfacial trapping versus free-carrier recombination).

3.3. Independent Prediction Test: Long-Train Potentiation

Reproducing the curves used for calibration does not yet demonstrate predictive power. As an out-of-sample test, the Device-A parameter set of Table 1–fitted exclusively to PPF and retention–was used to predict the response to a third independent protocol from [9]: optical potentiation by 50 light pulses (5 s on / 5 s off) at $V_D = 1.0$ V and $V_G = -2$ V. Figure 4 places the blind predictions next to the measured panel; both self-consistent transducer–parameter pairs of the Appendix are shown. Two lessons emerge. First, the temporal dynamics are predicted robustly: both forms yield a monotonic conductance staircase evolving on the correct several-hundred-second scale set by the fitted trap ladder. Second, the curvature of the staircase is not identifiable from PPF and retention alone: the triode pair predicts a concave staircase reaching half range by pulse 18, the saturation pair a nearly linear one (half range at pulse 26), and the measured curve–with its reported potentiation nonlinearity of 2.92 [9]–falls within the band spanned by the two, closer to the concave edge. The long-train measurement thus supplies exactly the constraint that the paired-pulse and retention protocols leave open, and closing it requires either such a potentiation trace or short-pulse transients in the

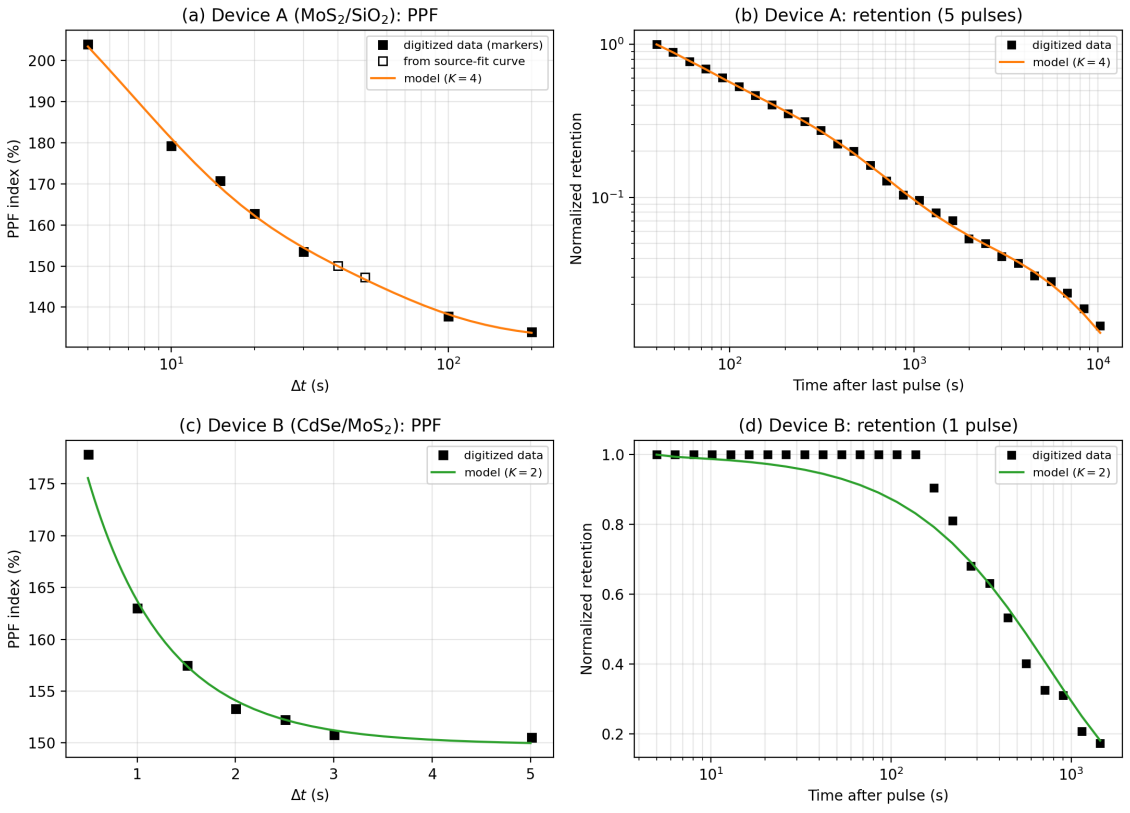


Figure 3: One compact model calibrated to two published devices (symbols: programmatically digitized data; lines: model). (**a**) Device A (monolayer MoS_2/SiO_2 [9], $K = 4$), paired-pulse facilitation with 5 s pulses; filled squares are digitized measurement markers, open squares the two points recovered from the source-fit curve (Section 3); note the ≈135% floor reproduced by the model. (**b**) Device A, normalized retention after five pulses over four decades of time. (**c**) Device B (CdSe-sensitized MoS_2 [12], $K = 2$), paired-pulse facilitation with 1 s pulses; the sharp sub-second component is reproduced by a fast level whose time constant is unconstrained (resolvability limit, Section 4). (**d**) Device B, normalized retention.

calibration set. Importantly, the array-level conclusions of Sections 5 and 6 are based on calibrated trap dynamics and on relative comparisons controlled by irradiance, which are common to both forms.

4. Analytical Insights

4.1. Apparent PPF Time Constants Are Emergent, Pulse-Width-Limited Quantities

Consider a single trap population probed by two pulses of width t_p separated by an interval Δt. During illumination, the occupancy approaches its steady state at rate $k_{\text{on}} = \sigma F + 1/\tau$; between pulses, it relaxes at rate $1/\tau$. Two regimes follow directly from Equation (2).

If $k_{\text{on}} t_p \gg 1$, the population equilibrates within each pulse: both pulses terminate at the same occupancy n^∞ regardless of history, so $A_2 \approx A_1$ and the population does not contribute to facilitation. In particular, any population with $\tau < t_p$ necessarily satisfies $k_{\text{on}} t_p > t_p/\tau > 1$ and is therefore invisible to a PPF measurement with pulse width t_p. In contrast, if $k_{\text{on}} t_p \ll 1$, the population accumulates linearly, each pulse adds $\delta n \approx \sigma F N t_p$, and the paired-pulse response from the dark baseline is

$$\frac{A_2}{A_1} \approx 1 + e^{-\Delta t/\tau} \tag{5}$$

(linear transduction, far from saturation).

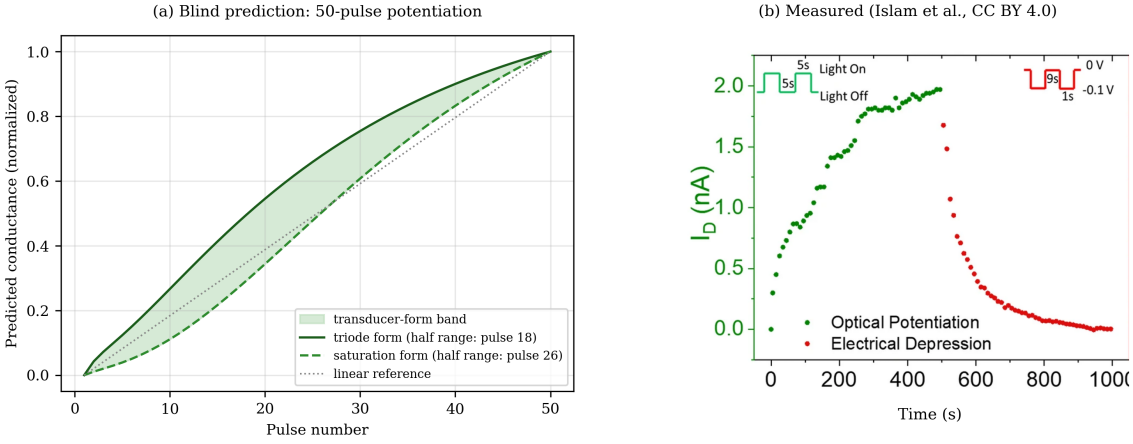


Figure 4: Out-of-sample prediction test. (**a**) Blind predictions of the 50-pulse optical potentiation protocol of Ref. [9] (5 s on / 5 s off), computed from parameters fitted only to PPF and retention: the two self-consistent transducer–parameter pairs (saturation-aware and triode; the Appendix) span the shaded band, and the measured curvature falls within it. (**b**) The corresponding measured panel, reproduced from Islam et al. [9] under the terms of the Creative Commons Attribution 4.0 license (optical potentiation, green, left axis; electrical depression, red, right axis).

Two consequences follow. First, the "fast" time constants obtained by double-exponential fitting of PPF data are emergent quantities: they reflect only those populations with $\tau \gtrsim t_p$, and any dynamics faster than the pulse width is aliased into the fitted amplitudes. For Device A ($t_p = 5$ s), the reported $\tau = 2.45$ s PPF component [9] therefore cannot correspond to a linearly responding trap level; for Device B ($t_p = 1$ s), the steep facilitation decay between $\Delta t = 0.5$ and 1 s (Fig. 3c) likewise lies below the resolvability limit: although a fast trap level reproduces it (Section 3), its time constant is unconstrained, and the EPSC overshoot visible in Ref. [12] suggests that free-carrier recombination contributes to the same feature. Second, protocol design matters: extracting genuine trap constants requires $t_p \ll \tau$, i.e., pulse widths much shorter than the fastest dynamics of interest.

4.2. The 200% Facilitation Bound and Transducer Superlinearity

Equation (5) is maximized at $\Delta t \to 0$, giving $A_2/A_1 \to 2$. Including capture saturation only reduces this bound, as the second-pulse fill $\delta n_2 = \sigma F(N - n)t_p$ decreases with residual occupancy. Hence, for any charge-trapping synapse read out by a transducer that is linear in trapped charge, the PPF index cannot exceed 200%, provided both amplitudes are referenced to the dark baseline, as in Refs. [9, 12] and throughout this work; if A_2 is instead measured from the pre-second-pulse level, the bound does not apply, and the two conventions must not be compared directly. Superlinear transduction relaxes the bound in a quantifiable way. In saturation, $I_D \propto V_{\mathrm{ov}}^2$, so the referenced baseline amplitude for a dark overdrive v and a charge-induced shift δ is $A \propto (v + \delta)^2 - v^2$; with linear accumulation ($\delta_2 \leq 2\delta_1$) the ratio $A_2/A_1 = 4(v + \delta_1)/(2v + \delta_1)$ interpolates between 200% (small signal, $\delta_1 \ll v$) and an upper limit of 400% (large signal, $\delta_1 \gg v$), while below the threshold the exponential dependence removes any finite bound. The maximum PPF of 203.5% measured for Device A [9] exceeds the linear-transduction bound and therefore necessarily encodes such transducer superlinearity–here supplied by the square-law and near-threshold regimes of Equation (4); consistently, the unconstrained fits place both devices slightly below threshold ($V_{GS} - V_{\mathrm{th0}} \approx -0.2$ V, Table 1). The measured PPF index of a phototransistor synapse is thus a product of trap kinetics and transducer nonlinearity, and reporting it without the operating point conflates the two. This observation also identifies gate bias as a legitimate design variable for facilitation-based temporal filtering.

5. Design–Technology Co-Optimization of a Photo-Programmed Array

5.1. Array Model and In Situ Training

To connect device physics to recognition metrics, the calibrated Device A model was deployed as the weight element of a single-layer optoelectronic classifier. Each synaptic weight is realized by a differential device pair, $W \propto G^+ - G^-$, programmed by optical pulses that advance the devices along their conductance ladder $G(p)$, calculated

from the physical model rather than assumed. Training follows the Manhattan rule–one programming pulse per selected weight in the direction of the gradient sign–in the standard potentiation-only differential scheme for crossbar hardware [16, 23, 24, 25]: a weight is increased by potentiating G^+ and decreased by potentiating G^-, so that the slow electrical erasure [9] is never invoked during learning and enters only through the periodic state-consistent refresh of saturated pairs, which is treated as an ideal reprogramming step. A direct consequence is that the array-level results are independent of the details of the electrical-depression characteristic; this was verified explicitly by sweeping the behavioral depression parameter by ±50%, which leaves every precision in this section unchanged. The task is ten-class handwritten-digit recognition on 8×8 images (1797 samples, 70/30 train/test split); an ideal floating-point softmax classifier trained by stochastic gradient descent provides the reference of 97.4% test accuracy.

5.2. Programming Irradiance as a DTCO Knob

Figure 5 summarizes the array-level results. Because per-pulse trap filling scales with irradiance (Equation (2)), the programming power directly shapes the 50-state conductance ladder–a physical handle invisible to LUT descriptions [16, 24] and complementary to recent material-level linearity engineering of optoelectronic synapses [26]. At nominal 10 mW / cm^2 the ladder is S-shaped, with update steps compressed to 0.9–1.5% of the range at both ends of the state space; at 1.5 mW / cm^2 the steps are redistributed toward the useful early range (1.2–4.2%). In situ training reaches $92.6 \pm 1.4\%$ and $93.4 \pm 1.0\%$ (mean ± s.d., five seeds), respectively, against the 97.4% floating-point reference: the low-irradiance ladder trains at least as accurately while–as shown below–costing 6.7× less energy per update, so the knob delivers its principal benefit on the energy axis without accuracy penalty. The variation from device-to-device of the update step is well tolerated: the accuracy is statistically flat up to $\sigma = 10\%$ ($92.6 \pm 1.1\%$, $92.8 \pm 1.2\%$ and $92.3 \pm 1.6\%$ in $\sigma = 0$, 5%, and 10%; the slight non-monotonicity in $\sigma = 5\%$ lies within the seed-to-seed uncertainty, eight seeds) and then gracefully degrades to $86.2 \pm 1.3\%$ in $\sigma = 30\%$, providing a tolerance budget for array design.

The energy analysis of Fig. 6 completes the co-optimization loop. The optical programming energy per pulse is $E = P\,A_{\mathrm{dev}}\,t_p$; Since the filling of the trap far from saturation depends on the fluence $P\,t_p$, the energy is set by the fluence × area and is independent of how the fluence is delivered. The linear capture requires $\sigma_i F t_p < 1$, which for the strongest fitted cross-section is a fluence condition, $\Phi < \Phi^* = E_{\mathrm{ph}}/\sigma_{\max} \approx 10$ mJ / cm^2 –satisfied by the low-power scheme (7.5 mJ / cm^2) but exceeded at the nominal one (50 mJ / cm^2), whose strongest level saturates within each pulse; and pulse widths approaching the fastest release time ($t_p \gtrsim \tau_1 = 7.3$ s) lose charge during Both limits are marked in Fig. 6a. For the 50 μm^2 device, one programming pulse costs 25 nJ at 10 mW/cm^2 versus 3.75 nJ at 1.5 mW/cm^2; over a full training run (307,200 pulses), the low-irradiance scheme reduces programming energy from 7.7 mJ to 1.15 mJ at equal-or-better accuracy–the accuracy–energy trade-off dissolves rather than binds. Electrical read energy is negligible at the fitted subthreshold operating point ($\sim 10^{-18}$ J per read at $V_{\mathrm{read}} = 0.1$ V, $t_{\mathrm{read}} = 1$ μs; even read currents of microamp-class would keep it below 0.1 pJ), orders of magnitude below the optical budget. Scaling the fluence–area product indicates a path from the current nJ-class updates toward the fJ class of mature memristive synapses [25, 27] through area reduction and capture-efficiency engineering; reducing the per-update energy of synaptic transistors is an active optimization target across device families [28].

6. Retention-Aware Training and Inference

The defining property of charge-trapping synapses, volatility, is precisely what LUT simulations cannot represent. In the present framework, every synapse carries the explicit occupancies n_i of all trap populations; a programming pulse advances them according to Equation (2), and any elapsed interval Δt relaxes them in closed form as $n_i \to n_i\, e^{-\Delta t/\tau_i}$. Training and inference thus become genuinely time-aware. Programming uses the iso-fluence equivalent of the low-irradiance scheme (50 ms at 150 mW/cm^2; identical 3.75 nJ per pulse), and the array is trained as in Section 5, reaching 95.2%–above the corresponding LUT-quantized result ($93.4 \pm 1.0\%$), since the continuous physical state removes quantization loss.

Figure 7 presents three analyzes enabled by this formulation. First, inference-time weight decay: after training, accuracy stays above 94% for more than half an hour, crosses 90% only after ≈2.2 h, and collapses within ≈8 h, tracking the slow-trap constants of Table 1; an engineered "deep-trap-only" variant (shallow populations passivated) still holds 95% at the moment the as-fitted mix has fallen to 90%, providing a quantitative target for trap engineering. The wall-clock latency per training batch grows from 0.01 s to 5 s, the final accuracy drops monotonically from $94.5 \pm 0.5\%$ to $87.0 \pm 1.5\%$, and to $83.7 \pm 0.3\%$ at 20 s (mean ± s.d., three seeds), because weights written early in an epoch relax before the epoch completes–a co-design constraint linking programming latency (Section 5) directly to learning

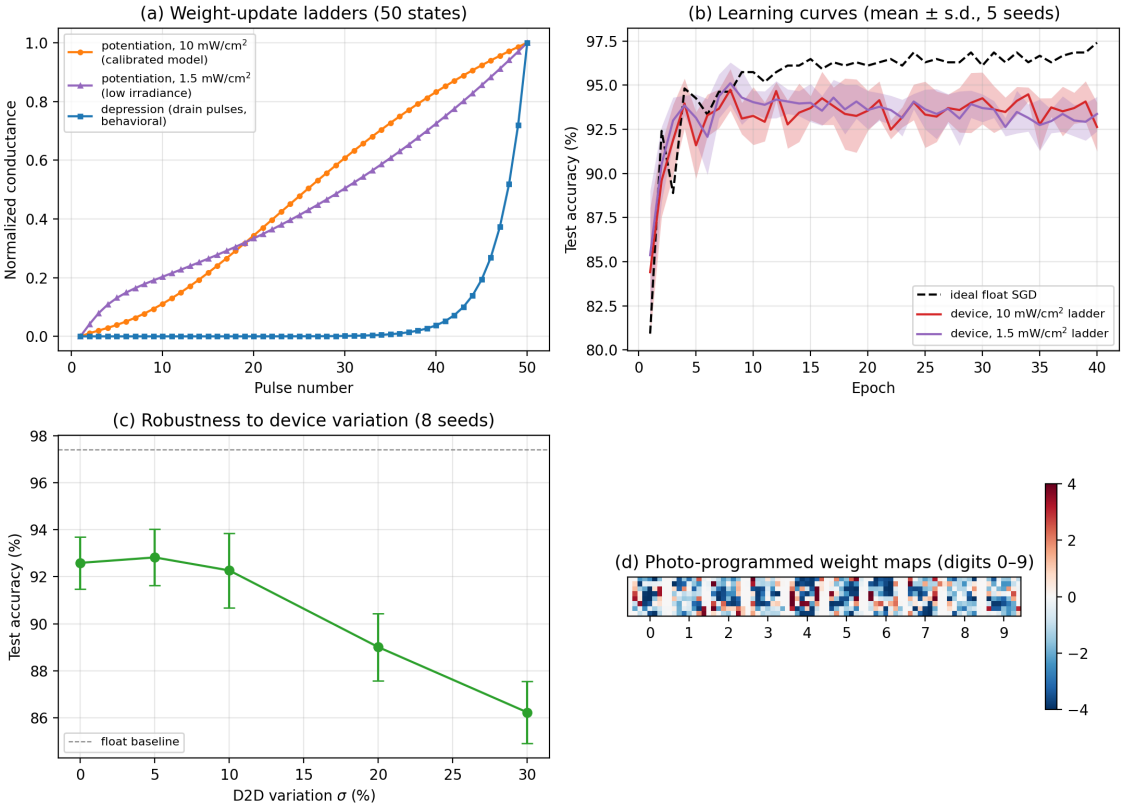


Figure 5: Photo-programmed synaptic array built from the calibrated Device A model. (**a**) Fifty-state weight-update ladders: potentiation at 10 and 1.5 mW/cm^2 programming irradiance (physical model) and the behavioral electrical-depression characteristic, shown for completeness (training updates are potentiation-only; see text). (**b**) In situ Manhattan-rule learning curves against the floating-point reference (mean ± s.d. over five seeds). (**c**) Final accuracy versus device-to-device variation of the update step (mean ± s.d., eight seeds). (**d**) Photo-programmed weight maps of the ten output classes.

Third, sustained inference with periodic optical refresh, simulated over a 6×10^4 s horizon that spans the full decay of panel (a): the accuracy floor traces a genuine design trade-off against standby power for the 640-synapse array (the power scales linearly with array size). Refresh every 300 s holds 94.4% at 0.94 μW; stretching the period to 3×10^3 s costs only 0.7 points (93.7% at 94 nW); at $T_{\mathrm{ref}} = 10^4$ s–of the order of τ_4–the floor drops to 88.0% at 28 nW; and beyond τ_4 ($T_{\mathrm{ref}} = 3 \times 10^4$ s) the weights decay fully between refreshes and the floor collapses to chance. The optical analog of DRAM refresh scheduling is thus benign down to tens of nanowatts, with the binding constraint set directly by the deepest fitted trap level.

7. Discussion

The results support a methodological claim: replacing empirical fits and lookup tables with a calibrated, analytically solvable trap-kinetics model changes the questions that can be asked of an optoelectronic synapse. The retention-aware formulation of Section 6 also builds a bridge toward reservoir computing, where the same volatility is a computational resource rather than a liability [18, 19]: the calibrated trap ladder supplies exactly the fading-memory kernel that reservoir approaches presently obtain empirically. Protocol dependence (Section 4), irradiance-driven update shaping (Section 5), and time-aware operation (Section 6) are all predictions of one set of parameters per device, obtained from two published curves each. A subtlety of Section 3 and the Appendix deserves one explicit remark: in-sample model selection and out-of-sample curvature constrain different aspects of the transducer–the information criteria compare how the two forms reproduce paired-pulse and retention shapes, whereas the long-train curvature probes the large-signal transfer characteristic that those protocols never exercise–so preferring the saturation form on the calibration data

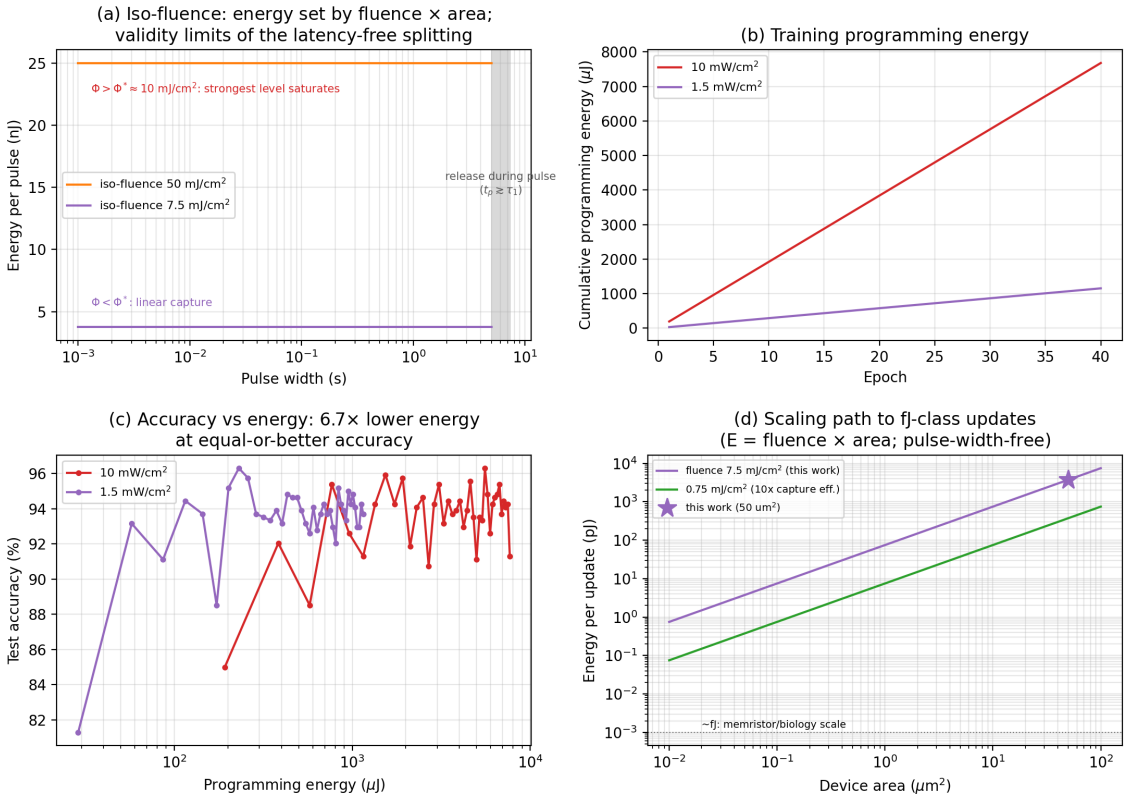


Figure 6: Energy analysis of the photo-programmed array. (**a**) At constant fluence, programming energy is independent of pulse width; the shaded region ($t_p \gtrsim \tau_1$) and the fluence criterion $\Phi \lesssim \Phi^*$ mark where the latency-free splitting breaks down (Section 5). (**b**) Cumulative programming energy during training for the two irradiance schemes. (**c**) Accuracy versus cumulative programming energy: the low-irradiance ladder improves both axes simultaneously. (**d**) Energy-per-update scaling with device area at constant fluence, indicating the path toward fJ-class updates.

while observing triode-like curvature out of sample is not a contradiction but a statement about which measurements constrain what.

Several limitations bound the present study and chart future work. First, electrical depression is modeled behaviorally; a physical de-trapping model driven by the drain field would close the last empirical element of the framework. Second, the fast facilitation (<1 s) of Device B, the reported sub-second PPF constants of Device A, and the residual curvature spread of the blind potentiation prediction (Fig. 4) all point to the same missing ingredient: a free-carrier recombination transient below the pulse-width resolvability limit of the available data. Extending the model with a single fast photoconductive channel, calibrated against short-pulse measurements, is straightforward within the same closed-form machinery. Third, the recognition benchmark uses 8×8 digit images to keep the variation and retention sweeps tractable; the training pipeline is dataset-agnostic, and full-MNIST evaluation is planned for the extended framework. Fourth, trap release is treated as temperature-independent; an Arrhenius extension would add testable thermal predictions and connect the fitted τ_i to trap energy depths. Finally, the closed-form segment propagation of Equation (2) maps directly onto a Verilog-A implementation, which would make the model usable in standard circuit-design flows and strengthen the DTCO link to peripheral CMOS.

8. Conclusions

A physics-based compact model of charge-trapping MoS_2 optoelectronic synapses was developed that couples exactly solvable multi-level trap kinetics to a smooth FET transducer. Calibrated to programmatically digitized data from two independently published devices, a single trap-ladder framework–with the trap count selected per device

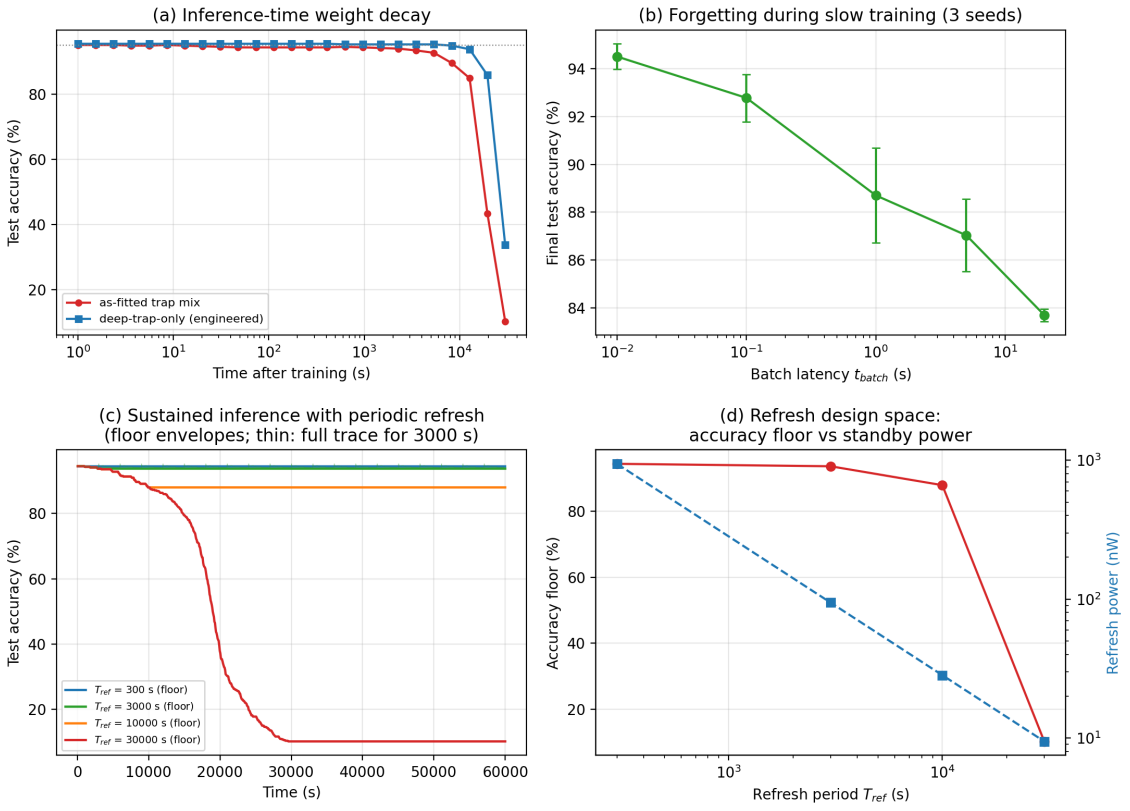


Figure 7: Retention-aware operation of the physically dynamical array. (**a**) Test accuracy versus time after training for the as-fitted trap mix and an engineered deep-trap-only device. (**b**) Final accuracy versus training batch latency (mean ± s.d., three seeds): slow programming causes forgetting during training. (**c**) Sustained inference under periodic optical refresh over a 6×10^4 s horizon (running-minimum floor envelopes; thin line: full trace for $T_{\mathrm{ref}} = 3000$ s). (**d**) Refresh design space: accuracy floor and standby refresh power versus refresh period, spanning the benign regime to the collapse beyond τ_4.

by information criteria–reproduced paired-pulse facilitation and multi-decade retention and independently recovered the empirically reported relaxation constants, while additionally capturing the facilitation floor that empirical double-exponential fits cannot represent. The model yielded two analytical results of general applicability–the pulse-width resolvability limit on apparent PPF time constants and the 200% facilitation bound of linearly transduced charge accumulation–and, exercised at system level, exposed programming irradiance as a design–technology co-optimization knob that reduces training energy 6.7-fold at equal-or-slightly-better recognition accuracy. Because every simulated synapse carries explicit trap occupancies, the framework further enabled retention-aware analysis of inference-time decay, slow-training forgetting, and optical-refresh scheduling, sustaining a 94% accuracy floor at sub-microwatt standby power for the 640-synapse demonstrator. Together, these results establish a reproducible modeling path from trap physics to system metrics for optoelectronic neuromorphic hardware.

Funding

The author declares that no funds, grants, or other support was received during the preparation of this manuscript.

Data Availability

The Python simulation framework, the programmatic figure-digitization scripts, the digitized data sets, and the scripts reproducing all figures of this article are available from the corresponding author upon reasonable request.

Table 2
Model-order selection: residual sum of squares and information criteria versus trap count K. Bold marks the selected order.

	Device A			Device B		
K	RSS	AIC	BIC	RSS	AIC	BIC
2	0.155	−188.5	−177.3	**0.114**	**−166.3**	**−156.0**
3	0.019	−260.7	−244.6	0.115	−160.2	−145.5
4	**0.003**	**−324.6**	**−303.7**	0.115	−154.1	−135.0
5	0.003	−318.8	−293.0	0.105	−151.1	−127.6

Acknowledgment

During the preparation of this manuscript, the author used the Claude (Anthropic) large language model to assist with code development, data analysis, and text drafting. The author has reviewed and edited all the results and assumes full responsibility for the content of this publication.

Appendixes

Model-Order Selection and Parameter Identifiability: The device, the fit of Section 3 was repeated for trap counts $K = 2$–5 with multi-start trust-region optimization (cold log-spaced starts plus warm starts derived from neighboring solutions K), and the resulting residual sums of squares were compared through the Akaike and Bayesian information criteria, $\mathrm{AIC} = n\ln(\mathrm{RSS}/n)+2k$ and $\mathrm{BIC} = n\ln(\mathrm{RSS}/n)+k\ln n$, with n the number of weighted residuals and $k = 3K + 1$ the parameter count (Table 2). Both criteria select $K = 4$ for Device A–adding a fifth level leaves the residual unchanged–and $K = 2$ for Device B, whose seven PPF points and plateau-dominated retention trace do not support additional levels.

Parameter uncertainties were estimated from the Gauss–Newton covariance $\mathbf{C} = s^2(\mathbf{J}^\top\mathbf{J})^+$ at the optimum, with $\mathbf{J}$ the residual Jacobian and $s^2 = \mathrm{RSS}/(n-k)$; since the parameters are fitted in $\log_{10}$ space, 95% confidence intervals are quoted as multiplicative factors. For Device A ($K = 4$), the release times are constrained to $\tau_1 = 7.3$ s ($\times/2.1$), $\tau_2 = 49$ s ($\times/4.7$), $\tau_3 = 548$ s ($\times/3.6$), and $\tau_4 = 8.5\times10^3$ s ($\times/4.1$): the timescale ordering and decade placement are robust, while individual constants are determined to within roughly half a decade. The correlation matrix shows the expected degeneracies: capture strengths correlate with the neighboring release times, the pair (σ_4, N_4) is almost perfectly anticorrelated ($r = -0.99$, only the product being constrained), and the operating-point offset correlates with the overall amplitude parameters. For Device B, the slow constant is determined to $\tau_2 \approx 10^3$ s ($\times/5.3$) while the fast constant carries a confidence factor exceeding ×20–quantitative confirmation that sub-second dynamics cannot be identified from 1-s-pulse data, in line with Section 4. The programming-history assumption for the Device-B retention trace was also tested: refitting under 5- and 10-pulse programming instead of a single pulse changes the slow constant from 1.03×10^3 s to at most 1.37×10^3 s–a factor of 1.3, well inside its confidence interval–while the fast constant and the operating point are unaffected. Finally, the entire analysis was repeated with the V_{DS}-linear (triode) transducer: the information-criterion selection ($K = 4$ for Device A, $K = 2$ for Device B) and all timescale orderings are unchanged, while the saturation-aware form of Equation (4) improves the Akaike criterion by 23 (Device A) and 12 (Device B) units and is therefore adopted throughout; the individual constants shift between the two forms within the confidence intervals quoted above.

CRediT authorship contribution statement

Muhammet Arucu: Conceptualization of this study, Methodology, Software, Data curation, Writing - Original draft preparation.

References

[1] Carver Mead. Neuromorphic electronic systems. *Proc. IEEE*, 78(10):1629–1636, 1990.
[2] Yang Chai. In-sensor computing for machine vision. *Nature*, 579(7797):32–33, 2020.
[3] Feichi Zhou, Zheng Zhou, Jiewei Chen, Tsz Hin Choy, Jinli Wang, Nian Zhang, Ziyuan Lin, Shimeng Yu, Jinfeng Kang, H.-S. Philip Wong, and Yang Chai. Optoelectronic resistive random access memory for neuromorphic vision sensors. *Nat. Nanotechnol.*, 14(8):776–782, 2019.

[4] Lukas Mennel, Joanna Symonowicz, Stefan Wachter, Dmitry K. Polyushkin, Aday J. Molina-Mendoza, and Thomas Mueller. Ultrafast machine vision with 2D material neural network image sensors. *Nature*, 579(7797):62–66, 2020.

[5] Changjin Wan et al. Artificial sensory memory. *Adv. Mater.*, 32(15):1902434, 2020.

[6] Oriol Lopez-Sanchez, Dominik Lembke, Metin Kayci, Aleksandra Radenovic, and Andras Kis. Ultrasensitive photodetectors based on monolayer MoS_2. *Nat. Nanotechnol.*, 8(7):497–501, 2013.

[7] Antonio Di Bartolomeo et al. Electrical transport and persistent photoconductivity in monolayer MoS_2 phototransistors. *Nanotechnology*, 28(21):214002, 2017.

[8] Jaewook Lee et al. Monolayer optical memory cells based on artificial trap-mediated charge storage and release. *Nat. Commun.*, 8:14734, 2017.

[9] Molla Manjurul Islam, Durjoy Dev, Adithi Krishnaprasad, Laurene Tetard, and Tania Roy. Optoelectronic synapse using monolayer MoS_2 field effect transistors. *Sci. Rep.*, 10(1):21870, 2020.

[10] Hong Kai He et al. Photonic potentiation and electric habituation in ultrathin memristive synapses based on monolayer MoS_2. *Small*, 14(15):1800079, 2018.

[11] Seung-Gyun Kim et al. Infrared detectable MoS_2 phototransistor and its application to artificial multilevel optic-neural synapse. *ACS Nano*, 13(9):10294–10300, 2019.

[12] Xiaohui Song, Xiaojing Lv, Mengjie He, Fei Mao, Jie Bai, Xuan Qin, Yanjie Hu, Zinan Ma, Zhen Liu, Xueping Li, Chenhai Shen, Yurong Jiang, Xu Zhao, and Congxin Xia. Artificial optoelectronic synapse based on CdSe nanobelt photosensitized MoS_2 transistor with long retention time for neuromorphic application. *Nanophotonics*, 13(22):4211–4224, 2024.

[13] Meng Zhang et al. MoS_2-based charge-trapping synaptic device with electrical and optical modulated conductance. *Nanophotonics*, 9(8):2475–2486, 2020.

[14] DaYea Oh, Haena Yim, So Yeon Yoo, Gwangtaek Oh, Chansoo Yoon, Ji-Won Choi, and Bae Ho Park. Synaptic MoS_2 transistors based on charge trapping two-dimensionally confined in $Sr_2CoNb_3O_{10}$ nanosheets. *Mater. Sci. Semicond. Process.*, 160:107424, 2023.

[15] Robert S. Zucker and Wade G. Regehr. Short-term synaptic plasticity. *Annu. Rev. Physiol.*, 64(1):355–405, 2002.

[16] Pai-Yu Chen, Xiaochen Peng, and Shimeng Yu. NeuroSim: A circuit-level macro model for benchmarking neuro-inspired architectures in online learning. *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, 37(12):3067–3080, 2018.

[17] Seunghwan Seo et al. Artificial optic-neural synapse for colored and color-mixed pattern recognition. *Nat. Commun.*, 9:5106, 2018.

[18] Chao Du, Fuxi Cai, Mohammed A. Zidan, Wen Ma, Seung Hwan Lee, and Wei D. Lu. Reservoir computing using dynamic memristors for temporal information processing. *Nat. Commun.*, 8:2204, 2017.

[19] Gouhei Tanaka, Toshiyuki Yamane, Jean Benoit Héroux, Ryosho Nakane, Naoki Kanazawa, Seiji Takeda, Hidetoshi Numata, Daiju Nakano, and Akira Hirose. Recent advances in physical reservoir computing: A review. *Neural Netw.*, 115:100–123, 2019.

[20] David V. Lang and Ralph A. Logan. Large-lattice-relaxation model for persistent photoconductivity in compound semiconductors. *Phys. Rev. Lett.*, 39(10):635–639, 1977.

[21] Patricia M. Mooney. Deep donor levels (DX centers) in III-V semiconductors. *J. Appl. Phys.*, 67(3):R1–R26, 1990.

[22] Yuzheng Guo, Xianlong Wei, Jiepeng Shu, Bo Liu, Jinbo Yin, Chenhui Guan, Yuehui Han, Song Gao, and Qing Chen. Charge trapping at the MoS_2–SiO_2 interface and its effects on the characteristics of MoS_2 metal-oxide-semiconductor field effect transistors. *Appl. Phys. Lett.*, 106(10):103109, 2015.

[23] Mirko Prezioso, Farnood Merrikh-Bayat, Brian D. Hoskins, Gina C. Adam, Konstantin K. Likharev, and Dmitri B. Strukov. Training and operation of an integrated neuromorphic network based on metal-oxide memristors. *Nature*, 521(7550):61–64, 2015.

[24] Geoffrey W. Burr, Robert M. Shelby, Severin Sidler, Carmelo di Nolfo, Junwoo Jang, Irem Boybat, Rohit S. Shenoy, Pritish Narayanan, Kumar Virwani, Emanuele U. Giacometti, Bülent N. Kurdi, and Hyunsang Hwang. Experimental demonstration and tolerancing of a large-scale neural network (165,000 synapses) using phase-change memory as the synaptic weight element. *IEEE Trans. Electron Devices*, 62(11):3498–3507, 2015.

[25] Shimeng Yu. Neuro-inspired computing with emerging nonvolatile memory. *Proc. IEEE*, 106(2):260–285, 2018.

[26] Peng Li, Xuanyu Shan, Ya Lin, Yi Du, Jiangang Ma, Zhongqiang Wang, Xiaoning Zhao, Ye Tao, Haiyang Xu, and Yichun Liu. Ultra-highly linear Ga_2O_3-based cascade heterojunctions optoelectronic synapse with thousands of conductance states for neuromorphic visual system. *Light Sci. Appl.*, 14:354, 2025.

[27] Sung Hyun Jo, Ting Chang, Idongesit Ebong, Bhavitavya B. Bhadviya, Pinaki Mazumder, and Wei Lu. Nanoscale memristor device as synapse in neuromorphic systems. *Nano Lett.*, 10(4):1297–1301, 2010.

[28] Dong-Hee Kim and Sung-Min Yoon. Improvement in energy consumption and operational stability of electrolyte-gated synapse transistors using atomic-layer-deposited HfO_2 thin films. *Mater. Sci. Semicond. Process.*, 153:107182, 2023.