An All-in-One Bio-inspired Neural Network

Shiva Subbulakshmi Radhakrishnan

Pennsylvania State University <https://orcid.org/0000-0003-1136-7425>

Akhil Dodda

The Pennsylvania State University <https://orcid.org/0000-0002-6022-4028>

Saptarshi Das (✉ sud70@psu.edu)

Pennsylvania State University <https://orcid.org/0000-0002-0188-945X>

Article

Keywords: artificial neural networks, two-dimensional layered materials, in-memory computing

Posted Date: October 25th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-969671/v1>

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

[Read Full License](#)

An All-in-One Bio-inspired Neural Network

Shiva Subbulakshmi Radhakrishnan¹, Akhil Dodda¹, and Saptarshi Das^{1,2,3,}*

¹Department of Engineering Science and Mechanics, Pennsylvania State University, University Park, PA 16802, USA

²Department of Materials Science and Engineering, Pennsylvania State University, University Park, PA 16802, USA

³Materials Research Institute, Pennsylvania State University, University Park, PA 16802, USA

Abstract: In spite of recent advancements in artificial neural networks (ANNs), the energy efficiency, multifunctionality, adaptability, and integrated nature of the biological neural networks largely remain unimitated in hardware neuromorphic computing systems. Here we exploit optoelectronic, computing, and programmable memory devices based on emerging two-dimensional (2D) layered materials such as MoS₂ to demonstrate a monolithically integrated, multi-pixel, and “all-in-one” bio-inspired neural network (BNN) capable of sensing, encoding, learning, forgetting, and inferring at miniscule energy expenditure. We also demonstrate learning adaptability and stimulate learning challenges under specific synaptic conditions to mimic biological learning. Our findings highlight the potential of in-memory computing and sensing based on emerging 2D materials, devices, and integrated circuits not only to overcome the bottleneck of von Neumann computing in conventional CMOS designs but also aid in eliminating peripheral components necessary for competing technologies such as memristors.

Biological neural networks comprising of billions of neurons connected via trillions of synapses are incredibly diverse, integrated, and energy efficient in processing information that involves sensing, encoding, storage, and computation. For example, sensory neurons receive external/internal stimuli from various sensory organs and convert the information into spike trains following various encoding algorithms, which are then communicated via interneurons to the central nervous system (CNS) where spike-based computation leads to memory formation (learning) and/or decision making (inference). Spikes are electrical impulses or digital point events in time that enable ultra-low-power neural computation as well as long-distance neural communication. Spiking activity between the pre-synaptic and post-synaptic neurons determines the potentiation or depression of their connection strengths or synaptic weights, which is ultimately responsible for learning and forgetting. Another key feature of biological neural network is neuroplasticity that allows adaptation to learning and decision making under changing environmental conditions. For example, eyes can identify patterns under both low-light (scotopic vision) as well as bright-light (photopic vision) conditions. Finally, the balance between the relative strength of potentiation and depression of synaptic connections is critical and any deviation can lead to neurological disorders including learning disability. Therefore, designing low-power neuromorphic hardware systems that resemble the functionality, organization, and plasticity of biological neural network can not only accelerate the development of hardware artificial intelligence (AI) and benefit edge computing but also offer a platform to model synaptic plasticity-related learning disorders of the CNS.

Artificial neural networks (ANNs) are highly simplified, but most prevalent abstraction of biological neural networks that have already demonstrated breakthroughs in many applications

including image classification, speech recognition, and game playing [1, 2]. However, hardware realization of ANNs using traditional complementary metal-oxide-semiconductor (CMOS) technology appears to consume orders of magnitude higher power compared to what the brain demands for similar tasks. One of the key differences is in the computing architecture, whereas CMOS-based computation embrace von Neumann architecture that physically separates the compute (logic) and storage (memory), biological neural networks dissolve such gap by placing neurons, the computational primitives, and synapses, the storage units, right next to each other.

Acknowledging the energy gap, field programmable gate arrays (FPGAs) [3] and crossbar architectures utilizing memristors [4, 5], resistive random-access memory (RRAM) [6], phase change memory (PCM) [7-9], etc. with tunable conductance states are accelerating the development of energy efficient and non von Neumann computing architectures. However, these non-von Neumann platforms still require CMOS based peripheral transducers for converting external stimuli into electrical impulses unlike biological neural networks where specialized afferent neurons transduce sensed information into electrical signals i.e. spike trains. Such pre-processing can ultimately limit the energy efficiency and scalability of emerging non von Neumann architectures [13, 14]. Finally, neuroplasticity of learning in changing environments, and modeling of learning disabilities even at a high level of abstraction is yet to be demonstrated.

Here we mitigate the aforementioned challenges by introduce a monolithically integrated, multi-pixel, and “all-in-one” bio-inspired neural network (BNN), which is capable of sensing, encoding, learning, forgetting, and inferring using monolayer MoS₂ based multifunctional field effect transistors (FETs). First, we use gate-tunable persistent photoconductivity in monolayer MoS₂

FET to convert optical information into graded potentials using a neuromorphic sensing module (SM). Next, we demonstrate MoS₂ based neuromorphic encoding module (EM) comprising of two MoS₂ FETs to transform the graded potentials into spike-count and spike-duration based programming voltages. And finally, we exploit electrical programmability of MoS₂ FET-based non-volatile synapses for realizing a neuromorphic learning module (LM) for spike-based learning, forgetting, and inference. Furthermore, we demonstrate low-power operation and adaptability of our BNN to learning under different ambient conditions mimicking neuroplasticity of biological neural networks. Our BNN hardware also offers a platform to model learning disabilities and disorders at a high level of abstraction. To the best of our knowledge, this is the first experimental demonstration of an integrated BNN exploiting in-memory computing and sensing based on emerging two-dimensional (2D) layered materials, devices, and circuits that can accelerate the development of energy efficient neuromorphic systems.

The motivation behind using two-dimensional (2D) layered MoS₂ as a hardware platform for neuromorphic computing is multifold. First, there are several demonstration of photodetectors [15], chemical sensors [16], biological sensors [16], touch sensors [17], and radiation sensors [18] using MoS₂ based devices, which can naturally serve as artificial sensory afferent neurons eliminating the need for peripheral sensors for MoS₂ based intelligent systems. Next, MoS₂ being a semiconductor, almost all peripheral analog or digital signal processing units can be build using MoS₂ FETs [19] eliminating the need for hybrid design involving CMOS circuitry. Additionally, the atomically thin body nature of MoS₂ allows aggressive channel length scaling without the loss of superior gate electrostatic benefiting high integration density. In fact, recent studies show high performance monolayer MoS₂ FETs with the channel and contact lengths scaled to 29 nm and 13

nm, respectively [20]. Moreover, some of the early criticism of 2D FETs have also been successfully addressed in recent years through the realization of low contact resistance [21], high ON current [22], integration of ultra-thin and high-k gate dielectric [23], and wafer scale growth using chemical vapor deposition (CVD) and metal organic CVD (MOCVD) [24, 25]. Similarly, MoS₂ based microprocessors [26], analogue operational amplifier [27], RF electronics components [28] and neuromorphic and biomimetic hardware platforms [29-31] have been reported. Finally, MoS₂ can enable flexible [32] and printable [39] optoelectronics adding value towards a MoS₂ based hardware platforms similar to ultra-thin silicon on insulator [33, 34].

Monolithic, multi-pixel, all-in-one, BNN platform:

Fig. 1a show neurobiological architecture for processing visual information and Fig. 1b-d, respectively, show the optical images of our multi-pixel (7×7) BNN hardware platform, each pixel comprising of four monolayer MoS₂ FETs (4T cell) that monolithically integrate the sensing module (SM), encoding module (EM), and learning module (LM), and of individual MoS₂ FETs, which are locally back-gated using a stack comprising of atomic layer deposition (ALD) grown 50 nm Al₂O₃ on sputter deposited 40/30 nm Pt/TiN. All back-gate islands were placed on a commercially purchased SiO₂/p⁺⁺-Si substrate (see *Supplementary Information 1-3* for enlarged optical images of the entire chip, 7×7 pixels, and each pixel). Within each pixel, the SM consists of 1 MoS₂ FET (T_{SM}), the EM consists of 2 MoS₂ FETs (T_{EM1} and T_{EM2}), and the LM consists of 1 MoS₂ FET (T_{LM}), which are connected using the circuit diagram shown in Fig. 1e. MoS₂ FETs used for the SMs and LMs have footprints ($W \times L$) of $5 \mu\text{m} \times 1 \mu\text{m}$ and MoS₂ FETs used for the EMs have a footprint of $5 \mu\text{m} \times 3 \mu\text{m}$ excluding the contact pads and each pixel has a footprint of $400 \mu\text{m} \times 600 \mu\text{m}$. Scalability of our devices was limited by our measurement setup, which require

large contact pads for probing the devices for demonstrative purpose. Nevertheless, the circuit schematic in Fig. 1e shows that each pixel is designed to achieve functional and organizational resemblance with different neuronal cells found in the vision pathways in primates, which is depicted schematically in Fig. 1a. For example, monolayer MoS₂ FET-based SM is functionally equivalent to photoreceptor cells (rods and cones) in the human eyes that convert external optical stimuli into corresponding graded potentials. Rods primarily enable low-light or scotopic vision whereas cones are responsible for bright-light or photopic vision, both of which can be achieved using our adaptive SM. Similarly, MoS₂ FET-based EM mimic the functionality of retinal ganglion cells that encode the graded potentials into spike trains and transmit to visual cortex or midbrain for higher order processing and computation. Finally, MoS₂ FET-based LM imitate the visual cortex where learning, forgetting, and inference take place. As we will elucidate later, the Al₂O₃/Pt/TiN gate stack shown in Fig. 1f allows non-volatile programming of our MoS₂ FETs owing to the trapping/detrapping of charge carriers at and near the MoS₂/Al₂O₃ interface when subjected to large positive and negative gate biases. This, in turn, empowers our BNN architecture to overcome the von Neumann bottleneck and enable in-memory sensing and computing capabilities, which are presently lacking for the conventional silicon technology. The programming capability is also central towards the realization of a reconfigurable BNN platform that allows adaptation to different learning conditions (e.g., scotopic conditions) similar to biological neuroplasticity, as well as offers a unique platform to model and study origin of various learning disabilities found in humans (e.g. autism disorder).

MoS₂ used in this study was obtained from 2D crystal consortium (2DCC) grown epitaxially on a sapphire substrate using MOCVD technique at 1000 °C [24, 35]. As we will elucidate, high-

Figure 1. Monolithically integrated, multi-pixel, all-in-one biomimetic spiking neural network (BNN). a) Biological neural network (BNN) for processing visual information. Optical images of b) 7×7 -pixel BNN architecture, c) each pixel comprising of four monolayer MoS₂ FETs (4T cell) that monolithically integrate the sensing module (SM), encoding module (EM), and learning module (LM), and d) individual monolayer MoS₂ field effect transistor (FET), which is locally back-gated using a stack comprising of atomic layer deposition (ALD) grown 50 nm Al₂O₃ on sputter deposited 40/30 nm Pt/TiN. All back-gate islands were placed on a commercially purchased SiO₂/p⁺⁺-Si substrate. e) Circuit schematic for each pixel showing the connection between SM, EM, and LM consisting of 1 (T_{SM}), 2 (T_{EM1} and T_{EM2}), and 1 (T_{LM}) MoS₂ FETs, respectively. MoS₂ FETs used for the SMs and LMs have footprints ($W \times L$) of $5 \mu\text{m} \times 1 \mu\text{m}$ and MoS₂ FETs used for the EMs have a footprint of $5 \mu\text{m} \times 3 \mu\text{m}$ excluding the contact pads and each pixel has a footprint of $400 \mu\text{m} \times 600 \mu\text{m}$. Each pixel is designed to achieve functional and organizational resemblance with different neuronal cells found in the visual BNN. For example, SM is functionally equivalent to photoreceptor cells (rods and cones) in the human eyes that convert external optical stimuli into corresponding graded potentials (V_{N3}) at node N_3 . Rods primarily enable low-light or scotopic vision whereas cones are responsible for bright-light or photopic vision, both of which can be achieved using T_{SM} by exploiting gate-tunable persistent photoconductivity. Similarly, EM mimics the functionality of retinal ganglion cells that encode the graded potentials into spike trains and transmit to visual cortex for further processing. Finally, LM imitates the visual cortex where learning, forgetting, and inference take place. f) 3D schematic of each MoS₂ FET with non-volatile programming capability, which empowers our BNN architecture to overcome the von Neumann bottleneck and enable in-memory sensing and computing.

temperature growth ensures high film quality and low device-to-device variability, which are critical for the successful demonstration of our BNN platform. The monolayer MoS₂ film was transferred from the growth substrate to the target application substrate, i.e. SiO₂/p⁺⁺-Si substrate with predefined islands of Al₂O₃/Pt/TiN for subsequent FET fabrication and monolithic integration of SM, EM, and LM. Details on fabrication of the back-gate stack, monolayer MoS₂ synthesis, film transfer, fabrication of MoS₂ FETs, and monolithic integration can be found in the **Methods** section.

Characterization and device-to-device variation of MoS₂ FETs:

Before diving deeper into each functional unit of our hardware BNN platform, i.e., SM, EM, and LM, it is important to thoroughly characterize the basic building blocks i.e. the MoS₂ FETs. Fig. 2a-b, respectively, show the Raman and photoluminescence (PL) spectra of a representative MoS₂ channel region. The Raman peak separation between the characteristics A_{1g} and E¹_{2g} modes of 17 cm⁻¹ and PL peak location at 1.82 eV are consistent with monolayer MoS₂ [36-38]. Fig. 2c-d, respectively, show the Raman and PL maps of 10 μm × 10 μm MoS₂ regions. Raman peak separation and PL peak position vary less than 4% over the entire map confirming high quality and uniformity of the monolayer film. In fact similar assessment on film uniformity for the entire chip can be made from Fig. 2e-f, which show Raman peak separation and PL peak position across 49 MoS₂ channels corresponding to each of the 7 × 7 pixels (see **Supplementary Information 4 & 5** for the Raman and PL scans for each of these 49 MoS₂ channels, respectively). The mean and standard deviation values were extracted to be 18 cm⁻¹ and 0.8 cm⁻¹, respectively, for Raman peak separation and 1.82 eV and 0.01 eV, respectively, for PL peak location.

Fig. 2g shows the transfer characteristics, i.e. source to drain current (I_{DS}) as a function of the local back-gate voltage (V_{BG}) at different drain biases (V_{DS}) for a representative MoS₂ FET with $L = 1$ μm and Fig. 2h shows the device-to-device variation in the transfer characteristics across 49 MoS₂ FETs corresponding to each of the 7×7 pixels (see **Supplementary Information 6** for the transfer characteristics for each of these 49 MoS₂ FETs). Note that, as expected, MoS₂ FETs show unipolar, n-type characteristics owing to the pinning of the metal Fermi level close to the conduction band allowing only electron transport through the channel. Fig. 2i show the map of electron field effect mobility values (μ_{FE}) extracted from the peak transconductance for these 49 MoS₂ FETs with mean of $\sim 21 \text{ cm}^2\text{V}^{-1}\text{s}^{-1}$ and standard deviation of $5.5 \text{ cm}^2\text{V}^{-1}\text{s}^{-1}$. Fig. 2j-l, respectively, show similar colormaps on device-to-device variation in current on/off ratio ($r_{ON/OFF}$), subthreshold slope (SS) over 3 orders of magnitude change in I_{DS} , and threshold voltage (V_{TH}) extracted at iso-current of $100 \text{ nA}/\mu\text{m}$ with mean values of 2.6×10^7 , 275 mV/decade , and 0.9 V , respectively, and standard deviation values of 0.8×10^7 , 59 mV/decade , and 0.2 V , respectively. Our μ_{FE} , $r_{ON/OFF}$, SS , and V_{TH} values and the corresponding device-to-device variations are on par with the state-of-the-art literature on large area grown MoS₂, which can be attributed to high quality growth, and relative damage-free transfer, and clean device fabrication. However, we also believe that it is possible to further reduce the device-to-device variation by improving the growth and the transfer process. **Supplementary Information 7** shows the output characteristics i.e. I_{DS} versus V_{DS} for different V_{BG} for a representative MoS₂ FET with $L = 1$ μm . While we mostly exploit off-state and subthreshold regime of FET operation in our SM, EM, and LM, on-state current reaching as high as $\sim 100 \text{ }\mu\text{A}/\mu\text{m}$ at $V_{DS} = 5 \text{ V}$ for an inversion charge carrier density of $\sim 1.5 \times 10^{13}/\text{cm}^2$ is yet another evidence of high film quality.

Figure 2. Characterization and device-to-device variation of MoS₂ FETs. a) Raman and b) photoluminescence (PL) spectra of a representative MoS₂ channel region. The Raman peak separation between the characteristics A_{1g} and E_{12g} modes of 17 cm⁻¹ and PL peak location at 1.82 eV are consistent with monolayer MoS₂. c) Raman and d) PL maps of 10 μm × 10 μm MoS₂ regions. Colormaps of distribution of e) Raman peak separation and f) PL peak position across 49 MoS₂ channels corresponding to each of the 7 × 7 pixels of our BNN architecture. The mean and standard deviation values were extracted to be 18 cm⁻¹ and 0.8 cm⁻¹, respectively, for Raman peak separation and 1.82 eV and 0.01 eV, respectively, for PL peak location. g) Transfer characteristics, i.e., source to drain current (I_{DS}) as a function of the local back-gate voltage (V_{BG}) at different drain biases (V_{DS}) for a representative MoS₂ FET with L = 1 μm. h) Device-to-device variation in the transfer characteristics across 49 MoS₂ FETs corresponding to each of the 7 × 7 pixels. Colormaps of distribution of i) electron field effect mobility values (μ_{FE}) extracted from the peak transconductance, j) current on/off ratio (r_{ON/OFF}), k) subthreshold slope (SS) over 3 orders of magnitude change in I_{DS}, and l) threshold voltage (V_{TH}) extracted at iso-current of 100 nA/μm for these 49 MoS₂ FETs. Extracted mean values for μ_{FE}, r_{ON/OFF}, SS, and V_{TH} were found to be 21 cm²V⁻¹s⁻¹, 2.6 × 10⁷, 275 mV/decade, and 0.9 V, respectively, with corresponding standard deviation values of 5.5 cm²V⁻¹s⁻¹, 0.8 × 10⁷, 59 mV/decade, and 0.2 V, respectively.

MoS₂ FET based neuromorphic sensing module (SM):

Monolayer MoS₂ based phototransistors have been studied extensively in the recent years including our own work [15, 39-43]. The phototransduction mechanism in MoS₂ FET is typically attributed to two mechanisms: photocarrier generation in the MoS₂ channel and photogating effect arising due to charge trapping/detrapping at the MoS₂/gate-dielectric interface. Fig. 3a shows the transfer characteristics with $V_{DS} = 1$ V for a representative monolayer MoS₂ FET before and after illumination from a blue light emitting diode (LED) with input currents ranging from $I_{LED} = 0.5$ mA (low-brightness) to $I_{LED} = 20$ mA (high-brightness) at different $V_{BG} = V_{write}$ for $t_{write} = 100$ ms. The corresponding incident optical power is in the range of 0.1-10 Wm⁻² obtained by calibrating using commercially purchased silicon PIN photodiode as described in **Supplementary Information 8**. Given the channel area of each MoS₂ FET used in the SM is $5 \mu\text{m} \times 1 \mu\text{m}$, the estimated incident power on each pixel is 0.5-50 pW. See **Supplementary Information 9** for the optical images showing corresponding LED brightness levels. Note that, instead of LASER illumination, conventionally used to study photoresponse in monolayer MoS₂ [43], we have used LED to provide optical stimuli since it represents more realistic lighting ambience where most neuromorphic sensors will be deployed.

Two distinct types of photoresponse are observed in Fig. 3a. For $V_{write} > 0$ V, i.e. illuminations in the on-state ($V_{write} = 2.0$ V) and in the subthreshold regime ($V_{write} = 0.5$ V) of the MoS₂ FET, there are no visible shift in the device characteristics post-illumination irrespective of the brightness level of the LED (I_{LED}). This can be ascribed to photocarrier generation in the MoS₂ channel, which are swept across by the applied V_{DS} and hence there is no persistent photocurrent beyond the optical exposure. However, for $V_{write} < 0$ V, i.e. illuminations in the off-state (V_{write}

= -1.5 V and $V_{\text{write}} = -2.5$ V) of the MoS₂ FET, there are significant shifts in the device characteristics post-illumination. This is a feature of photogating effect, where photocarrier trapping at the MoS₂/dielectric interface leads to the shift in the device threshold voltage (V_{TH}). The detrapping mechanism can be rather slow and can take hours to several days, which is why the V_{TH} shift is visible post-illumination. Higher I_{LED} , more negative V_{write} , and longer t_{write} naturally result in more photocarrier trapping (Q_{trap}) and hence larger V_{TH} shifts (ΔV_{TH}). **Supplementary Information 10** shows ΔV_{TH} and corresponding Q_{trap} ($= C_{\text{OX}}\Delta V_{\text{TH}}$) as a function of V_{write} and t_{write} for $I_{\text{LED}} = 20$ mA, where $C_{\text{OX}} \approx 2 \times 10^{-3} \text{ Fm}^{-2}$ is the back-gate oxide capacitance per unit area.

We exploit gate-tunable photogating effect in MoS₂ FET (T_{SM}) for the conversion of analog optical stimulus into graded potential, V_{N3} , at node N_3 using the circuit layout shown in Fig. 1d. A constant voltage, $V_{\text{N1}} = 5$ V, is applied to the node N_1 , which is the drain terminal of T_{SM} and a clocking signal toggling between V_{read} and V_{write} is applied to node N_2 , which is the local back-gate of T_{SM} with $\tau_{\text{CLK}} = 100$ ms. The source terminal of T_{SM} is connected to the local-back-gate of T_{EM1} at node, N_3 . Note that, T_{SM} and T_{EM1} forms a RC circuit, with T_{SM} serving as the resistor and the local back-gate of T_{EM1} serving as the capacitor. Prior to illumination, the time constant for charging the node, N_3 remains very high > 100 seconds since T_{SM} is biased in the off-state. Fig. 3b shows analog-valued and continuous-time input optical stimuli from the LED and Fig. 3c shows the corresponding temporal evolution of V_{N3} for different V_{write} with $V_{\text{read}} = 0$ V. Some key observations can be made from the results: 1) V_{N3} increases monotonically for any given I_{LED} and V_{write} owing to the photogating effect, which results in gradual negative shift in the V_{TH} of T_{SM} switching it from the off-state through subthreshold to the on-state and thereby reducing the

charging time constant for the node, N_3 , 2) for any given I_{LED} , V_{N3} increases faster for more negative V_{write} since more trap states are available at the MoS₂/dielectric interface resulting in greater V_{TH} shift and hence higher photoconductance, 3) the time required by V_{N3} to reach $V_{N1} =$

Figure 3. MoS₂ FET based neuromorphic sensing module (SM): a) Transfer characteristics of monolayer MoS₂ FET measured with $V_{DS} = 1$ V before and after illumination from a blue LED with input currents ranging from $I_{LED} = 0.5$ mA (low-brightness) to $I_{LED} = 20$ mA (high-brightness) at different $V_{BG} = V_{write}$ for $t_{write} = 100$ ms. b) Analog valued and continuous time input optical stimuli from the blue LED. c) Corresponding temporal evolution of the graded potential, V_{N3} , at node N_3 for different V_{write} obtained by using the circuit layout for SM shown in Fig. 1d. A constant voltage, $V_{N1} = 5$ V, is applied to the node N_1 , which is the drain terminal of T_{SM} and a clocking signal toggling between V_{read} and V_{write} is applied to node N_2 , which is the local back-gate of T_{SM} with $\tau_{CLK} = 100$ ms. The source terminal of T_{SM} is connected to the local-back-gate of T_{EM1} at node, N_3 . d) Time for V_{N3} to reach the same magnitude (τ_{SAT}) as a function of I_{LED} and V_{write} . e) Average energy consumption by the SM (E_{SM}) during each τ_{CLK} for different I_{LED} and V_{write} . f) Device-to-device variation in photoresponse of 49 MoS₂ FETs corresponding to the SMs of each of the 7×7 pixels of our BNN hardware after $t_{write} = 1$ second exposure to $I_{LED} = 20$ mA at $V_{write} = -2.5$ V. g) Colormap of distribution of ratio of post-illumination photoconductance to dark conductance (r_{PH}) measured at $V_{BG} = 0$ V. The mean and standard deviation values were found to be 6.7×10^3 and 3.8×10^3 , respectively.

5 V scales inversely with I_{LED} for any given V_{write} i.e. higher I_{LED} allows the graded potential to reach its maximum strength earlier and *vice versa*, and finally 4) a lower I_{LED} (e.g., 5 mA) can invoke similar response in V_{N3} like a higher I_{LED} (e.g., 20 mA), when the former is measured using a more negative $V_{\text{write}} = -2$ V compared to the later measured using $V_{\text{write}} = -1.5$ V allowing adaptation to different illumination levels. These observations are summarized in Fig. 3d, which shows the time for V_{N3} to reach the same magnitude as a function of I_{LED} and V_{write} . Also see ***Supplementary Video 1*** for time-evolution of the graded potential for various I_{LED} using different V_{write} .

Fig. 3e shows the average energy consumption by the SM (E_{SM}) during each τ_{CLK} for different I_{LED} and V_{write} . Even for the brightest LED illumination at the most negative V_{write} , $E_{\text{SM}} \sim 50$ fJ, which suggests energy efficient phototransduction by our MoS₂ FET based SM. Finally, Fig. 3f show the pre- and post-illumination transfer characteristics of 49 MoS₂ FETs corresponding to the SMs of each of the 7×7 pixels of our BNN hardware after $t_{\text{write}} = 1$ second exposure to $I_{\text{LED}} = 20$ mA at $V_{\text{write}} = -2.5$ V and Fig. 3g shows the colormap of ratio of post-illumination photoconductance to dark conductance (r_{PH}) measured at $V_{\text{BG}} = 0$ V (see ***Supplementary Information 11*** for the pre- and post-illumination transfer characteristics for each of these 49 MoS₂ FETs). The mean and standard deviation values were found to be 6.7×10^3 and 3.8×10^3 , respectively. Note that this is the first report on device-to-device variation in the photoresponse of MOCVD grown monolayer MoS₂ FETs.

MoS₂ FET-based neuromorphic encoding module (EM):

EM converts the graded potentials received from the SM into corresponding programming waveforms using spike-count and spike-duration based algorithms and transmit to the LM as summarized in Fig. 4a-j. Each EM comprises of two MoS₂ FETs, T_{EM1} and T_{EM2} connected in series as shown in Fig. 1d. Note that the local back-gate of T_{EM1} is shorted to its source at node, N_5 , which is also the drain of T_{EM2} and the output node of the EM. The drain terminal of T_{EM1} , i.e., N_4 is kept grounded, and the programming voltage, V_P is applied to the source terminal of T_{EM2} , i.e., N_6 . Furthermore, both FETs are pre-programmed such that T_{EM1} operate as a depletion mode (normally on) n-channel FET, whereas, T_{EM2} operates as an enhancement mode (normally off) n-channel FET as shown in Fig. 4a. The EM module, therefore, serves as an NMOS inverter with depletion load. Fig. 4b shows the input (V_{N3}) versus output (V_{N5}) characteristics of the EM for

different V_P values. Also note that the inverting threshold (V_{IT}), i.e., the magnitude of V_{N3} at which V_{N5} reaches $V_P/2$ can be adjusted by reconfiguring T_{EM1} and T_{EM2} . Fig. 4c-d, respectively, show the various programming states of T_{EM2} and corresponding EM characteristics for $V_P = -5$ V. Tunability in the EM characteristic is an additional benefit of our BNN platform allowing adaptation to various learning conditions as we will elucidate later. Finally, a constant V_P applied to node N_6 transforms the graded potential into spike-duration based programming voltage as shown in Fig. 4e, whereas a clocking signal toggling between V_P and 0 V with $\tau_{CLK} = 100$ ms applied to node N_6 transforms the graded potential into spike-count based programming voltage as shown in Fig. 4f (see **Supplementary Information 12** for the biasing configuration of EM for spike-duration and spike-count based encoding). Fig. 4g and Fig. 4i, respectively show the total spiking time (τ_{spike}), and total number of spikes (N_{spike}) as a function of V_{write} and I_{LED} , when

Figure 4. MoS₂ FET-based neuromorphic encoding module (EM). a) Transfer characteristics of MoS₂ FETs used as T_{EM1} and T_{EM2} in the EM. T_{EM1} is programmed to operate as a depletion mode (normally on) n-channel FET, whereas T_{EM2} operates as an enhancement mode (normally off) n-channel FET. Based on the circuit layout shown in Fig. 1d, the EM module serves as an NMOS inverter with depletion load. b) Input (V_{N3}) versus output (V_{N5}) characteristics of the EM for different V_P values applied to the source terminal of T_{EM2} , i.e., N_6 . The drain terminal of T_{EM1} , i.e., N_4 is kept grounded. c) Various programming states of T_{EM2} and d) corresponding EM characteristics for $V_P = -5$ V. The inverting threshold (V_{IT}), i.e., the magnitude of V_{N3} at which V_{N5} reaches $V_P/2$ can be adjusted by reconfiguring T_{EM1} and T_{EM2} . e) Spike-duration and f) spike-count based encoding of graded potential (V_{N3}) received from SM corresponding to different V_{write} and I_{LED} into programming voltage (V_{N5}) for transmitting to the learning module (LM). g) Total spiking time (τ_{spike}) and h) corresponding average encoding energy expenditure (E_{EM}) per clock cycle for spike-duration based encoding and i) total number of spikes (N_{spike}) and j) corresponding E_{EM} for spike-count based encoding as a function of V_{write} and I_{LED} . The input stimulus is presented for a duration of 10 s and $V_P = -6$ V. Spike-duration and spike numbers are counted once V_{N5} reaches 75% of V_P , i.e., -4.5 V.

the input stimulus is presented for a duration of 10 s and $V_p = -6$ V. Note that we start to count the spike time and spike number once V_{N5} reaches 75% of V_p , i.e. -4.5 V in Fig. 4e-f. As expected, for any given V_{write} , graded potentials received from the SM module that corresponds to higher values of input stimuli (I_{LED}) invoke longer τ_{spike} and more N_{spike} at the output of the EM for spike-duration and spike-count based encodings, respectively. Similarly, for any given I_{LED} , more negative V_{write} invokes longer τ_{spike} and more N_{spike} . Note that τ_{spike} and N_{spike} can also be controlled by configuring V_{IT} (Fig. 4d). For example, scotopic condition will benefit from lower V_{IT} since spikes will reach the programming voltage, V_p earlier for any given I_{LED} and V_{write} . Alternatively, a higher V_p value can be used to encode shorter duration or lower number of programming spikes. See *Supplementary Information 13* for encoding of the same graded potential using different V_p for both spike-duration and spike-count based adaptive encoding. The reconfigurability of the EM can also be exploited for modeling learning disabilities. For example if bright light is encoded into low-magnitude V_p spikes, potentiation of synapses can be severely limited invoking learning difficulty. Finally, the average encoding energy expenditure (E_{EM}) per clock cycle by the EM for spike-count and spike-duration based encoding module for different I_{LED} and V_{write} are shown in Fig. 4h and Fig. 4j, respectively. The relatively higher energy expenditure for the EM is a direct consequence of using a depletion mode NMOS inverter and can be reduced significantly by using a CMOS inverter. This will require the development of p-channel MoS₂ or use of another 2D material such as WSe₂.

MoS₂ FET-based neuromorphic learning module (LM):

Optical information encoded in spikes are delivered from the EM to the LM for pattern learning using memory augmented reinforcement. Each learning module comprises of one MoS₂ FET (T_{LM})

as shown in Fig. 1d, which serves as a non-volatile synapse with analog conductance states programmable by applying electrical voltage spikes to the local back-gate terminal, N_5 , which also serves as the output terminal of EM. Fig. 5a-o show that MoS₂ FET-based synapses allow both spike-count and spike-duration based non-volatile programming and can achieve both potentiation and depression analogous to chemical synapses in BNNs with low programming energy expenditure.

Fig 5a shows the potentiation of a representative MoS₂ electrical synapse from a low conductance state (LCS) after the application of a fixed number of programming spikes ($N_{\text{spike}} = 10$) of different amplitudes of negative polarity (V_P) with each spike applied for $t_{\text{spike}} = 100$ ms. Fig. 5b show the post-potentiated conductance states (G_P) measured at $V_{BG} = 0$ V as a function of N_{spike} for different V_P . As expected, lower number of N_{spike} invoke lower potentiation, i.e. smaller change in G_P and *vice versa*, which can be exploited for spike-count based learning. Similarly, Fig 5c shows the depression of a potentiated MoS₂ synapse i.e. from high conductance state (HCS) after the application of a fixed number of programming spikes ($N_{\text{spike}} = 10$) of different amplitudes of positive polarity (V_D) with each spike applied for $t_{\text{spike}} = 100$ ms. Fig. 5d show the post-depressed conductance states (G_D) measured at $V_{BG} = 0$ V as a function of N_{spike} for different V_D . As expected, lower number of N_{spike} invoke lower depression and *vice versa*, which can be exploited for spike-count based forgetting. As we will elucidate later forgetting capabilities enable relearning of new patterns using the same synapses that have learned a previous pattern. Also note that smaller number of N_{spike} can achieve higher potentiation/depression if encoded using higher $V_{P/D}$. As mentioned earlier, this aspect can be exploited to achieve learning plasticity.

Fig 5e-h show the spike-duration based potentiation and depression of MoS₂ synapses. Fig 5e shows the potentiation of MoS₂ synapse from LCS after the application of single spike of constant magnitude $V_P = -6$ V for different t_{spike} and Fig. 5f show the post-potentiated G_P measured at $V_{BG} = 0$ V as a function of t_{spike} for different V_P . Similarly, Fig 5g shows the depression of MoS₂ synapse from HCS after the application of single spike of constant magnitude $V_D = 6$ V for different t_{spike} and Fig. 5h show the post-depressed G_D measured at $V_{BG} = 0$ V as a function of t_{spike} for

Figure 5. Analog and non-volatile programming of MoS₂ FET-based synapses. a) Potentiation of a MoS₂ synapse from low conductance state (LCS) after the application of a fixed number of programming spikes ($N_{\text{spike}} = 10$) of different amplitudes of negative polarity (V_P) with each spike applied for $t_{\text{spike}} = 100$ ms. b) Post-potentiated conductance states (G_P) measured at $V_{BG} = 0$ V as a function of N_{spike} for different V_P . c) Depression of a MoS₂ synapse from high conductance state (HCS) after the application of a fixed number of programming spikes ($N_{\text{spike}} = 10$) of different amplitudes of positive polarity (V_D) with each spike applied for $t_{\text{spike}} = 100$ ms. d) Post-depressed conductance states (G_D) measured at $V_{BG} = 0$ V as a function of N_{spike} for different V_D . e) Potentiation of MoS₂ synapse from LCS after the application of single spike of constant magnitude $V_P = -6$ V for different t_{spike} . f) Post-potentiated G_P measured at $V_{BG} = 0$ V as a function of t_{spike} for different V_P . g) Depression of MoS₂ synapse from HCS after the application of single spike of constant magnitude $V_D = 6$ V for different t_{spike} . h) Post-depressed G_D measured at $V_{BG} = 0$ V as a function of t_{spike} for different V_D . Retention characteristics for i) 6 potentiated and j) 6 depressed conductance states for 100 seconds. k) Device-to-device variation in the pre- and post-programmed transfer characteristics and l) corresponding colormap of distribution of memory ratio (MR) measured at $V_{BG} = 0$ V for 49 monolayer MoS₂ FETs from each LM of our 7×7 BNN platform when programmed using $N_{\text{spike}} = 10$ with spike magnitude, $V_P = -8$ V, and spike width, $t_{\text{spike}} = 100$ ms. The mean and standard deviation values for MR were found to be 6×10^5 and standard deviations of 0.5×10^5 , respectively.

different V_D . Here, shorter t_{spike} invokes lower potentiation/depression and *vice versa*, which can be used for spike-duration based learning/forgetting. Note that similar to spike-count based learning/forgetting, higher potentiation/depression can be achieved for shorter spike durations when encoded using higher magnitude of $V_{P/D}$ enabling spike-duration based learning plasticity under scotopic condition.

The underlying mechanism behind the spike-count and spike-duration based potentiation and depression of MoS₂ synapses can be explained using the shift in V_{TH} observed in the transfer characteristics of MoS₂ FETs. The V_{TH} shift is attributed to charge trapping/detrapping at and near the MoS₂/Al₂O₃ interface, which is also responsible for the photogating effect described earlier. Negative shift in the threshold voltage (V_{TH}) with increasing magnitude of V_P and positive shift in the V_{TH} with increasing magnitude of V_D observed in the transfer characteristics of MoS₂ FET are indicative of electron trapping and de-trapping in the local back-gate stack, respectively. Interestingly, the trapping and de-trapping processes were found to be non-volatile as evident from the retention measurements displayed in Fig. 5i-j for 6 representative potentiated (G_P) and depressed (G_D) conductance states, respectively, for 100 seconds. We also examined long-term memory retention characteristics of two representative post-programmed analog conductance states for $\sim 10^4$ seconds as shown in **Supplementary Information 14**. The memory ratio (MR) between these two states was found to change from $\sim 1.1 \times 10^2$ to 0.6×10^2 following an exponential decay with time constant of 1.6×10^4 seconds. The projected time before the two states become indistinguishable or MR reaches 1 is found to be ~ 1 day. Note that, while conventional memory devices require non-volatile retention for years, many neuromorphic applications including those used by edge devices and sensors relax the requirement for long-term retention and can be well

served with short-term memory retention of several hours to days. The retention window demonstrated by the MoS₂ FETs was adequate for the successful realization of our proof-of-concept “*all-in-one*” BNN. Certainly, it is desirable and possible to improve the memory retention window by optimizing the design of the local back-gate stack e.g., by mimicking the floating-gate architecture used by the conventional FLASH memory devices.

The device-to-device variation in the pre and post-programmed transfer characteristics and corresponding colormap of MR measured at $V_{BG} = 0$ V for 49 monolayer MoS₂ FETs from each LM of our 7×7 BNN platform when programmed using $N_{spike} = 10$ with spike magnitude, $V_p = -8$ V, and spike width, $t_{spike} = 100$ ms are shown in Fig. 5k-l, respectively (see **Supplementary Information 15** for the pre- and post-programmed transfer characteristics for each of these 49 MoS₂ FETs). The mean and standard deviation values for MR were found to be 6×10^5 and standard deviations of 0.5×10^5 , respectively. Similar to photoresponse, this is the first report on device-to-device variation in the programmability of monolayer MoS₂ FETs.

Finally, Fig. 6a-b, respectively, show the spike-duration and spike-count based conductance evolution in MoS₂ FET based LM when input programming waveforms (V_{N5}) are received from the EM corresponding to different I_{LED} and V_{write} . As shown in Fig. 6c-d, for any given V_{write} , the spiking patterns received from the EM module that correspond to higher values of input stimuli (I_{LED}) results in higher values of final conductance and *vice versa* for both spike-duration and spike-count based learning, respectively. Similarly, for any given I_{LED} , more negative V_{write} results in better learning, i.e., higher final conductance. The average learning energy expenditure (E_{LM}) per clock cycle by the LM for spike-duration and spike-count based learning for different

I_{LED} and V_{write} are shown in Fig. 6e-f, respectively. The energy expenditure for the LM is found to be miniscule, ~ 50 fJ per clock cycle even for the brightest illumination and most negative V_{write} .

Figure 6. MoS₂ FET-based neuromorphic learning module (LM). a) Spike-duration and b) spike-count based conductance evolution in MoS₂ FET based LM when input programming waveforms (V_{N5}) are received from the EM corresponding to different I_{LED} and V_{write} . Final conductance state achieved by the LM for c) spike-duration and d) spike-count based input spiking patterns received from the EM module corresponding to different I_{LED} and V_{write} . The average learning energy expenditure (E_{LM}) per clock cycle by the LM for e) spike-duration and f) spike-count based learning for different I_{LED} and V_{write} .

Multi-pixel demonstration of analog image sensing, encoding, and leaning:

Fig. 7a-d and **Supplementary Video 2** show a complete demonstration of our BNN hardware involving multi-pixel and monolithically integrated SM, EM, and LM. A 7×7 analog input pattern obtained by illuminating the LED (Fig. 7a) is transduced into corresponding graded potentials using the SMs (Fig. 7b) and encoded into corresponding programming spikes following the spike-

count based encoding algorithm by the EMs (Fig. 7c), which are subsequently used by the LMs to potentiate the MoS₂ FET-based non-volatile synapses (Fig. 7d). Clearly, the input LED pattern is learnt by our 7×7 BNN hardware. For this demonstration, all synapses were initially programmed in their LCS and different LED illuminations were presented one by one to the corresponding pixels of our 7×7 BNN hardware. For simultaneous illumination, a lensing system will be needed to focus the image pixels on to the corresponding SMs of our BNN platform. In our future endeavor we will attempt to integrate the lensing system with the BNN hardware. Note that MoS₂ FETs used in the SMs are biased in the deep off-state by applying negative V_{write} to harness the photogating

Figure 7. Multi-pixel demonstration of analog image sensing, encoding, and learning. a) Analog 7×7 input pattern obtained by illuminating the blue LED. Temporal evolution of corresponding b) graded potential (V_{N3}) at the output of SMs, c) programming spike-count (N_{spike}) at the output of the EMs, and d) programmed conductance values at the output of LMs. Clearly, the input LED pattern is learnt by our 7×7 BNN hardware. For this demonstration, all MoS₂ FET based synapses belonging to the LMs were initially programmed in their LCS (100 pS) and different LED illuminations were presented one by one to the corresponding pixels of our 7×7 BNN hardware.

effect that results in the transduction of optical illuminations into corresponding graded potentials. However, the MoS₂ FETs used in the EM and LM are biased either in the subthreshold or in the on-state, where these devices remain insensitive to illumination and hence their operation is not impacted by illumination. Also note that, while the input pattern is learnt by our BNN architecture, device-to-device variation in the photogating effect, transfer characteristics, and programming of MoS₂ FETs are translated into variation in the graded potential, spike-count, and learnt conductance values corresponding to the same input LED signal as seen in Fig. 7. There is no doubt that further reduction in device-to-device variation is desirable. As MoS₂ technology matures further through the optimization of growth conditions to reduce point-defects etc., and cleaner and damage-free techniques are developed for large area transfer and polymer residues are eliminated from the device fabrication it will be possible to mitigate the device-to-device variation to a large extent and achieve near-ideal learning. Nevertheless, our proof-of-concept demonstration highlights the fully integrated nature of our MoS₂ FET-based hardware BNN that combines sensing, computing (encoding), and storage and thereby distinguishes it from other hardware BNN architectures based on CMOS or emerging technologies such as RRAM, PCM, memristor, all-optic, as well as hybrid approaches.

Importance of forgetting (synaptic depression) in learning and inference:

Forgetting has traditionally been considered to be a passive brain process, which ensures unused information fade over time so that neural resources can be reallocated for storing more important and newer information. When machines learn with unrestricted storage resources (e.g. cloud servers), forgetting is irrelevant. However, when storage capacity is either limited or not accessible, for example in internet of things (IoT) edge devices deployed in remote locations,

forgetting can play an active role in smart learning. Here we demonstrate the role of forgetting in relearning without any external supervision and by directly interacting with the changing environment.

Fig. 8a shows the schematic and optical image of a fully connected 2-layer BNN with 9 presynaptic input neurons and 1 postsynaptic output neuron for learning and inferring patterns from 3×3 pixelated images (see *Supplementary Information 16* for an enlarged version of the optical image). Fig. 8b shows the training and retraining schedule consisting of total $M = 40$ epochs, with each epoch having two cycles: potentiation and depression. See *Supplementary Information 17* showing the biasing configuring of the EM for introducing depression cycle. To introduce depression in EM, V_D is applied to the drain terminal of T_{EM1} , i.e. N_4 instead of keeping it grounded with clocking profiles as shown in Fig. S17a-b for spike-duration and spike-count based encoding, respectively. Fig. S17c-d, respectively, show the output (V_{N5}) of EM for a constant graded potential, $V_{N3} = 5$ V and for various combinations of V_D and V_P for spike-duration and spike-count based encoding, respectively. During the potentiation cycle, the image pattern to be learnt is presented to the corresponding synapses of the 9×1 BNN, whereas during the depression cycle all 9 synapses are uniformly depressed. The first pattern (left diagonal) is presented for 20 epochs followed by the second pattern (right diagonal) for another 20 epochs to test whether our BNN can forget previously learned pattern and relearn new patterns. Fig. 8c-d, respectively, show the spiking profiles used for spike-count and spike-duration based learning. For each type of learning, we consider three configurations of the BNN: 1) weak potentiation and strong depression, 2) strong potentiation and weak depression, and 3) strong potentiation and strong depression. For spike-count based learning, the strength of potentiation (V_P) and depression (V_D) are adjusted using the

Figure 8. Importance of forgetting (synaptic depression) in learning and inference. a) Schematic and optical image of a 2-layer BNN with 9 presynaptic neurons and 1 postsynaptic neuron for learning and inferring patterns from 3×3 pixelated images. b) Training and retraining schedule with $M = 40$ epochs, with each epoch having potentiation and depression cycles. During the potentiation, the pattern to be learned is presented to the BNN, whereas during the depression all synapses are uniformly depressed. Spiking profiles used for c) spike-count and d) spike-duration based learning. For each type of learning, three BNN configurations are used: 1) weak potentiation and strong depression, 2) strong potentiation and weak depression, and 3) strong potentiation and strong depression. The strength of potentiation (V_P) and depression (V_D) are adjusted using the spike magnitude and spike duration for spike-count and spike-duration based learnings, respectively. The time evolution of colormap of synaptic weights i.e., the conductance states of the 9 synapses during e) spike-count and f) spike-duration based learning. For each type of learning all synapses are initialized either in a high conductance state (HCS) with $G_{HCS} = 100$ nS, or a low conductance state (LCS) with $G_{LCS} = 100$ pS (also see the **Supplementary Video 3 and 4**). Learning of the left diagonal followed by relearning of the right diagonal when potentiation and depression are both strong for g) spike-count and h) spike-duration based learnings (also see the **Supplementary Video 5 and 6**).

spike magnitude, for example $V_P = -10$ V for strong and $V_P = -8$ V for weak potentiation and $V_D = 12$ V for strong and $V_D = 10$ V for weak depression. Similarly, for the pattern to be learned, each pixel in the 3×3 images is encoded with $N_{\text{spike}} = 10$ if it is bright and $N_{\text{spike}} = 0$ if it is dark. For spike-duration based learning the strength of potentiation and depression are adjusted using the spike duration, i.e. $t_{\text{spike}} = 800$ ms for strong and $t_{\text{spike}} = 100$ ms for weak potentiation/depression and for the pattern to be learned, each pixel in the 3×3 images are encoded with respective t_{spike} (weak/strong) if it is bright and $t_{\text{spike}} = 10$ ms if it is dark.

Fig. 8e-f, respectively, show the time evolution of colormap of synaptic weights i.e. the conductance states of the 9 synaptic devices during the spike-count and spike-duration based learning cycles. For each type of learning all synapses are initialized either in a high conductance state (HCS) with $G_{\text{HCS}} = 100$ nS, or a low conductance state (LCS) with $G_{\text{LCS}} = 100$ pS (also see the *Supplementary Video 3 and 4*). Following are the key observations. When potentiation is weak but depression is strong, it is difficult to learn irrespective of the initial state of the synapses, however, when potentiation is strong but depression is weak, learning from LCS is fast, but forgetting and hence relearning from HCS is slow. This is expected since synapses that are potentiated get stuck in their HCS owing to weak depression making it difficult for them to forget their respective states. Finally, if both potentiation and depression are strong, learning and forgetting become faster irrespective of the initial synaptic state. This is demonstrated in Fig. 8g-h, which show learning of the left diagonal followed by relearning of the right diagonal when potentiation and depression are both strong for spike-count and spike-duration based learnings, respectively (also see the *Supplementary Video 5 and 6*). Our findings indicate that the relative strengths of potentiation and depression play critical role in learning using BNN. This is similar to

BNN, where autism, or autism spectrum disorder (ASD), which includes a broad range of conditions such as challenges with learning social skills, repetitive behaviors, etc. are related to dysregulation or deficit in long term depression in several mouse models [44, 45]. Therefore, our hardware BNN platform offers a unique opportunity to bridge the gap between neuroscience of learning and machine learning.

Supplementary Information 18 shows inference using our BNN architecture. We have used a 9×2 fully connected neural network implemented using two sets of 9×1 synapses as shown in Fig. S18a. The synapses between the 9 presynaptic neurons and the “Yes” postsynaptic neuron are trained with the actual pattern, whereas the synapses between the 9 presynaptic neurons and the “No” postsynaptic neuron are trained with the inverse of the pattern to obtain the respective conductance maps ($G_{i-\text{Yes/No}}$, $i = 1, 2, 3, \dots, 8, 9$) as shown in Fig. S18b. Any input pattern from the LED is converted to corresponding graded potentials by the SMs and transduced into spike trains by the EMs. For spike-count based inference, the output voltage spikes (V_{ij} , $i = 1, 2, 3, \dots, 8, 9$; $j = 1, 2, 3, \dots, N_{\text{spike}}$) obtained at the output of the encoding module corresponding to each pixel of the 3×3 image are applied to the drain terminals of the 9 presynaptic neurons. The output currents from the common source terminal i.e. post-synaptic “Yes” and “No” neurons are integrated using capacitors ($C_{\text{Yes/No}}$) to obtain V_{Yes} and V_{No} as shown in Fig. S18c following Eq.1.

$$V_{\text{Yes}} = \frac{t_{\text{spike}}}{C_{\text{Yes}}} \sum_{i=1}^9 \sum_{j=1}^{N_{\text{spike}}} G_{i-\text{Yes}} V_{ij} \quad V_{\text{No}} = \frac{t_{\text{spike}}}{C_{\text{No}}} \sum_{i=1}^9 \sum_{j=1}^{N_{\text{spike}}} G_{i-\text{No}} V_{ij} \quad [1]$$

For spike-duration based inference, a similar approach is adopted, except for the fact that only one voltage spike (V_i , $i = 1, 2, 3, \dots, 8, 9$) is obtained at the output of the encoding module corresponding

to each pixel of the 3×3 image with different spiking durations. In this case, V_{Yes} and V_{No} are given by Eq. 2.

$$V_{\text{Yes}} = \frac{1}{C_{\text{Yes}}} \sum_{i=1}^9 \int_0^{t_{\text{spike}}} G_{i-\text{Yes}} V_i \quad V_{\text{No}} = \frac{1}{C_{\text{No}}} \sum_{i=1}^9 \int_0^{t_{\text{spike}}} G_{i-\text{No}} V_i \quad [2]$$

For the “Yes” neuron to be a winner, $V_{\text{Yes}} > V_{\text{No}}$ and $V_{\text{Yes}} \geq V_{\text{Win}}$, where V_{Win} is the winning threshold determined by the learned pattern. Clearly, the “Yes” neuron should be the winner only when the pattern similar to the learned one is inferred, whereas the “No” neuron should win for all other patterns. However, the experimental inference accuracy was found to be $\sim 96\%$. This is because the patterns which contain one or two off-diagonal pixels in addition to the diagonal pixels also make the “Yes” neuron the winner. There are total ${}^6C_1 + {}^6C_2 = 21$ such patterns, which accounts for $\sim 4\%$ of all $2^9 = 512$ patterns that are wrongly inferred. Note that if 3 or more pixels in addition to the diagonal pixels are bright, the “No” neuron wins. The inference accuracy was improved to 100% by making $V_{\text{No}} \geq V_{\text{Win}}$ even when only one off-diagonal pixel is present in the input pattern. This was accomplished through greater potentiation of the synaptic connections between the input neurons and the “No” neuron during the training with the inverse pattern resulting in an order of magnitude higher learned conductance value.

Conclusion

In conclusion, we have experimentally demonstrated a fully integrated, multi-pixel, and biomimetic BNN hardware platform based on monolayer MoS_2 that combines sensing, encoding, learning, and inference. We have employed both spike-count and spike-duration based encoding, learning, and inference inspired by the energy efficiency of spike-based computing in the brain. Similarly, we were able to show adaptive learning in photopic and scotopic conditions and impact

of relative strengths of synaptic potentiation and depression on learning and forgetting. Our accomplishments can be attributed to the unique photoresponse of monolayer MoS₂ based phototransistors for sensing, uniquely designed MoS₂ based neuromorphic circuit modules for encoding, and programmable and non-volatile MoS₂ synapses enabled by our local back-gate memory stack for unsupervised and adaptive learning. Our findings highlight the potential of in-memory computing and sensing based on emerging 2D materials, devices, and circuits that not only overcome the bottleneck of von Neumann computing in conventional CMOS designs but also aid in eliminating peripheral components necessary for competing technologies such as memristors, RRAM, PCM, etc. We believe that our MoS₂ based low-power and fully integrated hardware BNN system is more bio-realistic in terms of functionality, organization, and plasticity of BNN and, therefore, can not only accelerate the development of hardware artificial intelligence (AI) and benefit edge computing and smart sensing for Internet of Things (IoT), but also offer a platform for adaptive learning and for modeling plasticity-related learning disorders of the BNNs.

Methods

Fabrication of local back-gate islands: To define the back-gate island regions, the substrate 285 nm SiO₂ on p⁺⁺-Si was spin coated with bilayer photoresist consisting of Lift-Off-Resist (LOR 5A) and Series Photoresist (SPR 3012) baked at 185 °C and 95 °C, respectively. The bilayer photoresist was then exposed to Heidelberg Maskless Aligner (MLA 150) to define the island and developed using MF CD26 microposit, followed by a de-ionized (DI) water rinse. The back gate electrode of 20/50 nm TiN/Pt was deposited using reactive sputtering. The photoresist was removed using acetone and Photo Resist Stripper (PRS 3000) and cleaned using 2-propanol (IPA) and DI water. Atomic layer deposition (ALD) process was then implemented to grow 50 nm Al₂O₃ on the entire substrate including the island regions. To access the individual Pt back-gate electrodes etch patterns were defined using the same bilayer photoresist consisting of LOR 5A and SPR 3012. The bilayer photoresist was then exposed to MLA 150 and developed using MF CD26 microposit. 50 nm Al₂O₃ was subsequently dry etched using the BCl₃ chemistry at 5 °C for 20 seconds, which was repeated four times to minimize heating in the substrate. Next the photoresist was removed to give access to the individual Pt electrodes.

Large area monolayer MoS₂ film growth: Monolayer MoS₂ was deposited on epi-ready 2” c-sapphire substrate by metalorganic chemical vapor deposition (MOCVD). An inductively heated graphite susceptor equipped with wafer rotation in a cold-wall horizontal reactor was used to achieve uniform monolayer deposition as previously described [46]. Molybdenum hexacarbonyl (Mo(CO)₆) and hydrogen sulfide (H₂S) were used as precursors. Mo(CO)₆ maintained at 10°C and 950 Torr in a stainless-steel bubbler was used to deliver 0.036 sccm of the metal precursor for the growth, while 400 sccm of H₂S was used for the process. MoS₂ deposition was carried out at

1000°C and 50 Torr in H₂ ambient, where monolayer growth was achieved in 18 min. The substrate was first heated to 1000°C in H₂ and maintained for 10 min before the growth was initiated. After growth, the substrate was cooled in H₂S to 300°C to inhibit decomposition of the MoS₂ films. More details can be found in our earlier work [35, 40, 47].

MoS₂ film transfer to local back-gate islands: To fabricate the MoS₂ FETs, MOCVD grown monolayer MoS₂ film was transferred from the sapphire to SiO₂/p⁺⁺-Si substrate with local back-gate islands using PMMA (polymethyl-methacrylate) assisted wet transfer process. First, MoS₂ on sapphire substrate was spin coated with PMMA and then baked at 180 °C for 90 s. The corners of the spin-coated film were scratched using a razor blade and immersed inside 1 M NaOH solution kept at 90 °C. Capillary action causes the NaOH to be drawn into the substrate/film interface, separating the PMMA/ MoS₂ film from the sapphire substrate. The separated film was rinsed multiple times inside a water bath and finally transferred onto the SiO₂/p⁺⁺-Si substrate with local back-gate islands and then baked at 50 °C and 70 °C for 10 min each to remove moisture and residual PMMA, ensuring a pristine interface.

Fabrication of monolayer MoS₂ FET: To define the channel regions for the MoS₂ FETs, the substrate was spin-coated with PMMA and baked at 180 °C for 90 s. The resist was then exposed to electron beam (e-beam) and developed using 1:1 mixture of 4-methyl-2-pentanone (MIBK) and 2 propanol (IPA). The monolayer MoS₂ film was subsequently etched using sulfur hexafluoride (SF₆) at 5 °C for 30 s. Next, the sample was rinsed in acetone and IPA to remove the e-beam resist. To define the source and drain contacts, sample is then spin coated with methyl methacrylate (MMA) followed by A3 PMMA. Then using e-beam lithography source and drain contacts are

patterned and developed by using 1:1 mixture of MIBK and IPA for 60s. 40 nm of Nickel (Ni) and 30 nm of Gold (Au) are deposited using e-beam evaporation. Finally, lift-off process is performed to remove the evaporated Ni/Au except from the source/drain patterns by immersing the sample in acetone for 30 min followed by IPA for another 30 mins. Each island contains one MoS₂ FET to allow for individual gate control.

Monolithic Integration: Each pixel of our multi-pixel (7×7) BNN hardware consist of 4 MoS₂ FETs as shown using the circuit schematic in Fig. 1e. Within each pixel, the SM consists of 1 MoS₂ FET (T_{SM}), the EM consists of 2 MoS₂ FETs (T_{EM1} and T_{EM2}), and the LM consists of 1 MoS₂ FET (T_{LM}). To define the connections between the respective nodes of T_{SM} , T_{EM1} , T_{EM2} , and T_{LM} , the substrate was spin coated with MMA and PMMA, followed by the e-beam lithography and developing using 1:1 mixture of MIBK and IPA, and e-beam evaporation of 60 nm Au. Finally, the e-beam resist was rinsed away by lift-off process using acetone and IPA.

Electrical Characterization: Electrical characterization of the fabricated devices are performed using Lake Shore CRX-VF probe station under atmospheric condition using a Keysight B1500A parameter analyzer.

Data Availability: The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Code Availability: The codes used for plotting the data are available from the corresponding authors on reasonable request.

References

- [1] M. H. Hassoun, *Fundamentals of artificial neural networks*: MIT press, 1995.
- [2] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, *et al.*, "Mastering the game of Go without human knowledge," *Nature*, vol. 550, pp. 354-359, 2017/10/01 2017.
- [3] B. Glackin, T. M. McGinnity, L. P. Maguire, Q. Wu, and A. Belatreche, "A novel approach for the implementation of large scale spiking neural networks on FPGA hardware," in *International Work-Conference on Artificial Neural Networks*, 2005, pp. 552-563.
- [4] Z. Wang, S. Joshi, S. Savel'ev, W. Song, R. Midya, Y. Li, *et al.*, "Fully memristive neural networks for pattern classification with unsupervised learning," *Nature Electronics*, vol. 1, pp. 137-145, 2018.
- [5] M. Al-Shedivat, R. Naous, G. Cauwenberghs, and K. N. Salama, "Memristors empower spiking neurons with stochasticity," *IEEE journal on Emerging and selected topics in circuits and systems*, vol. 5, pp. 242-253, 2015.
- [6] V. Milo, G. Pedretti, R. Carboni, A. Calderoni, N. Ramaswamy, S. Ambrogio, *et al.*, "Demonstration of hybrid CMOS/RRAM neural networks with spike time/rate-dependent plasticity," in *2016 IEEE International Electron Devices Meeting (IEDM)*, 2016, pp. 16.8. 1-16.8. 4.
- [7] S. Kim, M. Ishii, S. Lewis, T. Perri, M. BrightSky, W. Kim, *et al.*, "NVM neuromorphic core with 64k-cell (256-by-256) phase change memory synaptic array with on-chip neuron circuits for continuous in-situ learning," in *2015 IEEE international electron devices meeting (IEDM)*, 2015, pp. 17.1. 1-17.1. 4.
- [8] I. Boybat, M. Le Gallo, S. Nandakumar, T. Moraitis, T. Parnell, T. Tuma, *et al.*, "Neuromorphic computing with multi-memristive synapses," *Nature communications*, vol. 9, pp. 1-12, 2018.
- [9] S. Ambrogio, N. Ciochini, M. Laudato, V. Milo, A. Pirovano, P. Fantini, *et al.*, "Unsupervised learning by spike timing dependent plasticity in phase change memory (PCM) synapses," *Frontiers in neuroscience*, vol. 10, p. 56, 2016.
- [10] P. Yao, H. Wu, B. Gao, S. B. Eryilmaz, X. Huang, W. Zhang, *et al.*, "Face classification using electronic synapses," *Nature communications*, vol. 8, pp. 1-8, 2017.
- [11] M. D. Pickett, G. Medeiros-Ribeiro, and R. S. Williams, "A scalable neuristor built with Mott memristors," *Nature materials*, vol. 12, pp. 114-117, 2013.
- [12] P. M. Sheridan, F. Cai, C. Du, W. Ma, Z. Zhang, and W. D. Lu, "Sparse coding with memristor networks," *Nature Nanotechnology*, vol. 12, pp. 784-789, 2017/08/01 2017.
- [13] M. Chu, B. Kim, S. Park, H. Hwang, M. Jeon, B. H. Lee, *et al.*, "Neuromorphic hardware system for visual pattern recognition with memristor array and CMOS neuron," *IEEE Transactions on Industrial Electronics*, vol. 62, pp. 2410-2419, 2014.
- [14] C. Li, M. Hu, Y. Li, H. Jiang, N. Ge, E. Montgomery, *et al.*, "Analogue signal and image processing with large memristor crossbars," *Nature Electronics*, vol. 1, p. 52, 2018.
- [15] Z. Yin, H. Li, H. Li, L. Jiang, Y. Shi, Y. Sun, *et al.*, "Single-layer MoS2 phototransistors," *ACS nano*, vol. 6, pp. 74-80, 2012.
- [16] L. Wang, Y. Wang, J. I. Wong, T. Palacios, J. Kong, and H. Y. Yang, "Functionalized MoS2 nanosheet-based field-effect biosensor for label-free sensitive detection of cancer marker proteins in solution," *Small*, vol. 10, pp. 1101-1105, 2014.
- [17] M. Park, Y. J. Park, X. Chen, Y. K. Park, M. S. Kim, and J. H. Ahn, "MoS2-based tactile sensor for electronic skin applications," *Advanced Materials*, vol. 28, pp. 2556-2562, 2016.
- [18] A. J. Arnold, T. Shi, I. Jovanovic, and S. Das, "Extraordinary Radiation Hardness of Atomically Thin MoS2," *ACS applied materials & interfaces*, vol. 11, pp. 8391-8399, 2019.
- [19] P.-C. Shen, C. Su, Y. Lin, A.-S. Chou, C.-C. Cheng, J.-H. Park, *et al.*, "Ultralow contact resistance between semimetal and monolayer semiconductors," *Nature*, vol. 593, pp. 211-217, 2021/05/01 2021.

- [20] Q. Smets, G. Arutchelvan, J. Jussot, D. Verreck, I. Asselberghs, A. N. Mehta, *et al.*, "Ultra-scaled MOCVD MoS₂ MOSFETs with 42nm contact pitch and 250 μ A/ μ m drain current," in *2019 IEEE International Electron Devices Meeting (IEDM)*, 2019, pp. 23.2. 1-23.2. 4.
- [21] A. Rai, A. Valsaraj, H. C. Movva, A. Roy, R. Ghosh, S. Sonde, *et al.*, "Air Stable Doping and Intrinsic Mobility Enhancement in Monolayer Molybdenum Disulfide by Amorphous Titanium Suboxide Encapsulation," *Nano Lett*, vol. 15, pp. 4329-36, Jul 8 2015.
- [22] C. D. English, K. K. H. Smithe, R. L. Xu, and E. Pop, "Approaching ballistic transport in monolayer MoS₂ transistors with self-aligned 10 nm top gates," in *2016 IEEE International Electron Devices Meeting (IEDM)*, 2016, pp. 5.6.1-5.6.4.
- [23] K. M. Price, K. E. Schauble, F. A. McGuire, D. B. Farmer, and A. D. Franklin, "Uniform Growth of Sub-5-Nanometer High- κ Dielectrics on MoS₂ Using Plasma-Enhanced Atomic Layer Deposition," *ACS applied materials & interfaces*, vol. 9, pp. 23072-23080, 2017.
- [24] 2DCC. *2d-crystal-consortium*. Available: <https://www.mri.psu.edu/2d-crystal-consortium/user-facilities/thin-films/list-thin-film-samples-available>
- [25] K. Kang, S. Xie, L. Huang, Y. Han, P. Y. Huang, K. F. Mak, *et al.*, "High-mobility three-atom-thick semiconducting films with wafer-scale homogeneity," *Nature*, vol. 520, p. 656, 2015.
- [26] S. Wachter, D. K. Polyushkin, O. Bethge, and T. Mueller, "A microprocessor based on a two-dimensional semiconductor," *Nature communications*, vol. 8, p. 14948, 2017.
- [27] D. K. Polyushkin, S. Wachter, L. Mennel, M. Paur, M. Paliy, G. Iannaccone, *et al.*, "Analogue two-dimensional semiconductor electronics," *Nature Electronics*, vol. 3, pp. 486-491, 2020/08/01 2020.
- [28] Q. Gao, Z. Zhang, X. Xu, J. Song, X. Li, and Y. Wu, "Scalable high performance radio frequency electronics based on large domain bilayer MoS₂," *Nature Communications*, vol. 9, p. 4778, 2018/11/14 2018.
- [29] S. Das, A. Dodda, and S. Das, "A biomimetic 2D transistor for audiomorphic computing," *Nature Communications*, vol. 10, p. 3450, 2019/08/01 2019.
- [30] A. Sebastian, A. Pannone, S. S. Radhakrishnan, and S. Das, "Gaussian synapses for probabilistic neural networks," *Nature communications*, vol. 10, pp. 1-11, 2019.
- [31] A. J. Arnold, A. Razavieh, J. R. Nasr, D. S. Schulman, C. M. Eichfeld, and S. Das, "Mimicking Neurotransmitter Release in Chemical Synapses via Hysteresis Engineering in MoS₂ Transistors," *ACS nano*, vol. 11, pp. 3110-3118, 2017.
- [32] S. J. Kim, K. Choi, B. Lee, Y. Kim, and B. H. Hong, "Materials for Flexible, Stretchable Electronics: Graphene and 2D Materials," *Annual Review of Materials Research*, vol. 45, pp. 63-84, 2015.
- [33] D.-H. Kim, J.-H. Ahn, W. M. Choi, H.-S. Kim, T.-H. Kim, J. Song, *et al.*, "Stretchable and foldable silicon integrated circuits," *Science*, vol. 320, pp. 507-511, 2008.
- [34] Y. M. Song, Y. Xie, V. Malyarchuk, J. Xiao, I. Jung, K.-J. Choi, *et al.*, "Digital cameras with designs inspired by the arthropod eye," *Nature*, vol. 497, pp. 95-99, 2013/05/01 2013.
- [35] A. Sebastian, R. Pendurthi, T. H. Choudhury, J. M. Redwing, and S. Das, "Benchmarking monolayer MoS₂ and WS₂ field-effect transistors," *Nature Communications*, vol. 12, p. 693, 2021/01/29 2021.
- [36] H. Li, Q. Zhang, C. C. R. Yap, B. K. Tay, T. H. T. Edwin, A. Olivier, *et al.*, "From Bulk to Monolayer MoS₂: Evolution of Raman Scattering," *Advanced Functional Materials*, vol. 22, pp. 1385-1390, Apr 10 2012.
- [37] S. Das, J. A. Robinson, M. Dubey, H. Terrones, and M. Terrones, "Beyond Graphene: Progress in Novel Two-Dimensional Materials and van der Waals Solids," *Annual Review of Materials Research*, vol. 45, pp. 1--27, 2015.

- [38] S. Tongay, J. Suh, C. Ataca, W. Fan, A. Luce, J. S. Kang, *et al.*, "Defects activated photoluminescence in two-dimensional semiconductors: interplay between bound, charged, and free excitons," *Scientific reports*, vol. 3, p. 2657, 2013.
- [39] J. R. Nasr, N. Simonson, A. Oberoi, M. W. Horn, J. A. Robinson, and S. Das, "Low-Power and Ultra-Thin MoS₂ Photodetectors on Glass," *ACS nano*, 2020.
- [40] A. Dodda, A. Oberoi, A. Sebastian, T. H. Choudhury, J. M. Redwing, and S. Das, "Stochastic resonance in MoS₂ photodetector," *Nature Communications*, vol. 11, p. 4406, 2020/09/02 2020.
- [41] D. Jayachandran, A. Oberoi, A. Sebastian, T. H. Choudhury, B. Shankar, J. M. Redwing, *et al.*, "A low-power biomimetic collision detector based on an in-memory molybdenum disulfide photodetector," *Nature Electronics*, pp. 1-10, 2020.
- [42] P. Han, L. S. Marie, Q. X. Wang, N. Quirk, A. El Fatimy, M. Ishigami, *et al.*, "Highly sensitive MoS₂ photodetectors with graphene contacts," *Nanotechnology*, vol. 29, p. 20LT01, 2018.
- [43] L. Mennel, J. Symonowicz, S. Wachter, D. K. Polyushkin, A. J. Molina-Mendoza, and T. Mueller, "Ultrafast machine vision with 2D material neural network image sensors," *Nature*, vol. 579, pp. 62-66, 2020.
- [44] C. Piochon, M. Kano, and C. Hansel, "LTD-like molecular pathways in developmental synaptic pruning," *Nature neuroscience*, vol. 19, p. 1299, 2016.
- [45] C. Hansel, "Deregulation of synaptic plasticity in autism," *Neuroscience letters*, vol. 688, pp. 58-61, 2019.
- [46] Y. Xuan, A. Jain, S. Zafar, R. Lotfi, N. Nayir, Y. Wang, *et al.*, "Multi-scale modeling of gas-phase reactions in metal-organic chemical vapor deposition growth of WSe₂," *Journal of Crystal Growth*, vol. 527, 2019.
- [47] D. Jayachandran, A. Oberoi, A. Sebastian, T. H. Choudhury, B. Shankar, J. M. Redwing, *et al.*, "A low-power biomimetic collision detector based on an in-memory molybdenum disulfide photodetector," *Nature Electronics*, vol. 3, pp. 646-655, 2020/10/01 2020.

AUTHOR INFORMATION

Corresponding Author

sud70@psu.edu, das.sapt@gmail.com

Author Contributions

S.D and S.S conceived the idea and designed the experiments. S.D, S.S, and A.D performed the experiments, analyzed the data, discussed the results, agreed on their implications. All authors contributed to the preparation of the manuscript.

Competing Interest

The authors declare no competing interests

Acknowledgement

The work was supported by Army Research Office (ARO) through Contract Number W911NF1920338. Authors also acknowledge Mr. Amritanand Sebastian and Mr. Thomas F Schranghamer for help with device fabrication. Authors also acknowledge the materials support from the National Science Foundation (NSF) through the Pennsylvania State University 2D Crystal Consortium–Materials Innovation Platform (2DCCMIP) under NSF cooperative agreement DMR-1539916.

Figure Captions

Figure 1. Monolithically integrated, multi-pixel, all-in-one biomimetic spiking neural network (BNN). a) Biological neural network (BNN) for processing visual information. Optical images of b) 7×7 -pixel BNN architecture, c) each pixel comprising of four monolayer MoS₂ FETs (4T cell) that monolithically integrate the sensing module (SM), encoding module (EM), and learning module (LM), and d) individual monolayer MoS₂ field effect transistor (FET), which is locally back-gated using a stack comprising of atomic layer deposition (ALD) grown 50 nm Al₂O₃ on sputter deposited 40/30 nm Pt/TiN. All back-gate islands were placed on a commercially purchased SiO₂/p⁺⁺-Si substrate. e) Circuit schematic for each pixel showing the connection between SM, EM, and LM consisting of 1 (T_{SM}), 2 (T_{EM1} and T_{EM2}), and 1 (T_{LM}) MoS₂ FETs, respectively. MoS₂ FETs used for the SMs and LMs have footprints ($W \times L$) of $5 \mu\text{m} \times 1 \mu\text{m}$ and MoS₂ FETs used for the EMs have a footprint of $5 \mu\text{m} \times 3 \mu\text{m}$ excluding the contact pads and each pixel has a footprint of $400 \mu\text{m} \times 600 \mu\text{m}$. Each pixel is designed to achieve functional and organizational resemblance with different neuronal cells found in the visual BNN. For example, SM is functionally equivalent to photoreceptor cells (rods and cones) in the human eyes that convert external optical stimuli into corresponding graded potentials (V_{N3}) at node N_3 . Rods primarily enable low-light or scotopic vision whereas cones are responsible for bright-light or photopic vision, both of which can be achieved using T_{SM} by exploiting gate-tunable persistent photoconductivity. Similarly, EM mimics the functionality of retinal ganglion cells that encode the graded potentials into spike trains and transmit to visual cortex for further processing. Finally, LM imitates the visual cortex where learning, forgetting, and inference take place. f) 3D schematic of each MoS₂ FET with non-volatile programming capability, which empowers our BNN

architecture to overcome the von Neumann bottleneck and enable in-memory sensing and computing.

Figure 2. Characterization and device-to-device variation of MoS₂ FETs. a) Raman and b) photoluminescence (PL) spectra of a representative MoS₂ channel region. The Raman peak separation between the characteristics A_{1g} and E^1_{2g} modes of 17 cm^{-1} , and PL peak location at 1.82 eV are consistent with monolayer MoS₂. c) Raman and d) PL maps of $10\text{ }\mu\text{m} \times 10\text{ }\mu\text{m}$ MoS₂ regions. Colormaps of distribution of e) Raman peak separation and f) PL peak position across 49 MoS₂ channels corresponding to each of the 7×7 pixels of our BNN architecture. The mean and standard deviation values were extracted to be 18 cm^{-1} and 0.8 cm^{-1} , respectively, for Raman peak separation and 1.82 eV and 0.01 eV , respectively, for PL peak location. g) Transfer characteristics, i.e., source to drain current (I_{DS}) as a function of the local back-gate voltage (V_{BG}) at different drain biases (V_{DS}) for a representative MoS₂ FET with $L = 1\text{ }\mu\text{m}$. h) Device-to-device variation in the transfer characteristics across 49 MoS₂ FETs corresponding to each of the 7×7 pixels. Colormaps of distribution of i) electron field effect mobility values (μ_{FE}) extracted from the peak transconductance, j) current on/off ratio ($r_{ON/OFF}$), k) subthreshold slope (SS) over 3 orders of magnitude change in I_{DS} , and l) threshold voltage (V_{TH}) extracted at iso-current of $100\text{ nA}/\mu\text{m}$ for these 49 MoS₂ FETs. Extracted mean values for μ_{FE} , $r_{ON/OFF}$, SS , and V_{TH} were found to be $21\text{ cm}^2\text{V}^{-1}\text{s}^{-1}$, 2.6×10^7 , 275 mV/decade , and 0.9 V , respectively, with corresponding standard deviation values of $5.5\text{ cm}^2\text{V}^{-1}\text{s}^{-1}$, 0.8×10^7 , 59 mV/decade , and 0.2 V , respectively.

Figure 3. MoS₂ FET based neuromorphic sensing module (SM): a) Transfer characteristics of monolayer MoS₂ FET measured with $V_{DS} = 1\text{ V}$ before and after illumination from a blue LED

with input currents ranging from $I_{\text{LED}} = 0.5 \text{ mA}$ (low-brightness) to $I_{\text{LED}} = 20 \text{ mA}$ (high-brightness) at different $V_{\text{BG}} = V_{\text{write}}$ for $t_{\text{write}} = 100 \text{ ms}$. b) Analog valued and continuous time input optical stimuli from the blue LED. c) Corresponding temporal evolution of the graded potential, V_{N3} , at node N_3 for different V_{write} obtained by using the circuit layout for SM shown in Fig. 1d. A constant voltage, $V_{\text{N1}} = 5 \text{ V}$, is applied to the node N_1 , which is the drain terminal of T_{SM} and a clocking signal toggling between V_{read} and V_{write} is applied to node N_2 , which is the local back-gate of T_{SM} with $\tau_{\text{CLK}} = 100 \text{ ms}$. The source terminal of T_{SM} is connected to the local back-gate of T_{EM1} at node, N_3 . d) Time for V_{N3} to reach the same magnitude (τ_{SAT}) as a function of I_{LED} and V_{write} . e) Average energy consumption by the SM (E_{SM}) during each τ_{CLK} for different I_{LED} and V_{write} . f) Device-to-device variation in photoresponse of 49 MoS₂ FETs corresponding to the SMs of each of the 7×7 pixels of our BNN hardware after $t_{\text{write}} = 1 \text{ second}$ exposure to $I_{\text{LED}} = 20 \text{ mA}$ at $V_{\text{write}} = -2.5 \text{ V}$. g) Colormap of distribution of ratio of post-illumination photoconductance to dark conductance (r_{PH}) measured at $V_{\text{BG}} = 0 \text{ V}$. The mean and standard deviation values were found to be 6.7×10^3 and 3.8×10^3 , respectively.

Figure 4. MoS₂ FET-based neuromorphic encoding module (EM). a) Transfer characteristics of MoS₂ FETs used as T_{EM1} and T_{EM2} in the EM. T_{EM1} is programmed to operate as a depletion mode (normally on) n-channel FET, whereas T_{EM2} operates as an enhancement mode (normally off) n-channel FET. Based on the circuit layout shown in Fig. 1d, the EM module serves as an NMOS inverter with depletion load. b) Input (V_{N3}) versus output (V_{N5}) characteristics of the EM for different V_{p} values applied to the source terminal of T_{EM2} , i.e., N_6 . The drain terminal of T_{EM1} , i.e., N_4 is kept grounded. c) Various programming states of T_{EM2} and d) corresponding EM characteristics for $V_{\text{p}} = -5 \text{ V}$. The inverting threshold (V_{IT}), i.e., the magnitude of V_{N3} at which V_{N5}

reaches $V_p/2$ can be adjusted by reconfiguring T_{EM1} and T_{EM2} . e) Spike-duration and f) spike-count based encoding of graded potential (V_{N3}) received from SM corresponding to different V_{write} and I_{LED} into programming voltage (V_{N5}) for transmitting to the learning module (LM). g) Total spiking time (τ_{spike}) and h) corresponding average encoding energy expenditure (E_{EM}) per clock cycle for spike-duration based encoding and i) total number of spikes (N_{spike}) and h) corresponding E_{EM} for spike-count based encoding as a function of V_{write} and I_{LED} . The input stimulus is presented for a duration of 10 s and $V_p = -6$ V. Spike-duration and spike numbers are counted once V_{N5} reaches 75% of V_p , i.e., -4.5 V.

Figure 5. Analog and non-volatile programming of MoS₂ FET-based synapses. a) Potentiation of a MoS₂ synapse from low conductance state (LCS) after the application of a fixed number of programming spikes ($N_{spike} = 10$) of different amplitudes of negative polarity (V_p) with each spike applied for $t_{spike} = 100$ ms. b) Post-potentiated conductance states (G_p) measured at $V_{BG} = 0$ V as a function of N_{spike} for different V_p . c) Depression of a MoS₂ synapse from high conductance state (HCS) after the application of a fixed number of programming spikes ($N_{spike} = 10$) of different amplitudes of positive polarity (V_D) with each spike applied for $t_{spike} = 100$ ms. d) Post-depressed conductance states (G_D) measured at $V_{BG} = 0$ V as a function of N_{spike} for different V_D . e) Potentiation of MoS₂ synapse from LCS after the application of single spike of constant magnitude $V_p = -6$ V for different t_{spike} . f) Post-potentiated G_p measured at $V_{BG} = 0$ V as a function of t_{spike} for different V_p . g) Depression of MoS₂ synapse from HCS after the application of single spike of constant magnitude $V_D = 6$ V for different t_{spike} . h) Post-depressed G_D measured at $V_{BG} = 0$ V as a function of t_{spike} for different V_D . Retention characteristics for i) 6 potentiated and j) 6 depressed conductance states for 100 seconds. k) Device-to-device variation in the pre- and post-programmed

transfer characteristics and 1) corresponding colormap of distribution of memory ratio (MR) measured at $V_{BG} = 0$ V for 49 monolayer MoS₂ FETs from each LM of our 7×7 BNN platform when programmed using $N_{spike} = 10$ with spike magnitude, $V_P = -8$ V, and spike width, $t_{spike} = 100$ ms. The mean and standard deviation values for MR were found to be 6×10^5 and standard deviations of 0.5×10^5 , respectively.

Figure 6. MoS₂ FET-based neuromorphic learning module (LM). a) Spike-duration and b) spike-count based conductance evolution in MoS₂ FET based LM when input programming waveforms (V_{N5}) are received from the EM corresponding to different I_{LED} and V_{write} . Final conductance state achieved by the LM for c) spike-duration and d) spike-count based input spiking patterns received from the EM module corresponding to different I_{LED} and V_{write} . The average learning energy expenditure (E_{LM}) per clock cycle by the LM for e) spike-duration and f) spike-count based learning for different I_{LED} and V_{write} .

Figure 7. Multi-pixel demonstration of analog image sensing, encoding, and leaning. a) Analog 7×7 input pattern obtained by illuminating the blue LED. Temporal evolution of corresponding b) graded potential (V_{N3}) at the output of SMs, c) programming spike-count (N_{spike}) at the output of the EMs, and d) programmed conductance values at the output of LMs. Clearly, the input LED pattern is learnt by our 7×7 BNN hardware. For this demonstration, all MoS₂ FET based synapses belonging to the LMs were initially programmed in their LCS (100 pS) and different LED illuminations were presented one by one to the corresponding pixels of our 7×7 BNN hardware.

Figure 8. Importance of forgetting (synaptic depression) in learning and inference. a) Schematic and optical image of a 2-layer BNN with 9 presynaptic neurons and 1 postsynaptic neuron for learning and inferring patterns from 3×3 pixelated images. b) Training and retraining schedule with $M = 40$ epochs, with each epoch having potentiation and depression cycles. During the potentiation, the pattern to be learned is presented to the BNN, whereas during the depression all synapses are uniformly depressed. Spiking profiles used for c) spike-count and d) spike-duration based learning. For each type of learning, three BNN configurations are used: 1) weak potentiation and strong depression, 2) strong potentiation and weak depression, and 3) strong potentiation and strong depression. The strength of potentiation (V_P) and depression (V_D) are adjusted using the spike magnitude and spike duration for spike-count and spike-duration based learnings, respectively. The time evolution of colormap of synaptic weights i.e., the conductance states of the 9 synapses during e) spike-count and f) spike-duration based learning. For each type of learning all synapses are initialized either in a high conductance state (HCS) with $G_{HCS} = 100$ nS, or a low conductance state (LCS) with $G_{LCS} = 100$ pS (also see the *Supplementary Video 3 and 4*). Learning of the left diagonal followed by relearning of the right diagonal when potentiation and depression are both strong for g) spike-count and h) spike-duration based learnings (also see the *Supplementary Video 5 and 6*).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryVideo1.mov](#)
- [SupplementaryVideo2.mov](#)
- [SupplementaryVideo3.mp4](#)
- [SupplementaryVideo4.mp4](#)
- [SupplementaryVideo5.mp4](#)
- [SupplementaryVideo6.mp4](#)
- [SupplementaryInformation.pdf](#)