REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggesstions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA, 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any oenalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
10-04-2023	Final Report	30-Sep-2019 - 31-Dec-2022

4. TITLE AND SUBTITLE
Final Report: Gaussian Synapses for Probabilistic Neural Network, Thrust Area: Novel Methods of Computing

5a. CONTRACT NUMBER
W911NF-19-2-0338

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
Pennsylvania State University
Office of Sponsored Programs
110 Technology Center Building
University Park, PA 16802 -7000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
75870-NC-ACI.7

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Saptarshi Das
UU	UU	UU	UU		19b. TELEPHONE NUMBER 814-863-2639

RPPR Final Report
as of 10-Apr-2023

Agency Code: 21XD

Proposal Number: 75870NCACI **Agreement Number: W911NF-19-2-0338**
INVESTIGATOR(S):

Name: Saptarshi Das
Email: sud70@psu.edu
Phone Number: 8148632639
Principal: Y

Organization: **Pennsylvania State University**
Address: Office of Sponsored Programs, University Park, PA 168027000
Country: USA
DUNS Number: 003403953 EIN: 246000376
Report Date: 31-Mar-2023 Date Received: 10-Apr-2023
Final Report for Period Beginning 30-Sep-2019 and Ending 31-Dec-2022
Title: Gaussian Synapses for Probabilistic Neural Network, Thrust Area: Novel Methods of Computing
Begin Performance Period: 30-Sep-2019 **End Performance Period:** 31-Dec-2022
Report Term: 0-Other
Submitted By: Saptarshi Das Email: sud70@psu.edu
Phone: (814) 863-2639
Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: 2 **STEM Participants:**

Major Goals: The recent decline in energy, size, and complexity scaling of Si complementary metal oxide semiconductor (CMOS) technology and the associated von Neumann architecture poses a significant challenge towards achieving Exascale Computing infrastructure, which is essential for defense and national security, healthcare, education, and agriculture. Recent progress in neuromorphic computing based on artificial neural networks (ANNs) that mimic the fundamental computing unit of the brain i.e. neurons connected to other neurons via synapse has indeed enabled the complexity
scaling but has failed to address the energy and size scaling aspect. This explains why even the most advanced neuromorphic chips such as the True North from IBM will require almost 1000000X enhancement in energy efficiency when scaled up to the complexity of the human brain. Furthermore, fundamental challenges associated with ANNs such as slow learning, incremental adaption, and false convergence can ultimately limit the complexity scaling as well. In this project, we propose to overcome the aforementioned limitations of the state-of-the-art neuromorphic chips by introducing an aggressively scalable, extremely energy efficient, and in-memory analog device,
namely, the Gaussian synapse, based on two-dimensional (2D) materials that enable hardware implementation of probabilistic machine learning to solve a diverse set of discrete and continuous domain computational problems. In particular, atomically thin 2D materials facilitate size scaling, analog Gaussian synapses offer aggressive energy scaling by utilizing subthreshold thermionic or steep subthreshold tunneling transport to go beyond Boltzmann tyranny, and finally in-memory implementation of various probabilistic and non-probabilistic machine learning models advances complexity scaling. We will demonstrate how the area and energy efficiency at the device level provide cascading benefits for computing at the circuit, architecture, and system levels. In short, the successful execution of our proposed effort will ensure that the growth in computing power remains on an exponential trajectory, while the energy, area, and dollar cost continues to diminish.

Accomplishments: 1. Demonstration of MoTe2-based reconfigurable Gaussian synapse
2. Graphene Memristors: We have demonstrated graphene-based multi-level (>16) and non-volatile memristive synapses with arbitrarily programmable conductance states with desirable retention and programming endurance. We also demonstrated that graphene memristors enable weight assignment based on k-means clustering, which offers greater computing accuracy when compared with uniform weight quantization for vector-matrix multiplication, an essential component for any neural network.
3. Spike-based Computing: The ultimate energy benefits of Gaussian synapses can be achieved through delta synapses or spike-based probabilistic computing. We have been able to demonstrate a biomimetic device based on dual gated field effect transistor (FET) capable of encoding analog signals into stochastic spike trains following various neural encoding algorithms such as rate-based coding, spike timing-based coding, and spike count-based

coding. Two important aspects of neural encoding, namely, dynamic range and encoding precision are also captured in our demonstration. Furthermore, the spike encoding energy was found to be as frugal as ~200 pJ/spike. Finally, we show fast (~200 timesteps) encoding of the MNIST data set using our biomimetic device followed by more than 91% accurate inference using a trained SNN.

4. We have been able to demonstrate hardware realization of Gaussian synapses using n-type MoS2 and p-type V-doped WSe2 FETs. We have integrated memory with compute to demonstrate in-memory Gaussian synapses with reconfigurable means and standard deviations. We have simulated Kernel density estimation (KDE) using Gaussian kernels for regression analysis.
5. We have developed in-memory computing primitives for hardware acceleration of simulated annealing (SA). We exploited subthreshold conduction in complementary two-dimensional (2D) field effect transistors (p-type WSe2 and n-type MoS2) integrated with analog, non-volatile, and programmable floating-gate memory stack to develop in-memory computing primitives and annealing schedules necessary for energy and area efficient hardware acceleration of SA for the Ising spin systems. We experimentally demonstrated > 800X search acceleration for 4×4 ferromagnetic, antiferromagnetic, and a spin glass system using SA compared to an exhaustive search at a minuscule total energy expenditure of ~120 nJ.
6. We have also demonstrated an “all-in-one” biomimetic hardware spiking neural network (SNN) which is capable of sensing, encoding, and spike-based unsupervised learning and inference using monolayer MoS2-based multifunctional optoelectronic and programmable memory devices. We use photogating effect in the MoS2 phototransistor to directly sense and encode optical information into graded spike trains. We developed MoS2-based encoding cells to implement spike-count and spike-timing-based encoding algorithms. And finally, we use MoS2-based electrically programmable non-volatile synapses for spike-based unsupervised learning and inference. Furthermore, we demonstrate low-power operation and adaptability of our SNN to learning under different ambient conditions mimicking the neuroplasticity of biological neural network (BNN). Our SNN hardware also offers a platform to model learning disabilities and disorders of BNN at a high level of abstraction
7. We completed our proposed task of the hardware realization of Gaussian synapses using n-type MoS2 and p-type V-doped WSe2 FETs through heterogeneous integration. We also completed the integration of logic and memory to demonstrate in-memory Gaussian synapses with reconfigurable means and standard deviations. In addition, we also demonstrated logic circuits such as inverters and multiplexers, and neuromorphic computing primitives such as sigmoid and hyperbolic tangent activation functions using the non–von Neumann 2D CMOS platform. This work was published in Small 18 (33), 2202590, 2022.
8. We also completed our proposed task of a hybrid hardware/software co-acceleration of simulated annealing (SA). We exploited the subthreshold conduction and in-memory compute capability of the non–von Neumann 2D CMOS platform to develop hardware primitives and annealing schedules necessary for energy and area-efficient hardware acceleration of SA for the Ising spin systems. We experimentally demonstrated > 800X search acceleration for 4×4 ferromagnetic, antiferromagnetic, and a spin glass system using SA compared to an exhaustive search at a minuscule total energy expenditure of ~120 nJ. This work was published as Editor's Choice in Advanced Materials 34 (4), 2107076, 2022.
9. We also completed the hardware realization of a sparse spike-timing-based encoding of visual information for the acceleration of emerging spiking neural networks (SNNs). Here we exploited the in-sensor compute capability of our photosensitive 2D memtransistors based on monolayer MoS2. Note that the encoding of information into sparse spike trains constitutes the basis of energy-efficient neural computation in animal brains. Specifically, we have designed and experimentally demonstrated a medium-scale integrated circuit composed of two cascaded three-stage inverters and one XOR logic gate fabricated using a total of twenty-one (21) photosensitive 2D memtransistors for spike-timing-based encoding of visual information at a minuscule energy expenditure of few microjoules. This work has been accepted for publication in Advanced Materials 2202535, 2022.
10. We also completed a hardware demonstration of near-sensor compute using our photosensitive 2D memtransistors based on monolayer MoS2 for the acceleration of a bio-inspired encryption algorithm. Our hardware platform consisted of a total of 320 fully integrated 2D memtransistors and consumed energy in the range of hundreds of picojoules. This work was published in Nature Communications 13, 3587, 2022.
11. We also completed hardware acceleration of Bayesian networks (BNs) by exploiting programming stochasticity in 2D memtransistor based on monolayer MoS2. While Si-based CMOS technology has been used for hardware implementation of BNs, the lack of inherent stochasticity makes it area and energy inefficient. We circumvented these challenges by demonstrating a low-power and compact s-bit (stochastic bit, where the probability of obtaining ‘1’ in a bit stream can be reconfigured) generator circuit that exploits cycle-to-cycle fluctuation in the post-programmed conductance state of 2D memtransistors. Next, the s-bit generators were monolithically integrated with 2D memtransistor-based logic gates to implement BNs. This work was published in Nature Communications 13, 5578, 2022.

RPPR Final Report
as of 10-Apr-2023

Training Opportunities: The following graduate students were trained in device fabrication, material characterization, electrical measurements, analytical modeling, and numerical simulations. They also made presentations in different forums and wrote several manuscripts.

1. Akhil Dodda (Graduated with a Ph.D. degree and currently working at Western Digital)
2. Amritanand Sebastian (Graduated with a Ph.D. degree and currently working at Intel)
3. Shiv Subbulakshmi Radhakrishnan
4. Darsith Jayachandran
5. Aaryan Oberoi
6. Thomas Schranghamer
7. Rahul Pendurthi
8. Andrew Pannone
9. Harikrishnan Ravichandran
10. Yikai Zheng

Results Dissemination: 22. Ravichandran, H., Zheng, Y., Schranghamer, T. F., Trainor, N., Redwing, J. M., & Das, S. A Monolithic Stochastic Computing Architecture for Energy Efficient Arithmetic. Advanced Materials, 2206168.

21. Jayachandran, D., Pannone, A., Das, M., Schranghamer, T. F., Sen, D., & Das, S. Insect-Inspired, Spike-Based, in-Sensor, and Night-Time Collision Detector Based on Atomically Thin and Light-Sensitive Memtransistors. ACS Nano. .

20. Dodda, A., Jayachandran, D., Pannone, A., Trainor, N., Stepanoff, S. P., Steves, M. A., Subbulakhshmi Radhakrishnan, S., Bachu, S., Ordonez, C. W., Shallenberger, J. R., Redwing, J. M., Knappenberger, K., Wolfe, D. E., & Das, S. Active pixel sensor matrix based on monolayer MoS2 phototransistor array. Nature Materials, 21, 1379–1387. (Cover Article)

19. Dodda, A., Jayachandran, D., Pannone, A., Subbulakhshmi Radhakrishnan, S., Trainor, N., Redwing, J. M., & Das, S. A Bio-inspired and Low-power 2D Machine Vision with Adaptive Machine Learning and Forgetting. ACS Nano, 16, 12, 20010–20020, (2022).

18. Subbulakshmi Radhakrishnan, S., Dodda, A., & Das, S. An All-in-One Bio-inspired Neural Network. ACS Nano, 16, 12, 20100–20115, (2022).

17. Chakrabarti, S., Wali, A., Ravichandran, H., Kundu, S., Schranghamer, T. F., Basu, K., & Das, S. Logic Locking of Integrated Circuits Enabled by Nanoscale MoS2-Based Memtransistors. ACS Applied Nano Materials, 5, 10, 14447-14455.

16. Sebastian, A., Pendurthi, R., Kozhakhmetov, A., Trainor, N., Robinson, J. A., Redwing, J. M., & Das, S. Two-dimensional materials-based probabilistic synapses and reconfigurable neurons for measuring inference uncertainty using Bayesian neural networks. Nature Communications, 13, 6139, (2022).

15. Zheng, Y., Ravichandran, H., Schranghamer, T. F., Trainor, N., Redwing, J. M., & Das, S. Hardware implementation of Bayesian network based on two-dimensional memtransistors. Nature communications, 13, 5578, (2022).

14. Das, S., & Das, S. Digital Keying Enabled by Reconfigurable Two?dimensional Modulators. Advanced Materials, 2203753, (2022).

13. Pendurthi, R., Jayachandran, D., Kozhakhmetov, A., Trainor, N., Robinson, J. A., Redwing, J. M., & Das, S. Heterogeneous integration of atomically thin semiconductors for non?von Neumann CMOS. Small, 18(33), 2202590, (2022).

12. Dodda, A., Trainor, N., Redwing, J., & Das, S. All-in-one, bio-inspired, and low-power crypto engines for near-sensor security based on two-dimensional memtransistors. Nature Communications, 13, 3587, (2022).

11. Subbulakshmi Radhakrishnan, S., Chakrabarti, S., Sen, D., Das, M., Schranghamer, T., Sebastian, A., & Das. S. A Sparse and Spike-timing-based Adaptive Photo Encoder for Augmenting Machine Vision for Spiking Neural Networks. Advanced Materials, 34, 48, 2202535, (2022).

10. Sebastian, A., Das, S. & Das, S. An Annealing Accelerator for Ising Spin Systems Based on In-Memory Complementary 2D FETs. Advanced Materials, 34 (4), 2107076, (2022).

9. Das, S, Sebastian, A., Pop, E., McClellan, C. J., Franklin, A. D., Grasser, T., Knobloch, T., Illarionov, Y., Penumatcha, A. V., Appenzeller, J., Chen, Z., Zhu, W., Asselberghs, I., Li, L., Avci, U. E., Bhat, N., Anthopoulos, T. D., & Singh, R. Transistors based on two-dimensional materials for future integrated circuits. Nature Electronics, 4 (11), 786-799, (2021).

8. Dodda, A., & Das, S. Demonstration of Stochastic Resonance, Population Coding, and Population Voting Using Artificial MoS2 Based Synapses. ACS Nano, 15 (10), 16172–16182, (2021).

7. Kozhakhmetov,A., Stolz, S., Tan, A. M. Z., Pendurthi, R., Bachu, S., Turker, F., Alem, N., Kachian, J., Das, S.,

Hennig, R. G., Gröning, O., Schuler, B., & Robinson, J. A. Controllable p?Type Doping of 2D WSe2 via Vanadium Substitution. Advanced Functional Materials, 31 (42), 2105252, (2021).

6. Schranghamer, T. F., Sharma, M., Singh, R., & Das, S. Review and comparison of layer transfer methods for two-dimensional materials for emerging applications. Chemical Society Reviews, 50, 11032-11054, (2021).

5. Subbulakshmi Radhakrishnan, S., Sebastian, Oberoi, A., Das, S., & Das, S. A biomimetic neural encoder for spiking neural network. Nature Communications, 12 (1), 2143, (2021).

4 . A., Sebastian, Pendurthi, R., A., Choudhury, T. H., Redwing, J. M., & Das, S. Benchmarking monolayer MoS2 and WS2 field-effect transistors. Nature Communications, 12 (1), 693, (2021).

3. Schranghamer, T. F., Oberoi, A., & Das, S. Graphene memristive synapses for high precision neuromorphic computing. Nature Communications, 11 (1), 5474, (2020).

2. Jayachandran, D., Oberoi, A., Sebastian, A., Choudhury, T. H., Shankar, B., Redwing, J. M., & Das, S. A Low Power Biomimetic Collision Detector Based on In-memory MoS2 Photodetector. Nature Electronics, 3 (10), 646-655, (2020).

1. Dodda, A., Oberoi, A., Sebastian, A., Choudhury, T. H., Redwing, J. M., & Das, S. Stochastic resonance in MoS2 photodetector. Nature Communications, 11 (1), 4406 (2020).

Honors and Awards: Nothing to Report

Protocol Activity Status:

Technology Transfer: Nothing to Report

PARTICIPANTS:

Participant Type: Graduate Student (research assistant)
Participant: Akhil Dodda
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Amritanand Sebastian
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Harikrishnan Ravichandran
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Andrew Pannone
Person Months Worked: 12.00 **Funding Support:**

Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Yikai Zheng
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Shiva Subbulakshmi Radhakrishnan
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Darsith Jayachandran
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Aaryan Oberoi
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Thomas Schranghamer
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Rahul Pendurthi
Person Months Worked: 12.00 **Funding Support:**
Project Contribution:
National Academy Member: N

ARTICLES:

RPPR Final Report
as of 10-Apr-2023

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: Chemical Society Reviews
Publication Identifier Type: DOI Publication Identifier: 10.1039/D1CS00706H
Volume: Issue: First Page #:
Date Submitted: 9/5/21 12:00AM Date Published: 8/16/21 4:00AM
Publication Location:
Article Title: Review and comparison of layer transfer methods for two-dimensional materials for emerging applications
Authors: Thomas F. Schranghamer, Madan Sharma, Rajendra Singh, Saptarshi Das
Keywords: Two-dimensional materials, layer-transfer
Abstract: Here we benchmark device-to-device variation in field-effect transistors (FETs) based on monolayer MoS2 and WS2 films grown using metal-organic chemical vapor deposition process. Our study involves 230 MoS2 FETs and 160 WS2 FETs with channel lengths ranging from 5??m down to 100?nm. We use statistical measures to evaluate key FET performance indicators for benchmarking these two-dimensional (2D) transition metal dichalcogenide (TMD) monolayers against existing literature as well as ultra-thin body Si FETs. Our results show consistent performance of 2D FETs across 1?×?1?cm2 chips owing to high quality and uniform growth of these TMDs followed by clean transfer onto device substrates. We are able to demonstrate record high carrier mobility of 33?cm2?V?1?s?1 in WS2 FETs, which is a 1.5X improvement compared to the best reported in the literature. Our experimental demonstrations confirm the technological viability of 2D FETs in future integrated circuits.
Distribution Statement: 3-Distribution authorized to U.S. Government Agencies and their contractors
Acknowledged Federal Support: **Y**

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: Nature Communications
Publication Identifier Type: DOI Publication Identifier: 10.1038/s41467-021-22332-8
Volume: 12 Issue: 1 First Page #:
Date Submitted: 9/5/21 12:00AM Date Published: 4/1/21 4:00AM
Publication Location:
Article Title: A biomimetic neural encoder for spiking neural network
Authors: Shiva Subbulakshmi Radhakrishnan, Amritanand Sebastian, Aaryan Oberoi, Sarbashis Das, Saptarshi D
Keywords: Two-dimensional materials, neural encoder
Abstract: Spiking neural networks (SNNs) promise to bridge the gap between artificial neural networks (ANNs) and biological neural networks (BNNs) by exploiting biologically plausible neurons that offer faster inference, lower energy expenditure, and event-driven information processing capabilities. However, implementation of SNNs in future neuromorphic hardware requires hardware encoders analogous to the sensory neurons, which convert external/internal stimulus into spike trains based on specific neural algorithm along with inherent stochasticity. Unfortunately, conventional solid-state transducers are inadequate for this purpose necessitating the development of neural encoders to serve the growing need of neuromorphic computing. Here, we demonstrate a biomimetic device based on a dual gated MoS2 field effect transistor (FET) capable of encoding analog signals into stochastic spike trains following various neural encoding algorithms such as rate-based encoding, spike timing-based encoding, and
Distribution Statement: 1-Approved for public release; distribution is unlimited.
Acknowledged Federal Support: **Y**

RPPR Final Report
as of 10-Apr-2023

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: Nature Communications
Publication Identifier Type: DOI Publication Identifier: 10.1038/s41467-020-20732-w
Volume: 12 Issue: 1 First Page #:
Date Submitted: 9/5/21 12:00AM Date Published: 1/1/21 5:00AM
Publication Location:
Article Title: Benchmarking monolayer MoS2 and WS2 field-effect transistors
Authors: Amritanand Sebastian, Rahul Pendurthi, Tanushree H. Choudhury, Joan M. Redwing, Saptarshi Das
Keywords: MoS2, WS2, FETs, Benchamrking
Abstract: Here we benchmark device-to-device variation in field-effect transistors (FETs) based on monolayer MoS2 and WS2 films grown using metal-organic chemical vapor deposition process. Our study involves 230 MoS2 FETs and 160 WS2 FETs with channel lengths ranging from 5??m down to 100?nm. We use statistical measures to evaluate key FET performance indicators for benchmarking these two-dimensional (2D) transition metal dichalcogenide (TMD) monolayers against existing literature as well as ultra-thin body Si FETs. Our results show consistent performance of 2D FETs across 1?×?1?cm2 chips owing to high quality and uniform growth of these TMDs followed by clean transfer onto device substrates. We are able to demonstrate record high carrier mobility of 33?cm2?V?1?s?1 in WS2 FETs, which is a 1.5X improvement compared to the best reported in the literature. Our experimental demonstrations confirm the technological viability of 2D FETs in future integrated circuits.
Distribution Statement: 1-Approved for public release; distribution is unlimited.
Acknowledged Federal Support: **Y**

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: Nature Communications
Publication Identifier Type: DOI Publication Identifier: 10.1038/s41467-020-19203-z
Volume: 11 Issue: 1 First Page #:
Date Submitted: 9/5/21 12:00AM Date Published: 10/1/20 4:00AM
Publication Location:
Article Title: Graphene memristive synapses for high precision neuromorphic computing
Authors: Thomas F. Schranghamer, Aaryan Oberoi, Saptarshi Das
Keywords: Memristive synapse, graphene
Abstract: Memristive crossbar architectures are evolving as powerful in-memory computing engines for artificial neural networks. However, the limited number of non-volatile conductance states offered by state-of-the-art memristors is a concern for their hardware implementation since trained weights must be rounded to the nearest conductance states, introducing error which can significantly limit inference accuracy. Moreover, the incapability of precise weight updates can lead to convergence problems and slowdown of on-chip training. In this article, we circumvent these challenges by introducing graphene-based multi-level (>16) and non-volatile memristive synapses with arbitrarily programmable conductance states. We also show desirable retention and programming endurance. Finally, we demonstrate that graphene memristors enable weight assignment based on k-means clustering, which offers greater computing accuracy when compared with uniform weight quantization for vector matrix multiplication.
Distribution Statement: 1-Approved for public release; distribution is unlimited.
Acknowledged Federal Support: **Y**

RPPR Final Report
as of 10-Apr-2023

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: Nature Electronics
Publication Identifier Type: DOI Publication Identifier: 10.1038/s41928-020-00466-9
Volume: 3 Issue: 10 First Page #: 646
Date Submitted: 9/5/21 12:00AM Date Published: 8/1/20 4:00AM
Publication Location:
Article Title: A low-power biomimetic collision detector based on an in-memory molybdenum disulfide photodetector
Authors: Darsith Jayachandran, Aaryan Oberoi, Amritanand Sebastian, Tanushree H. Choudhury, Balakrishnan S
Keywords: In-memory compute, 2D materials, Collision detector
Abstract: Accurately detecting a potential collision and triggering a timely escape response is critical in the field of robotics and autonomous vehicle safety. The lobula giant movement detector (LGMD) neuron in locusts can detect an approaching object and prevent collisions within a swarm of millions of locusts. This single neuronal cell performs nonlinear mathematical operations on visual stimuli to elicit an escape response with minimal energy expenditure. Collision avoidance models based on image processing algorithms have been implemented using analogue very-large-scale-integration designs, but none is as efficient as this neuron in terms of energy consumption or size. Here we report a nanoscale collision detector that mimics the escape response of the LGMD neuron. The detector comprises a monolayer molybdenum disulfide photodetector stacked on top of a non-volatile and programmable floating-gate memory architecture. It consumes a small amount of energy (in the range of nanojoules).
Distribution Statement: 1-Approved for public release; distribution is unlimited.
Acknowledged Federal Support: **Y**

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: Nature Communications
Publication Identifier Type: DOI Publication Identifier: 10.1038/s41467-020-18195-0
Volume: 11 Issue: 1 First Page #:
Date Submitted: 9/5/21 12:00AM Date Published: 9/1/20 4:00AM
Publication Location:
Article Title: Stochastic resonance in MoS2 photodetector
Authors: Akhil Dodda, Aaryan Oberoi, Amritanand Sebastian, Tanushree H. Choudhury, Joan M. Redwing, Sapta
Keywords: stochastic resonance, photodetector
Abstract: In this article, we adopt a radical approach for next generation ultra-low-power sensor design by embracing the evolutionary success of animals with extraordinary sensory information processing capabilities that allow them to survive in extreme and resource constrained environments. Stochastic resonance (SR) is one of those astounding phenomena, where noise, which is considered detrimental for electronic circuits and communication systems, plays a constructive role in the detection of weak signals. Here, we show SR in a photodetector based on monolayer MoS2 for detecting ultra-low-intensity subthreshold optical signals from a distant light emitting diode (LED). We demonstrate that weak periodic LED signals, which are otherwise undetectable, can be detected by a MoS2 photodetector in the presence of a finite and optimum amount of white Gaussian noise at a frugal energy expenditure of few tens of nano-Joules. The concept of SR is generic in nature and can be extended beyond photodetector
Distribution Statement: 1-Approved for public release; distribution is unlimited.
Acknowledged Federal Support: **Y**

RPPR Final Report
as of 10-Apr-2023

Partners

,

I certify that the information in the report is complete and accurate:
Signature: Saptarshi Das
Signature Date: 4/10/23 3:18PM

ADVANCED MATERIALS
www.advmat.de

A Monolithic Stochastic Computing Architecture for Energy Efficient Arithmetic

*Harikrishnan Ravichandran, Yikai Zheng, Thomas F Schranghamer, Nicholas Trainor, Joan M. Redwing, and Saptarshi Das**

As the energy and hardware investments necessary for conventional high-precision digital computing continue to explode in the era of artificial intelligence (AI), a change in paradigm that can trade precision for energy and resource efficiency is being sought for many computing applications. Stochastic computing (SC) is an attractive alternative since, unlike digital computers, which require many logic gates and a high transistor volume to perform basic arithmetic operations such as addition, subtraction, multiplication, sorting, etc., SC can implement the same using simple logic gates. While it is possible to accelerate SC using traditional silicon complementary metal–oxide–semiconductor (CMOS) technology, the need for extensive hardware investment to generate stochastic bits (s-bits), the fundamental computing primitive for SC, makes it less attractive. Memristor and spin-based devices offer natural randomness but depend on hybrid designs involving CMOS peripherals for accelerating SC, which increases area and energy burden. Here, the limitations of existing and emerging technologies are overcome, and a standalone SC architecture embedded in memory and based on 2D memtransistors is experimentally demonstrated. The monolithic and non-von-Neumann SC architecture occupies a small hardware footprint and consumes a miniscule amount of energy (<1 nJ) for both s-bit generation and arithmetic operations, highlighting the benefits of SC.

1. Introduction

The aggressive downscaling of feature sizes in silicon-based complementary metal–oxide–semiconductor (CMOS) technology over the past five decades has led to an exponential growth in the computing power of modern-day computers. Today, computers can fly jets, control industrial processes, and solve optimization problems. In fact, computers can also beat professional players in the game of "Go" and predict complex structures of proteins thanks to remarkable progress in the field of artificial intelligence (AI). The ongoing revolution in AI is directly linked to the unfathomable data processing power by computers enabling the implementation of deep learning and various other sophisticated machine learning algorithms.[1,2] However, there is significant infrastructure cost associated with advanced AI and computing systems. For example, any mathematical algorithm implemented using hardware requires arithmetic operations such as addition, subtraction, multiplication, sorting, etc., which are executed using logic circuits consisting of hundreds of transistors that occupy large areas and consume significant amounts of energy. Furthermore, the conventional von Neumann computing architecture necessitates frequent data shuttling between the arithmetic and the memory units to run algorithms, adding further area and energy overheads. Needless to say, these challenges are aggravated as the data size grows exponentially for both AI and non-AI platforms.[3,4] Therefore, a new paradigm that can drastically reduce the area and energy cost of arithmetic operations can not only benefit cloud computing using supercomputers but also enable edge computing in resource-constrained internet of things (IoT) devices.

Stochastic computing (SC) is an attractive alternative, where arithmetic operations can be performed using simple logic gates with high energy and area efficiency.[5,6] For example, a simple two-bit multiplication in a conventional CMOS-based full adder circuit requires 78 transistors whereas a SC unit can execute the same operation using a single *AND* gate. Similarly, stochastic addition and subtraction can be performed using multiplexer (*MUX*) and *XOR* gates, respectively. The key difference is that, unlike classical computing systems which

H. Ravichandran, Y. Zheng, T. F. Schranghamer, S. Das
Engineering Science and Mechanics
Penn State University
University Park, PA 16802, USA
E-mail: sud70@psu.edu
N. Trainor, J. M. Redwing, S. Das
Materials Science and Engineering
Penn State University
University Park, PA 16802, USA
N. Trainor, J. M. Redwing, S. Das
Materials Research Institute
Penn State University
University Park, PA 16802, USA
S. Das
Electrical Engineering and Computer Science
Penn State University
University Park, PA 16802, USA
The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/adma.202206168.

DOI: 10.1002/adma.202206168

© 2022 Wiley-VCH GmbH

Check for updates
15214095, 2023, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/adma.202206168 by Pennsylvania State University, Wiley Online Library on [08/02/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

represent information in the form of binary logic ("1"s and "0"s), SC encodes information through stochastic bit (s-bit) streams that are interpreted as probabilities that fall in the interval [0,1] in the unipolar format. For instance, the bit-stream $A = \{1\ 0\ 1\ 1\ 0\ 1\ 0\ 0\}$ encodes the value $p_A = 0.5$ since there are four 1's present within the bit-stream of length 8-bit. Note that SC is capable of handling negative numbers by encoding information in bipolar format where the numerical value of A is interpreted as $2p_A - 1$, thereby changing the effective range of SC to [–1,1]. Thus, in the bipolar format, the same bitstream A will encode the value $p_A = 0$. In other words, a bit-stream containing only zeros has a unipolar value 0 and a bipolar value –1. The logic gates required to implement stochastic arithmetic operation also changes depending on the format of the number used. For instance, as mentioned earlier, an *AND* gate can perform multiplication of unipolar stochastic numbers while an *XNOR* gate is required to perform multiplication in bipolar format.[7] For the current study, we have investigated the implementation of stochastic arithmetic pertaining to unipolar stochastic numbers. An attractive feature of SC is its resilience to error tolerance[8] since there is no distinction between the most and the least significant bits; in other words, all s-bits carry equal weight. While promising, the application of SC has largely been limited to specialized domains such as image and audio processing where a finite amount of error or loss in precision is acceptable.[9,10] Such limitations primarily stem from the requirement of having a much longer bit-stream for more accurate probability estimation, leading to a corresponding increase in the computation time and energy. Despite these shortcomings, SC is becoming popular for many AI applications, which deal with large volumes of audio-visual information.[9–12]

The concept of SC is well known and extensively studied. CMOS-based,[13,14] memristor-based,[15–19] and spintronic-based[20–24] SC architectures have already been demonstrated in the past. However, CMOS-based SC architectures require several hundred transistors to generate s-bits, limiting their area and energy efficiency.[25] Stochastic switching in memristors and spintronic devices offers an excellent mechanism to generate fast and random bits.[26–28] However, these emerging devices rely on CMOS peripherals to control the probability of switching for the conversion of random bits into s-bits and for subsequent logic operations using those s-bits, which can ultimately limit the area and energy efficiency.

Here, we overcome the abovementioned limitations by introducing a standalone, in-memory SC architecture based on 2D memtransistors. Memtransistors are programmable field effect transistors (FETs) made from ultrathin-body semiconducting channel materials such as monolayer molybdenum disulfide (MoS_2). We exploit programming stochasticity in memtransistors and design a circuit comprising six memtransistors to obtain s-bits; we then integrate the s-bit generators with 2D-memtransistor-based logic gates such as *AND*, *MUX*, *XOR*, and *OR* gates to demonstrate arithmetic operations such as addition, subtraction, multiplication, and sorting.

2. Fabrication and Characterization of 2D Memtransistors

Optical images of a representative monolayer MoS_2-based 2D memtransistor and the hardware platform for the acceleration of the SC architecture are shown in **Figure 1**a,b. To enable individual electrostatic gate control, each memtransistor is locally back-gated using a stack comprising atomic layer deposition (ALD) grown 50 nm Al_2O_3 on sputter-deposited 50/20 nm Pt/TiN, placed on a commercially purchased SiO_2/p^{++}-Si substrate. The Al_2O_3/Pt/TiN gate stack plays a vital role as it enables stochastic programming of the memtransistor, which in conjunction with the memtransistor-based logic gates allows hardware implementation of SC. Details on the synthesis of monolayer MoS_2, method of film transfer, fabrication of the local back-gate islands, and fabrication of MoS_2 memtransistors, and SC architecture can be found in the Experimental Section and in our previous works.[29–35] Electrical characterization, which includes transfer characteristics, output characteristics, device-to-device variation, analog programming, and short- and long-term retention characteristics of MoS_2 memtransistors can be found in Figures S1 and S2 (Supporting Information).

3. Programming Stochasticity in 2D Memtransistors and s-Bit Generation

Next, we exploit programming stochasticity in 2D memtransistors for the generation of high-quality random bits following our earlier work.[34,36] The programming stochasticity arises due to inherent randomness in the trapping/detrapping of charged carriers in the gate oxide of our 2D memtransistors when subjected to bias voltages. To characterize the programming stochasticity, a representative 2D memtransistor is subjected to 100 cycles of program and erase operations by the application of a negative "Write" voltage pulse, $V_P = -10$ V and a positive "Erase" voltage pulse, $V_E = 10$ V, each for $\tau_s = 100$ µs. Figure 1c shows the transfer characteristics of a representative MoS_2 memtransistor measured after each cycle and Figure S3 (Supporting Information) shows the distribution of conductance (G_{MT}) measured using $V_{BG} = 0$ V. Clearly, the cycle-to-cycle variability in post-programmed and post-erased G_{MT} follow Gaussian random distributions. While programming stochasticity is detrimental for conventional computing, it offers a unique opportunity for SC.

To translate the conductance fluctuation into s-bits, we deploy an s-bit generator module consisting of six memtransistors (MT1, MT2, MT3, MT4, MT5, and MT6). Figure 1d,e, respectively, shows the optical image and the corresponding circuit diagram, of an s-bit generator module which consists of three vital units, namely, a voltage divider, constructed using MT1 and MT2, an inverting amplifier, constructed using MT3 and MT4 and a thresholding inverter, constructed using MT5 and MT6. Figures S4 and S5 (Supporting Information) describe the conversion of conductance fluctuation into s-bit, and Figure 1f shows the probability of obtaining "1" in the bit stream (p_s) as a function of the inversion threshold, V_{IT}, of the thresholding inverter. The seamless generation of s-bits is enabled by the programmability of V_{IT} of the 2D-memtransistor-based inverters, a crucial feature that distinguishes them from conventional CMOS-based inverters. Note that p_s increases monotonically with V_{IT} and is bounded between [0,1]. Figure S6 (Supporting Information) explains in more detail how to set and reset p_s to a desirable value. Note that the same s-bit

© 2022 Wiley-VCH GmbH

15214095, 2023, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/adma.202206168 by Pennsylvania State University, Wiley Online Library on [08/02/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

15214095, 2023, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/adma.202206168 by Pennsylvania State University, Wiley Online Library on [08/02/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

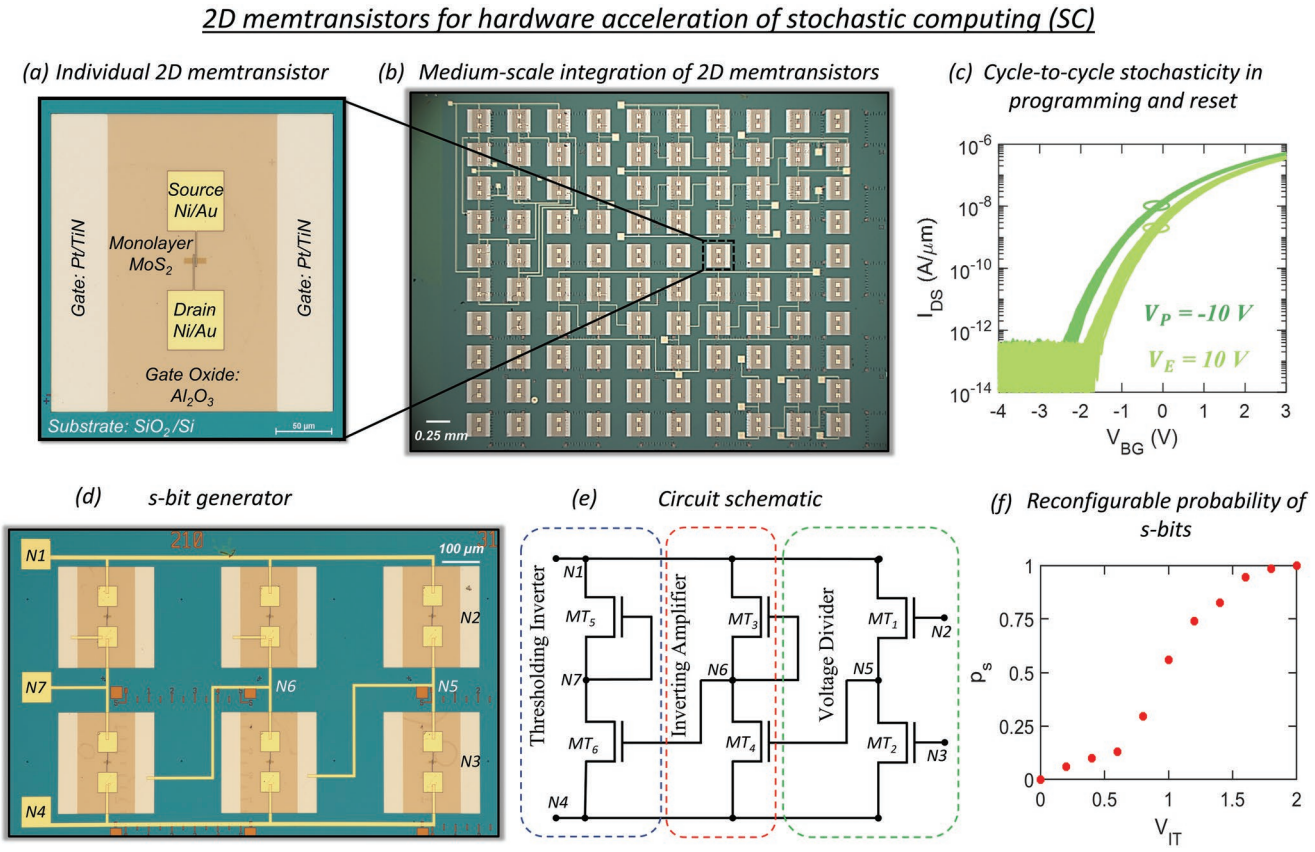


Figure 1. 2D memtransistor-based hardware platform for the acceleration of stochastic computing (SC). a,b) Optical images of a representative 2D memtransistor based on monolayer MoS_2 (a) and medium-scale integrated circuits based on 2D memtransistors (b) for the hardware acceleration of SC. c) Transfer characteristics of a representative 2D memtransistor, measured each time after the application of programming and erase voltage pulses of $V_P = -10$ V and $V_E = 10$ V, respectively, each for $\tau_s = 100$ µs, for a total of 100 cycles. d) Optical image and e) corresponding circuit diagram of the s-bit generator consisting of six memtransistors (MT1, MT2, MT3, MT4, MT5, and MT6). Each s-bit generator has three units: a voltage divider, an inverting amplifier, and a thresholding inverter. f) Probability of obtaining "1" in the s-bit stream (p_s) as a function of inversion threshold, V_{IT}, of the thresholding inverter. This shows the ability of the s-bit generator to transform the cycle-to-cycle conductance fluctuations in 2D memtransistor into bit-streams with reconfigurable p_s that lie between [0,1].

generator can be utilized to generate bipolar s-bits with reconfigurable p_s-values that lie between [–1,1]. Figure S7 (Supporting Information) explains the details of bipolar s-bit generation. To enable this, the voltage applied to node N4 of the s-bit generator is held constant at –1 V instead of 0 V. Nevertheless, it is evident that our six 2D memtransistor-based s-bit generator can seamlessly produce unipolar and bipolar s-bits.

The average energy expenditure for s-bit generation ($E_{s\text{-bit}}$) was calculated using Equation (1);

$$E_{s\text{-bit}} = C_G\left[V_P^2 + V_E^2 + V_{DD}^2\right] + \frac{1}{N}\sum_{i=1}^{N} I_{N1N4-i} V_{DD} \tau_{clk};\ C_G = \varepsilon_0 \varepsilon_{ox} WL / t_{ox} \quad (1)$$

In Equation (1), C_G is the gate capacitance, I_{N1N4-i} is the current flowing through the s-bit generator during each τ_{clk} (see Figure S4, Supporting Information), $\varepsilon_0 = 8.85 \times 10^{-12}$ F m^{-1} is the vacuum permittivity, and $\varepsilon_{ox} = 10$ and $t_{ox} = 50$ nm are, respectively, the relative permittivity and thickness of Al_2O_3. We found that $E_{s\text{-bit}} < 2$ pJ clock-cycle^{-1}, which supports our claim on energy efficient s-bit generation. Note that the second term in Equation (1) accounts for about 0.3 pJ (we have used $N = 200$ to calculate the average current in the s-bit generator per clock cycle). Therefore, it is possible to reduce the energy expenditure even further through scaling of t_{ox}, which will scale the program/erase voltages accordingly. Note that the footprint of each s-bit generator is only 30 µm^2 as each memtransistor has an active device area of ≈5 µm^2 excluding the large contact pads. With monolayer 2D materials offering aggressive dimensional scalability, the active device footprint can be reduced significantly without compromising the s-bit quality.

4. Stochastic Arithmetic Modules

4.1. Multiplication

Stochastic multiplication can be accomplished using a simple *AND* gate as shown in **Figure 2**a. The stochastic output, $C(p_C)$, of an *AND* gate with two stochastic input variables, $A(p_A)$ and $B(p_B)$, is given by Equation (2).

$$C = AB \quad (2a)$$

$$p_C = p_A p_B \quad (2b)$$

© 2022 Wiley-VCH GmbH

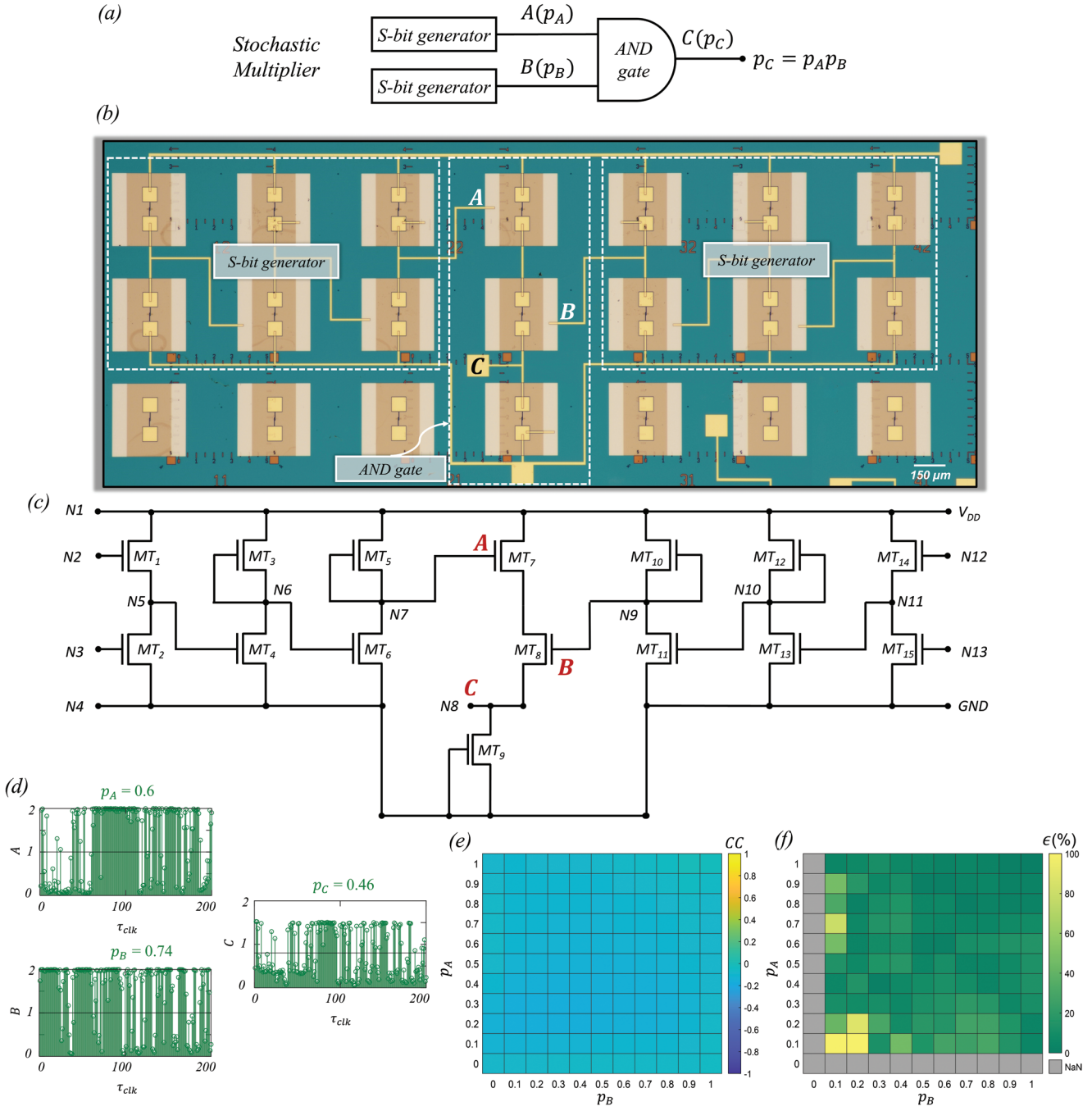


Figure 2. Stochastic multiplier. a) Schematic, b) optical image, and c) corresponding circuit configuration of a stochastic multiplier consisting of two s-bit generator and one *AND* gate with a total of 15 memtransistors. d) Representative stochastic bit-streams for the random variables, $A(p_A)$ and $B(p_B)$ obtained from their respective s-bit generators and the corresponding output bit-stream for $C(p_C)$. e) Colormap of correlation coefficient (*CC*) for different combinations of p_A and p_B lying between [0,1]. f) Colormap of percentage error (ε) for multiplication between all combinations of p_A and p_B lying in the interval [0,1] . Lower values of ε are a direct consequence of near ideal CC values close to zero indicating mutual independence of *A* and *B*, which is critical for accurate multiplication. We have used bit-streams of length 200-bit to evaluate the probability values associated with the random variable.

In Equation (2a) and (2b), p_A, p_B, and p_C, are the probabilities associated with the random variables *A*, *B*, and *C*, respectively. Note that Equation (2) is valid only if the random variables, *A* and *B*, are mutually independent or uncorrelated. Figure 2b,c, respectively, shows the optical image and corresponding circuit configuration of a stochastic multiplier consisting of two s-bit generators and an *AND* gate for a total of 15 memtransistors. The *AND* gate consists of three memtransistors, MT7, MT8, and MT9. Inputs A and B are applied to the local backgates of MT7 and MT8, which are connected in series with MT9 at node N8. The source and gate terminals of MT9 are shorted and connected to the ground. As such, MT9 operates

© 2022 Wiley-VCH GmbH

15214095, 2023, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/adma.202206168 by Pennsylvania State University, Wiley Online Library on [08/02/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

as a load resistor. The output, *C*, of the *AND* gate is obtained at node N8. Figure 2d shows the representative stochastic bit-streams for the random variables $A(p_A = 0.6)$ and $B(p_B = 0.74)$ obtained from their respective s-bit generators by programming V_{IT} and the corresponding output bit-stream for C with $p_C = 0.46$. As mentioned earlier, to obtain an accurate multiplication product, *A* and *B* must be mutually independent. Figure 2e shows the colormap of the correlation coefficient (CC) between the s-bit streams used as *A* and *B* for different combinations of p_A and p_B lying between [0,1]. Figure 2f shows the colormap of percentage error (ε) for multiplication following Equation (3) obtained for different combinations of p_A and p_B. We have used bit-streams of length 200-bit to evaluate the corresponding probability values.

$$\varepsilon = \left| 1 - \frac{(p_C)_{obtained}}{(p_C)_{expected}} \right| \times 100\% \tag{3}$$

In Equation (3), $(p_C)_{obtained}$ and $(p_C)_{expected}$ are the experimentally obtained and theoretically predicted output of the stochastic computation following Equation (2). Lower values of ε are a direct consequence of near ideal CC-values close to zero, indicating mutual independence of *A* and *B* as shown in Figure 2e, which is critical for accurate multiplication. Clearly, our 15 memtransistor circuit can perform stochastic multiplication with high accuracy.

Note that the accuracy can be increased by increasing the length of s-bit streams at the expense of longer computation time, since one s-bit is generated every τ_{clk}. The average energy expenditure for the multiplication operation when 200 τ_{clk} are used is ≈0.8 nJ. Certainly, the energy expense can be further reduced by shortening the length of the s-bit streams at the cost of reduced precision.

4.2. Addition

Stochastic addition operations can be accomplished using a *MUX* as shown in **Figure 3**a. The stochastic output, $C(p_C)$, of a *MUX* with two stochastic input variables, $A(p_A)$ and $B(p_B)$, and a stochastic select line, $S(p_S)$, is given by Equation (4).

$$C = SA + S^cB \tag{4a}$$

$$p_C = p_S p_A + (1 - p_S) p_B \tag{4b}$$

$$p_C = 0.5(p_A + p_B);\ if\ \ p_S = 0.5 \tag{4c}$$

Clearly, for $p_S = 0.5$, one can achieve scaled addition. Figure 3b,c, respectively, shows the optical image and corresponding circuit configuration of a stochastic adder consisting of three s-bit generator modules and one 2×1 *MUX* for a total of 22 memtransistors. The 2×1 *MUX* consists of four memtransistors, MT19, MT20, MT21, and MT22. Note that memtransistors MT19 and MT20 form a NOT gate with stochastic variable *S* as the input and S^c as the output. *S* and S^c are applied to the local back-gates of MT21 and MT22, respectively, which are connected in series at node N19. The stochastic variable *A* is connected to the source terminal of MT21 at node N12, whereas the stochastic variable B is connected to the drain terminal of MT22 at node N13. The output of the *MUX*, i.e., *C*, is obtained at node N19. Figure 3d shows the representative stochastic bit-streams for the random variables $S(p_S = 0.5)$, $A(p_A = 0.28)$, and $B(p_B = 0.55)$ obtained from their respective s-bit generation modules at nodes N7, N12, and N13 and the corresponding output bit-stream for *C* with $p_C = 0.41$. Figure 3e shows the colormap of percentage error (ε) for scaled addition (Equation (3)) for different combinations of p_A and p_B lying in the interval [0,1] for $p_S \approx 0.5$. Clearly, our 22 memtransistor module can perform stochastic addition with high accuracy. The average energy expenditure for the scaled addition operation is ≈1.2 nJ.

4.3. Subtraction

While the circuits used for stochastic multiplication and addition require the stochastic inputs to be independent or uncorrelated to achieve accurate results, stochastic subtraction benefits greatly from the correlation between the stochastic inputs. In fact, Alaghi and Hayes[37] have shown that correlated inputs can drastically alter the functionality of a stochastic circuit, thereby simplifying the hardware acceleration of specific arithmetic operations. For example, if a *XOR* gate (**Figure 4**a) is implemented using two uncorrelated stochastic input variables, $A(p_A)$ and $B(p_B)$, the stochastic output, $C(p_C)$, will be given by Equation (6).

$$C = AB^C + A^CB \tag{6a}$$

$$p_C = p_A(1 - p_B) + p_B(1 - p_A) \tag{6b}$$

However, when *A* and *B* are highly correlated, a *XOR* gate implements absolute-valued subtraction following Equation (9).

$$p_C = |p_A - p_B| \tag{7}$$

As an example, if $A = 01110110$ and $B = 01100010$ are two correlated stochastic streams representing $p_A = 5/8$ and $p_B = 3/8$, then $C = 00010100$ and $p_C = 2/8$. Note that conventional implementation of this function requires one *NOT* gate, one 2×1 *MUX*, and one finite state machine (FSM), increasing the area and energy overhead.[10]

Figure 4b,c, respectively, shows the optical image and corresponding circuit configuration of a *XOR* gate with a total of nine memtransistors. Note that memtransistor pairs MT1 and MT2 and MT5 and MT6 are *NOT* gates used to invert *A* to A^c and *B* to B^c, respectively. *A* and B^c are applied to the local back-gates of MT3 and MT4, respectively, which are connected in series. Similarly, A^c and *B* are applied to the local back-gates of MT7 and MT8, respectively, which are also connected in series. Finally, the series connections of MT3 and MT4 and MT7 and MT8 are connected in parallel between nodes N1 and N5. The drain terminal of MT9 is connected to N5, whereas the source and gate terminals are shorted to the ground. The overall circuit accomplishes the *XOR* logic for inputs *A* and *B* at node

© 2022 Wiley-VCH GmbH

15214095, 2023, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/adma.202206168 by Pennsylvania State University, Wiley Online Library on [08/02/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

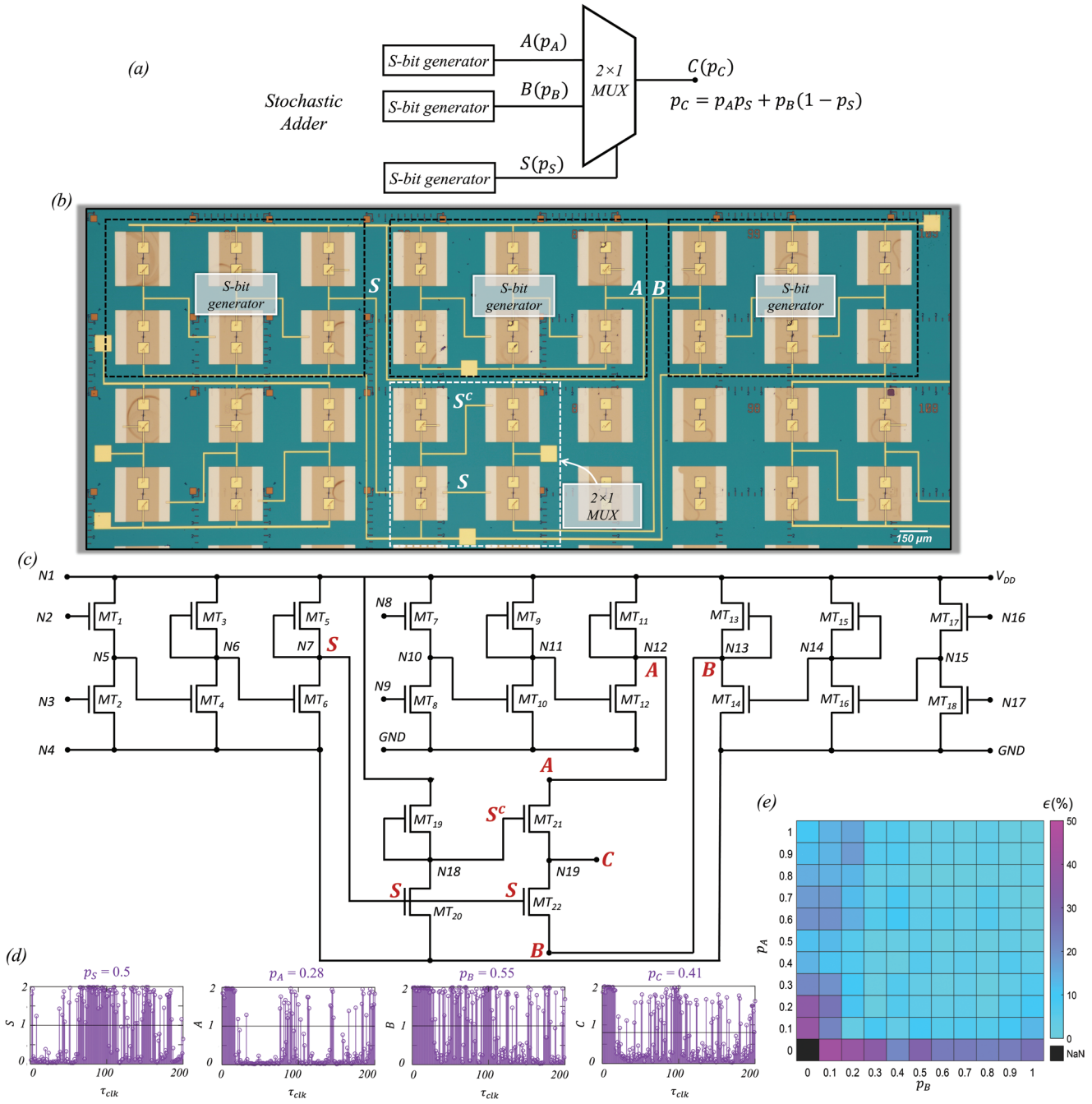


Figure 3. Stochastic adder. a) Schematic, b) optical image, and c) corresponding circuit configuration of a stochastic adder consisting of three s-bit generator and one 2×1 *MUX* gate with a total of 22 memtransistors. d) Representative stochastic bit-streams for the random variables $S(p_S)$, $A(p_A)$, and $B(p_B)$ obtained from their respective s-bit generation modules at nodes N7, N12, and N13 and the corresponding output bit-stream for $C(p_C)$. e) Colormap of percentage errors (ε) for scaled addition for different combinations of p_A and p_B lying between [0,1], and for $p_S \approx 0.5$.

N5. Figure 4d shows the representative stochastic bit-streams for the correlated random variables $A(p_A = 0.85)$, $B(p_B = 0.93)$, and the output of the *XOR* gate, $C(p_C = 0.08)$, which is close to $|p_A - p_B|$. Clearly, our nine memtransistor circuit can perform stochastic subtraction when the stochastic bit-streams are correlated. Note that the CC between A and B was intentionally made high, ≈0.88, by using a correlator circuit, as described below.

While the s-bit generators produce uncorrelated bit-streams, correlated random variables can be created by using an *OR* gate, as shown in Figure 4e. The optical image and corresponding circuit configuration of an *OR* gate comprising three memtransistors are shown in Figure 4f,g, respectively. Two mutually independent/uncorrelated stochastics inputs, A and B, obtained from the s-bit generators are applied to the local back-gates of MT1 and MT2, which are connected in parallel among

© 2022 Wiley-VCH GmbH

15214095, 2023, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/adma.202206168 by Pennsylvania State University, Wiley Online Library on [08/02/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

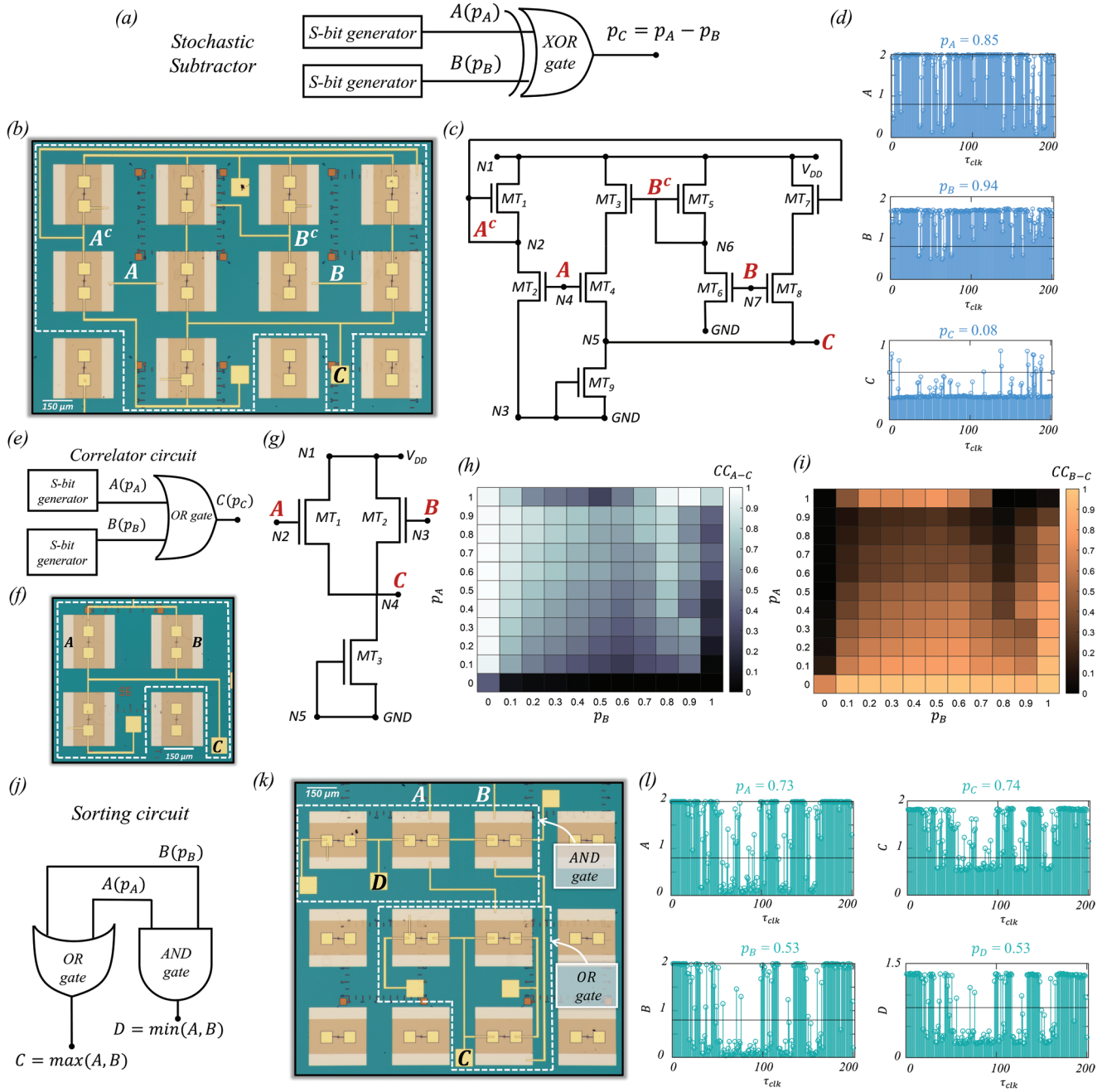


Figure 4. Stochastic subtraction and sorting using correlated s-bits. a) Schematic, b) optical image, and c) corresponding circuit configuration for stochastic subtraction using one *XOR* gate consisting of 9 memtransistors. d) Representative stochastic bit-streams for the random variables $A(p_A)$ and $B(p_B)$, which are highly correlated with CC = 0.88, and the corresponding output bit-stream for $C(p_C)$. e) Schematic, f) optical image, and g) corresponding circuit configuration for a correlator circuit based on *OR* gate consisting of three memtransistors. h) Colormaps of correlation coefficient between the output *C* and input *A* (CC_{A-C}) and input *B* (CC_{B-C}) for different combinations of p_A and p_B in the interval [0,1]. i) Schematic, j) optical image, and k) corresponding circuit configuration of a sorting circuit comprising of one *OR* gate and one *AND* gate. l) Representative stochastic bit-streams for the correlated random variables *A*, *B*, and the sorted output *C* for maximum and *D* for minimum values, respectively.

themselves and in series with MT3. As explained earlier, MT3 operates as a load resistor and the entire circuit serves as an *OR* gate. Interestingly, the output, *C*, obtained at node N4 becomes correlated with either or both *A* and *B*. Figure 4h,i shows the correlation coefficient between *C* and *A*, i.e., CC_{A-C}, and *C* and *B*, i.e., CC_{B-C}, respectively, for different values of p_A and p_B lying between [0,1]. Clearly, CC_{A-C} and CC_{B-C} values range from ≈0 to ≈1. Also note that lower p_A values ensure higher correlation between *C* and *B*, and vice versa. Nevertheless, the correlator circuit allows us to obtain a correlated bit-stream with desirable correlation coefficients. The average energy expenditure for obtaining a correlated bit stream is ≈0.8 nJ.

© 2022 Wiley-VCH GmbH

15214095, 2023, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/adma.202206168 by Pennsylvania State University, Wiley Online Library on [08/02/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

15214095, 2023, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/adma.202206168 by Pennsylvania State University, Wiley Online Library on [08/02/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

4.4. Sorting

As we have shown earlier, an *AND* gate functions as a stochastic multiplier for uncorrelated bit-streams. However, when the inputs become highly correlated, it gives the minimum of the two stochastic streams. For example, if A = 01101110 and B = 01100100 are two correlated stochastic streams representing p_A = 5/8 and p_B = 3/8, then C = 01100100 and p_C = 3/8. Similarly, an *OR* gate gives the maximum value of two stochastic streams, i.e., C = 01101110 and p_C = 5/8. In contrast, conventional implementation[9] with uncorrelated inputs requires FSM-based stochastic hyperbolic tangent (tanh) function along with the three *MUX*s, which again increases area and energy overhead. Figure 4j,k, respectively, shows the schematic and optical image of a six memtransistor sorting circuit, which finds the minimum and maximum between two stochastic variables, A and B. Figure 4l shows the representative stochastic bit-streams for the correlated random variables, A and B, and the sorted output C for maximum and D for minimum values, respectively.

Table 1 summarizes the SC architectures for different arithmetic operations involving medium scale integration (MSI) of 2D memtransistors along with their respective energy expenditure. Our future goal is to expand the SC architecture to accelerate Bayesian neural networks, invertible logic, and solve various combinatorial optimization problems such as the traveling salesman problem. While our ultimate objective is to realize all peripherals using 2D memtransistors, in the short-term 2D memtransistor-based stochastic computing hardware can benefit from integration with mature Si CMOS technology. In fact, it is possible that the 2D memtransistor and CMOS technology can synergistically co-exist. Also note that very large-scale integration (VLSI) of 2D memtransistors is non-trivial as multiple challenges must be overcome.[38] While there has been tremendous progress on large-area growth of a wide range of 2D materials,[39] there is still room to minimize growth defects to achieve higher performance and increase growth uniformity to ensure low device-to-device variation. At the same time, large-area transfer[40] of 2D materials must be improved for cleaner and mechanical damage-free transfer, ensuring high yield during device fabrication. Finally, the future roadmap for 2D memtransistors will involve scaling of channel lengths and oxide thicknesses. While earlier experimental reports,[41,42] including some of our own work[29] and theoretical projections,[43–45] do indicate that 2D-material-based FETs can meet the requirements set forth by the International Roadmap for Devices and Systems (IRDS 2028), programmability of scaled memtransistors needs to be investigated.

5. Conclusion

We have exploited cycle-to-cycle variability in the programmed conductance states of monolayer-MoS_2-based 2D memtransistors to create s-bits with a reconfigurable probability of obtaining "1" in the bit-stream using an s-bit generator circuit comprising six memtransistors. The s-bit generator was subsequently combined with 2D-memtransistor-based logic gates to demonstrate a standalone SC architecture that can perform accurate arithmetic operations such as addition, subtraction, multiplication, and sorting. Our SC architecture consumes miniscule energy (≈1 nJ) to perform arithmetic operations and uses a limited number of memtransistors to achieve a small active-area footprint. Our demonstration offers a way to accelerate SC using a non-von-Neumann platform based on novel 2D materials and devices.

6. Experimental Section

Fabrication of Local Back-Gate Islands: To define the back-gate island regions, the substrate (285 nm SiO_2 on p^{++}-Si) was spin coated with bilayer photoresist consisting of Lift-Off-Resist (LOR 5A) and Series Photoresist (SPR 3012) baked at 185 and 95 °C, respectively. The bilayer photoresist was then exposed using a Heidelburg Maskless Aligner (MLA 150) to define the island regions 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. An atomic layer deposition (ALD) process was then implemented to grow 50 nm Al_2O_3 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 again exposed using the MLA 150 and developed using MF CD26 microposit. Al_2O_3 (50 nm) was subsequently dry etched using a BCl_3 chemistry at 5 °C for 20 s, which was repeated four times to minimize heating in the substrate. Finally, the photoresist was removed to give access to the individual Pt electrodes.

Large-Area Monolayer MoS_2 Film Growth: Monolayer MoS_2 was deposited on epi-ready 2″ *c*-sapphire substrate by metal–organic 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 (H_2S) were used as precursors. $Mo(CO)_6$ maintained at 10 °C and 650 Torr in a stainless-steel bubbler was used to deliver 1.1×10^{-3} sccm of the metal precursor for the growth, while 400 sccm of H_2S was used for the process. MoS_2 deposition was carried out at 1000 °C and 50 Torr in H_2 ambient, where monolayer growth was achieved in 18 min. The substrate was first heated to 1000 °C in H_2 and maintained for 10 min before the growth was initiated. After growth, the substrate was cooled in H_2S to 300 °C to inhibit the decomposition of the MoS_2 films. More details can be found in earlier works.[29,47,48]

Table 1. Summary of the SC architecture for different arithmetic operations.

Arithmetic operation	# of s-bit generators	Logic gates	# of memtransistors	Average energy expenditure
Multiplication	2	*AND*	15	≈0.8 nJ
Addition	3	2 × 1 *MUX*	22	≈1.2 nJ
Subtraction (correlated s-bits)	2	*XOR*	24 (including correlator circuit)	≈0.8 nJ
Sorting (correlated s-bits)	2	*OR* , *AND*	21 (including correlator circuit)	≈0.8 nJ

© 2022 Wiley-VCH GmbH

MoS_2 Film Transfer to Local Back-Gate Islands: To fabricate the 2D memtransistors, MOCVD-grown monolayer MoS_2 film was transferred from the sapphire growth substrate to the SiO_2/p^{++}-Si application substrate with local back-gate islands using a poly(methyl methacrylate) (PMMA) assisted wet transfer process. First, the as-grown MoS_2 on the sapphire substrate was spin-coated with PMMA and baked at 180 °C for 90 s. The corners of the spin-coated film were scratched using a razor blade and immersed inside 2 M NaOH solution kept at 90 °C. Capillary action caused the NaOH to be drawn into the substrate/film interface, separating the PMMA/MoS_2 film from the sapphire substrate. The separated film was rinsed multiple times inside separate water baths and finally transferred onto the application substrate and baked at 50 and 70 °C for 10 min each to remove moisture and promote adhesion, ensuring a pristine interface. Finally, the PMMA layer was removed using acetone and the film was cleaned using IPA.

Fabrication of 2D Memtransistors: To define the channel regions for the memtransistors, the substrate was spin-coated with PMMA and baked at 180 °C for 90 s. The resist was then patterned using an electron beam (e-beam) and developed using 1:1 mixture of 4-methyl-2-pentanone (MIBK) and IPA. The monolayer MoS_2 film was subsequently etched using sulfur hexafluoride (SF_6) 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 was then spin-coated with methyl methacrylate (MMA) followed by A3 PMMA. E-beam lithography was then used to define the source and drain contacts, with development again being performed by using 1:1 mixture of MIBK and IPA for 60 s. 40 nm of nickel (Ni) and 30 nm of gold (Au) were deposited using e-beam evaporation to act as the contact metals. Finally, a lift-off process was 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 min. Each island contains one memtransistor to allow for individual gate control.

Monolithic Integration: To define the connections between the respective memtransistors, the substrate was spin-coated with MMA and PMMA, patterned using e-beam lithography, and developed using 1:1 mixture of MIBK and IPA. E-beam evaporation was used then used to deposit 60 nm Ni and 30 nm Au. Finally, the e-beam resist was rinsed away in a lift-off process using acetone and IPA.

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

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

The work was supported by Army Research Office (ARO) through Contract Number W911NF1920338. The 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.

Author Contributions

S.D conceived the idea and designed the experiments. H.R., Y.Z., and T.F.S. fabricated the memtransistors. H.R., Y.Z., and S.D performed the measurements, analyzed the data, discussed the results, and agreed on their implications. N.T. grew MOCVD MoS_2. All authors contributed to the preparation of the final manuscript.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords

2D materials, arithmetic, memtransistors, stochastic computing

Received: July 13, 2022
Revised: October 4, 2022
Published online: December 8, 2022

[1] D. A. Reed, J. Dongarra, *Commun. ACM* **2015**, *58*, 56.
[2] X. Liao, K. Lu, C. Yang, J. Li, Y. Yuan, M. Lai, L. Huang, P. Lu, J. Fang, J. Ren, J. Shen, *Front. Inf. Technol. Electron. Eng.* **2018**, *19*, 1236.
[3] M. Yamaoka, T. Okuyama, S. Tanaka, M. Hayashi, C. Yoshimura, *Hitachi Rev.* **2015**, *64*, 525.
[4] G. Indiveri, S.-C. Liu, *Proc. IEEE* **2015**, *103*, 1379.
[5] B. R. Gaines, in *Proc. April 18–20, 1967, Spring Joint Computer Conf.*, ACM, New York **1967**, pp. 149–156.
[6] W. Poppelbaum, C. Afuso, J. Esch, in *Proc. November 14–16, 1967*, Fall Joint Computer Conf, ACM, New York **1967**, pp. 635–644.
[7] A. Alaghi, W. Qian, J. P. Hayes, *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **2018**, *37*, 1515.
[8] A. Alaghi, J. Hayes, *ACM Trans. Embed. Comput. Syst.* **2013**, *12*, 92:1.
[9] P. Li, D. J. Lilja, W. Qian, K. Bazargan, M. D. Riedel, *IEEE Trans. Very Large Scale Integr. VLSI Syst* **2013**, *22*, 449.
[10] P. Li, D. J. Lilja, in *2011 IEEE 29th Int. Conf. on Computer Design (ICCD)*, IEEE, Piscataway, NJ, USA **2011**, pp. 154–161.
[11] S. Y. Jang, Y. H. Yoon, S. E. Lee, in *2020 IEEE International Conference on Consumer Electronics (ICCE)*, **2020**, https://doi.org/10.1109/ICCE46568.2020.9042978.
[12] A. Alaghi, C. Li, J. P. Hayes, in *2013 50th ACM/EDAC/IEEE Design Automation Conf. (DAC)*, IEEE, Piscataway, NJ, USA **2013**, https://doi.org/10.1145/2463209.2488901.
[13] S. C. Smithson, N. Onizawa, B. H. Meyer, W. J. Gross, T. Hanyu, *IEEE Trans. Circuits Syst. I* **2019**, *66*, 2263.
[14] A. Ardakani, F. Leduc-Primeau, N. Onizawa, T. Hanyu, W. J. Gross, *IEEE Trans. Very Large Scale Integr. VLSI Syst.* **2017**, *25*, 2688.
[15] P. Knag, W. Lu, Z. Zhang, *IEEE Trans. Nanotechnol.* **2014**, *13*, 283.
[16] S. Gaba, P. Knag, Z. Zhang, W. Lu, in *2014 IEEE Int. Symp. on Circuits and Systems (ISCAS)*, IEEE, Piscataway, NJ, USA **2014**, pp. 2592–2595.
[17] S. Gaba, P. Sheridan, J. Zhou, S. Choi, W. Lu, *Nanoscale* **2013**, *5*, 5872.
[18] C. Lammie, J. K. Eshraghian, W. D. Lu, M. R. Azghadi, *IEEE Trans. Circuits Syst. II* **2021**, *68*, 1650.
[19] P. K. Reddy Boppidi, B. Suresh, G. Abhijith, P. M. Preetam Raj, S. Banerjee, S. Kundu, *Mater. Sci. Semicond. Process.* **2022**, *139*, 106355.
[20] R. Venkatesan, S. Venkataramani, X. Fong, K. Roy, A. Raghunathan, in *2015 Design, Automation & Test in Europe Conf. Exhibition (DATE)*, IEEE, Piscataway, NJ, USA **2015**, pp. 1575–1578.
[21] G. Finocchio, M. Di Ventra, K. Y. Camsari, K. Everschor-Sitte, P. K. Amiri, Z. Zeng, *arXiv: 1910.07176*, **2019**.

© 2022 Wiley-VCH GmbH

15214095, 2023, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/adma.202206168 by Pennsylvania State University, Wiley Online Library on [08/02/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

[22] J. Hu, B. Li, C. Ma, D. Lilja, S. J. Koester, *IEEE Trans. Electron. Devices* **2019**, *66*, 3620.
[23] J. Kaiser, W. A. Borders, K. Y. Camsari, S. Fukami, H. Ohno, S. Datta, *Phys. Rev. Appl.* **2022**, *17*, 014016.
[24] E. Becle, G. Prenat, P. Talatchian, L. Anghel, I. L. Prejbeanu, *IEEE Trans. Circuits Syst. I* **2022**, *69*, 3251.
[25] K. Yang, D. Fick, M. B. Henry, Y. Lee, D. Blaauw, D. Sylvester, in *2014 IEEE Int. Solid-State Circuits Conf. Dig. Tech. Pap. (ISSCC)*, IEEE, Piscataway, NJ, USA **2014**, pp. 280–281.
[26] W. Sun, B. Gao, M. Chi, Q. Xia, J. J. Yang, H. Qian, H. Wu, *Nat. Commun.* **2019**, *10*, 3453.
[27] A. Jaiswal, X. Fong, K. Roy, *IEEE J. Emerging Sel. Top. Circuits Syst.* **2016**, *6*, 120.
[28] A. Sengupta, P. Panda, P. Wijesinghe, Y. Kim, K. Roy, *Sci. Rep.* **2016**, *6*, 30039.
[29] A. Sebastian, R. Pendurthi, T. H. Choudhury, J. M. Redwing, S. Das, *Nat. Commun.* **2021**, *12*, 693.
[30] R. Pendurthi, D. Jayachandran, A. Kozhakhmetov, N. Trainor, J. A. Robinson, J. M. Redwing, S. Das, *Small* **2022**, *18*, 2202590.
[31] S. Subbulakshmi Radhakrishnan, S. Chakrabarti, D. Sen, M. Das, T. F. Schranghamer, A. Sebastian, S. Das, *Adv. Mater.* **2022**, *n/a*, 2202535.
[32] A. Dodda, N. Trainor, J. M. Redwing, S. Das, *Nat. Commun.* **2022**, *13*, 3587.
[33] A. Oberoi, A. Dodda, H. Liu, M. Terrones, S. Das, *ACS Nano* **2021**, *15*, 19815.
[34] Y. Zheng, H. Ravichandran, T. F. Schranghamer, N. Trainor, J. M. Redwing, S. Das, *Nat. Commun.* **2022**, *13*, 5578.
[35] S. Chakrabarti, A. Wali, H. Ravichandran, S. Kundu, T. F. Schranghamer, K. Basu, S. Das, *ACS Appl. Nano Mater.* **2022**, *5*, 14447.
[36] A. Wali, H. Ravichandran, S. Das, *ACS Nano* **2021**, *15*, 17804.
[37] A. Alaghi, J. P. Hayes, in *2013 IEEE 31st Int. Conf. on Computer Design (ICCD)*, IEEE, Piscataway, NJ, USA **2013**, pp. 39–46.
[38] S. Das, A. Sebastian, E. Pop, C. J. McClellan, A. D. Franklin, T. Grasser, T. Knobloch, Y. Illarionov, A. V. Penumatcha, J. Appenzeller, Z. Chen, W. Zhu, I. Asselberghs, L.-J. Li, U. E. Avci, N. Bhat, T. D. Anthopoulos, R. Singh, *Nat. Electron.* **2021**, *4*, 786.
[39] 2DCC, *2d-crystal-consortium*; Available at: https://www.mri.psu.edu/2d-crystal-consortium/user-facilities/thin-films/list-thin-film-samples-available.
[40] T. F. Schranghamer, M. Sharma, R. Singh, S. Das, *Chem. Soc. Rev.* **2021**, 11032.
[41] Q. Smets, G. Arutchelvan, J. Jussot, D. Verreck, I. Asselberghs, A. N. Mehta, A. Gaur, D. Lin, S. E. Kazzi, B. Groven, M. Caymax, I. Radu, in *2019 IEEE Int. Electron Devices Meeting (IEDM)*, IEEE, Piscataway, NJ, USA **2019**, https://doi.org/10.1109/IEDM19573.2019.8993650.
[42] I. Asselberghs, Q. Smets, T. Schram, B. Groven, D. Verreck, A. Afzalian, G. Arutchelvan, A. Gaur, D. Cott, T. Maurice, S. Brems, K. Kennes, A. Phommahaxay, E. Dupuy, D. Radisic, J. F. d. Marneffe, A. Thiam, W. Li, K. Devriendt, C. Huyghebaert, D. Lin, M. Caymax, P. Morin, I. P. Radu, in *2020 IEEE Int Electron Devices Meeting (IEDM)*, IEEE, Piscataway, NJ, USA, **2020**, https://doi.org/10.1109/IEDM13553.2020.9371926.
[43] T. K. Agarwal, B. Soree, I. Radu, P. Raghavan, G. Iannaccone, G. Fiori, W. Dehaene, M. Heyns, *Sci. Rep.* **2017**, *7*, 5016.
[44] D. E. Nikonov, I. A. Young, *IEEE J. Explor. Solid-State Comput. Devices Circuits* **2015**, *1*, 3.
[45] S. S. Sylvia, K. Alam, R. K. Lake, *IEEE J. Explor. Solid-State Comput. Devices Circuits* **2016**, *2*, 28.
[46] Y. Xuan, A. Jain, S. Zafar, R. Lotfi, N. Nayir, Y. Wang, T. H. Choudhury, S. Wright, J. Feraca, L. Rosenbaum, J. M. Redwing, V. Crespi, A. C. T. van Duin, *J. Cryst. Growth* **2019**, *527*, 125247.
[47] D. Jayachandran, A. Oberoi, A. Sebastian, T. H. Choudhury, B. Shankar, J. M. Redwing, S. Das, *Nat. Electron.* **2020**, *3*, 646.
[48] A. Dodda, A. Oberoi, A. Sebastian, T. H. Choudhury, J. M. Redwing, S. Das, *Nat. Commun.* **2020**, *11*, 4406.

© 2022 Wiley-VCH GmbH

15214095, 2023, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/adma.202206168 by Pennsylvania State University, Wiley Online Library on [08/02/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

nature communications

Article

https://doi.org/10.1038/s41467-022-33699-7

Two-dimensional materials-based probabilistic synapses and reconfigurable neurons for measuring inference uncertainty using Bayesian neural networks

Received: 2 February 2022

Accepted: 27 September 2022

Published online: 17 October 2022

Check for updates

Amritanand Sebastian[1] ✉, Rahul Pendurthi[1], Azimkhan Kozhakhmetov[2], Nicholas Trainor[2,3], Joshua A. Robinson[2,4,5], Joan M. Redwing[2,3,6] & Saptarshi Das[1,2,6] ✉

Artificial neural networks have demonstrated superiority over traditional computing architectures in tasks such as pattern classification and learning. However, they do not measure uncertainty in predictions, and hence they can make wrong predictions with high confidence, which can be detrimental for many mission-critical applications. In contrast, Bayesian neural networks (BNNs) naturally include such uncertainty in their model, as the weights are represented by probability distributions (e.g. Gaussian distribution). Here we introduce three-terminal memtransistors based on two-dimensional (2D) materials, which can emulate both probabilistic synapses as well as reconfigurable neurons. The cycle-to-cycle variation in the programming of the 2D memtransistor is exploited to achieve Gaussian random number generator-based synapses, whereas 2D memtransistor based integrated circuits are used to obtain neurons with hyperbolic tangent and sigmoid activation functions. Finally, memtransistor-based synapses and neurons are combined in a cross-bar array architecture to realize a BNN accelerator for a data classification task.

Machine learning has seen unprecedented growth and success in recent years owing to the development of artificial neural networks (ANNs). By mimicking the biological neural architecture and employing deep learning algorithms, ANNs have demonstrated notable advantages over standard computing methods for tasks such as image classification, facial recognition, data mining, weather forecasting, and stock market prediction[1–5]. While ANNs offer high performance, especially in terms of high prediction accuracy, they often suffer from overfitting due to a lack of generalization as they do not model uncertainty. Large datasets and various regularization techniques are often required to reduce overfitting in ANNs[6]. However, this can limit the use of ANN in applications where the data is scarce. In addition, uncertainty estimation is important in applications like autonomous driving, and medical diagnostics, where machine learning must be complemented with uncertainty-aware models or human intervention[7,8]. The integration of probabilistic computing paradigms with ANNs allows regularization and enables us to model uncertainty in predictions[9–12]. This is achieved in Bayesian neural networks (BNNs) by incorporating Bayes theorem to the traditional neural network scheme[12,13]. BNNs are capable of modeling uncertainty and avoiding overfitting, while working well with small

[1]Deparment of Engineering Science and Mechanics, Penn State University, University Park, PA 16802, USA. [2]Department of Materials Science and Engineering, Penn State University, University Park, PA 16802, USA. [3]2D Crystal Consortium Materials Innovation Platform, Penn State University, University Park, PA 16802, USA. [4]Department of Chemistry, Penn State University, University Park, PA, USA. [5]Department of Physics, Penn State University, University Park, PA, USA. [6]Department of Electrical Engineering and Computer Science, Penn State University, University Park, PA, USA. ✉e-mail: amritsebastian@gmail.com; sud70@psu.edu

datasets[14]. In fact, BNNs are extremely powerful as they represent an ensemble model, which is equivalent to the combinations of numerous ANNs, but with a small number of parameters. Unlike ANNs, where the synaptic weights are point estimates (single-valued), in BNNs, the weights (W) are represented by probability distributions, as shown in Fig. 1.

Over the years we have witnessed the development of neural network accelerators aimed at improving the size, energy consumption, and speed of neural networks, especially for edge computing applications[15–17]. Since the training process in neural networks is energy and resource intensive, these works typically rely on off-chip training and on-chip inference. Hence, BNN accelerators have also mostly focused on implementing Bayesian inference on-chip[18–23]. A crucial component of the BNN accelerator is an on-chip Gaussian random number generator (GRNG)-based synapse that can sample weights from a Gaussian distribution. In addition, a BNN requires a circuit to implement a neuron, i.e., to perform the multiply and accumulate (MAC) operation and the neural activation. BNN implementations based on Si complementary metal oxide semiconductor (CMOS) and field-programmable gate array (FPGA) typically require elaborate hardware for GRNGs, MAC operation, and the activation function, rendering them area and energy inefficient[18–20]. Moreover, these demonstrations are based on the von-Neumann architecture with separate memory and logic units, requiring frequent shuttling of data between the two. BNN accelerators based on emerging and non-von Neumann memristive and spintronic synapses utilize cycle-to-cycle variability in switching to generate Gaussian random numbers (GRNs)[21–23]. However, these GRNG-based synapses are limited to mean (μ) of 0 and standard deviation (σ) of 1 and require extensive CMOS-based peripheral circuitry to obtain unrestricted μ and σ values. For example, multiplication and addition operations are used to transform $N(0,1)$ to $N(\mu,\sigma) = \sigma^{*}N(0,1) + \mu$. Finally, two-terminal memristors also lack the capability to emulate neurons for the activation functions. Therefore, energy and area efficient acceleration of BNNs will benefit from a standalone hardware platform, which can offer both neurosynaptic functionalities as well as programmable stochasticity.

In this work, we introduce three-terminal memtransistor technology based on two-dimensional (2D) monolayer MoS_2 and WSe_2 offering all computational primitives needed for a BNN accelerator. First, we realize an ultra-low power GRNG-based synapse by exploiting the cycle-to-cycle variability in programming/erasing operation in the MoS_2 memtransistor. Next, using a circuit comprising of two memtransistors we achieve reconfigurable μ and σ. Activation functions such as hyperbolic tangent (tanh) and sigmoid are also realized using the integration of n-type MoS_2 and V-doped p-type WSe_2 memtransistors. Finally, we demonstrate a crossbar array architecture in order to implement on-chip BNN inference. Furthermore, the entire network is simulated using LTSpice and uncertainty decomposition is performed to identify the various sources of uncertainty.

Results

2D memtransistors

This 2D memtransistor has a back-gated geometry, where, Al_2O_3 is used as the gate dielectric and TiN/Pt is used as the gate electrode (see the "Methods" section and our earlier publications[24–27]for details on fabrication). We have used monolayer MoS_2[28,29] and V-doped WSe_2[30] grown using metal organic chemical vapor deposition (MOCVD) described in our previous reports as the channel materials. The choice of 2D materials is motivated by recent demonstrations highlighting the technological viability of 2D materials[28,31,32] and their wide-scale adoption in brain-inspired computing[33–42]. The transfer characteristics, i.e., drain current (I_{DS}) versus gate-to-source voltage (V_{GS}) for drain-to-source voltage (V_{DS}) of 1 V for 250 MoS_2 memtransistor demonstrating unipolar n-type behavior and the distribution of their threshold voltage ($V_{TH,n}$) are shown in Fig. 2a, b. The transfer characteristics for 20 V-doped WSe_2 memtransistors demonstrating unipolar p-type behavior is shown in Fig. 2c. Here the WSe_2 is doped with 1.1% V to realize unipolar p-type characteristics[30]. Figure 2d shows the distribution of the threshold voltage ($V_{TH,p}$) for WSe_2 memtransistors. Here, the threshold voltages are extracted using the constant-current method, at 0.1 nA μm^{-1}. The device-to-device variation is seen to be low in both cases.

The memtransistors offers analog and non-volatile charge-trap memory. The memtransistor can be programmed i.e., the threshold voltage of the device can be decreased by applying a program pulse to the back-gate with a large negative voltage (V_P). Figure 2e demonstrates the post-programmed transfer characteristics of a MoS_2 memtransistor for V_{DS} = 0.1 V, measured after programming with different V_P. Similarly, MoS_2 memtransistor can be erased i.e., the threshold voltage of the device can be increased by applying an erase pulse to the back-gate with a large positive voltage (V_E), as shown in Supplementary Fig. 1b. A pulse duration (t) of 100 μs is used for both programming and erasing. The dependence of programming and erasing, on t_P is shown in Supplementary Fig. 1c, d, respectively. The non-volatile nature of the MoS_2 memtransistor is shown in Fig. 2f, where the retention characteristics for five different conductance states are demonstrated for 2000 s. The characterization of memory of WSe_2 memtransistors is shown in Supplementary Fig. 2. The working principle of this analog and non-volatile memory has been described in detail in our earlier report[35].

Gaussian random number generator-based synapse

We have demonstrated Gaussian synapses for a probabilistic neural network in our prior work, where the synapses were realized by mimicking the Gaussian function[33]. However, BNN accelerators require GRNGs and they typically rely on techniques such as cumulative density function inversion, central limit theorem (CLT)-based approximation, and the Wallace method to sample standard GRNs[18–20]. These methods typically require linear feedback shift registers, multipliers, and adders, involving numerous transistors to implement the GRNGs, rendering them area and energy inefficient. In contrast, here we use

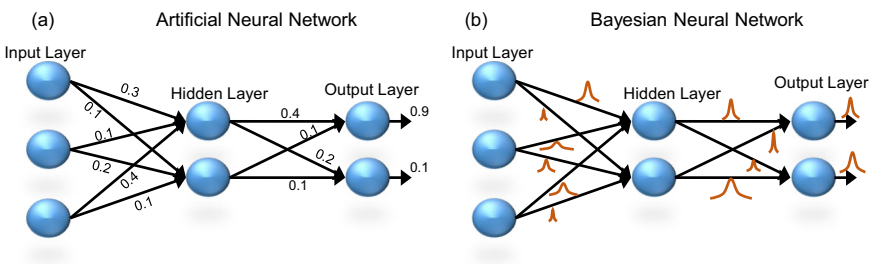


Fig. 1 | Comparison of an artificial neural network (ANN) and a Bayesian neural network (BNN). Schematic of **a** an ANN and **b** a BNN. The synapses of ANN are represented by single-valued weights while the synapses of a BNN is represented by probability distributions.

https://doi.org/10.1038/s41467-022-33699-7

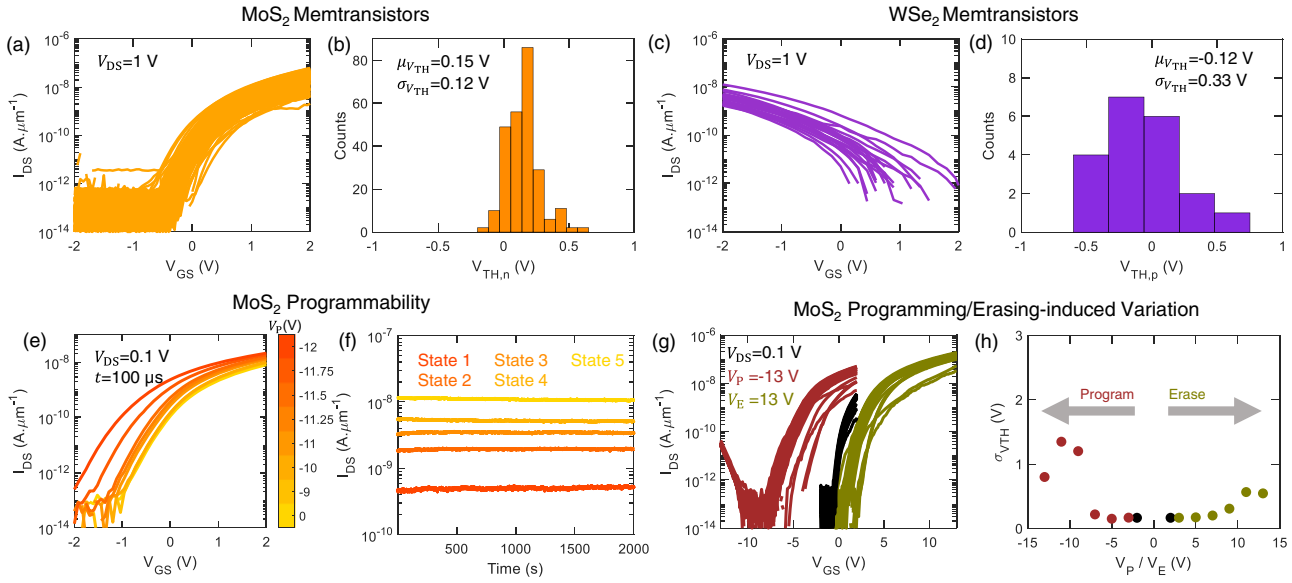


Fig. 2 | Programmable memtransistors. a Device-to-device variation in the transfer characteristics, i.e., drain current (I_{DS}) versus gate-to-source voltage (V_{GS}) for drain-to-source voltage (V_{DS}) of 1 V for 250 MoS_2 memtransistors, demonstrating unipolar *n*-type transport. Distribution of **b** threshold voltage ($V_{TH,n}$) extracted from 250 MoS_2 memtransistors. **c** Device-to-device variation in the transfer characteristics for 20 V-doped WSe_2 memtransistors, demonstrating unipolar *p*-type transport. Distribution of **d** threshold voltage ($V_{TH,p}$) extracted from 20 WSe_2 memtransistors. The corresponding means (μ) and standard deviations (σ) are denoted in the inset. Memtransistors offers analog and non-volatile memory, where their threshold voltage can be adjusted by applying a programming pulse to the back gate. **e** Transfer characteristics of a post-programmed MoS_2 memtransistor obtained by applying negative programming voltages (V_P) of different amplitudes for pulse duration (t) of 100 µs. **f** Analog retention characteristics, i.e., post-programmed I_{DS} versus time measured at V_{GS} = 0 V, for five different states. **g** The impact of programming/erasing on device-to-device variation, demonstrated using different V_{GS} sweep ranges. **h** $\sigma_{V_{TH,n}}$ as a function of V_P and erasing voltage (V_E) applied during V_{GS} sweeps. An increase in variation is seen for high V_P/V_E.

cycle-to-cycle variation in the programmability of our MoS_2 memtransistor to generate GRNs. While cycle-to-cycle variation is undesirable for traditional computing, it can be exploited to reduce the design complexity of a BNN accelerator[21,22,43]. To demonstrate the effect of programming on variation, we use dynamic programming on 40 MoS_2 memtransistors, where we measure the transfer characteristics with different V_{GS} sweep ranges. To evaluate the effect of V_P (V_E), the maximum positive (negative) V_{GS} is fixed at +2 V (−2 V), while the maximum negative (positive) V_{GS} is stepped from −3 to −13 V (3 to 13 V). As shown in Fig. 2g, high V_P and V_E (±13 V) increases the device-to-device variation (post programming/erasing). High V_P and V_E (beyond ±7 V) results in significant V_{TH} shift (see Supplementary Fig. 3), while also increasing $\sigma_{V_{TH}}$, as shown in Fig. 2h. This increase in device-to-device variation for high V_P and V_E is also accompanied by an increase in the cycle-to-cycle variation.

To utilize the cycle-to-cycle variation, the gate of a MoS_2 memtransistor is subjected to successive erase-program-read pulse cycles with V_E = 13 V, V_P = −13 V, and read voltage (V_R) of 0 V as shown in Fig. 3a. The corresponding V_{DS} values were 0, 0 and 0.1 V, respectively. The conductance (G) of the memtransistor, measured at each read step, is shown in Fig. 3b for 200 cycles. As evident from the histogram shown in Fig. 3c, G follows a Gaussian distribution, with mean, μ_G = 3.5 nS and, standard deviation, σ_G = 0.9 nS. The quantile-quantile (Q-Q) plot of G further confirms the Gaussian distribution. The quantiles of G (represented using circles) are plotted against the theoretical quantiles from a Gaussian distribution as shown in Fig. 3d. As expected, it closely follows a straight line. Note that the slope of the Q-Q plot represents σ_G and G corresponding to quantile 0 represents μ_G. Further characterization of 2D memtransitor-based GRNG has been done in our previous report[44]. Here, the random distributions are observed as an effect of random nature of charge trapping which is typically observed in charge-trap memory devices[45,46]. Hence, these memtransistors can also be replaced with standard three-terminal charge-trap flash memories such as TaN-Al_2O_3-Si_3N_4-SiO_2-Si (TANOS) and Si-SiO_2-Si_3N_4-SiO_2-Si (SONOS)[47–49]. Nevertheless, MoS_2 memtransistors are used to generate a physical random variable that samples analog conductance values from a Gaussian distribution, i.e., $G \sim N(\mu_G,\sigma_G)$. Moreover, the MoS_2 memtransistor can be used as a synapse, which scales the input by its synaptic weight. If input is applied as voltage to the drain terminal of the memtransistor, the output current is scaled by G, i.e., $I_{DS} = G.V_{DS}$, as shown in Fig. 3a. Therefore, by combining the cycle-to-cycle variation in G with the synaptic functionality of the memtransistor, we are able to realize a GRNG-based synapse. To implement a BNN accelerator, it's important to tune both μ_G and σ_G of the GRNG-based synapse independently. μ_G and σ_G can be tuned by modulating V_P in the erase-program-read pulse cycle, as shown in Fig. 3e. However, μ_G and σ_G are found to be coupled, and the coefficient of variation ($C_v = \mu_G/\sigma_G$) depends on V_P, as shown in Fig. 3f. A similar trend is seen in μ_G, σ_G and C_v as a function of V_E, as shown in Supplementary Fig. 4.

Figure 4a shows the design of our GRNG-based synapse with independent control over its μ and σ, using two MoS_2 memtransistors, T_+ and T_-. While prior demonstrations rely on additional mathematical manipulations of the generated GRNs to establish control over their μ and σ, we are able to achieve it without any additional manipulations or circuitry[21–23,43]. It is common practice in neural network accelerators to use two devices per synapse in order to map both positive and negative weights[49]. Here, the input to the synapse, V_{in} is applied as $+V_{in}$ and $-V_{in}$ to T_+ and T_-, respectively, as shown in Fig. 4a. The current at the output node (I_{out}) is then given by sum of currents through T_+ and T_- i.e., I_{T_+} and I_{T_-}, according to the Kirchhoff's current law (KCL) given by Eq. 1.

$$I_{out} = I_{T_+} + I_{T_-} = G_+.V_{in} - G_-.V_{in} = (G_+ - G_-).V_{in} = G_{eff}.V_{in} \quad (1)$$

Here G_+ and G_- are the conductance of T_+ and T_- respectively and G_{eff} is the effective conductance of the synapse. While the conductance of a device is always positive, by modulating G_+ and G_-,

https://doi.org/10.1038/s41467-022-33699-7

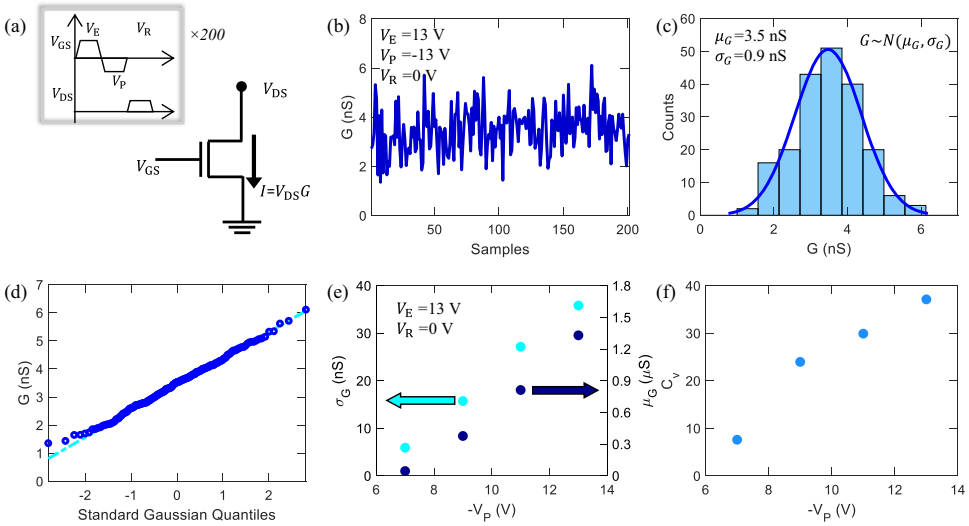


Fig. 3 | Gaussian random number generator (GRNG) using MoS2 memtransistor. a Schematic of a MoS_2 memtransistor used as a GRNG. To generate gaussian random numbers (GRNs), the gate of a MoS_2 memtransistor is subjected to an erase-program-read pulse cycle 200 times with V_E of 13 V, V_P of −13 V, and read voltage (V_R) of 0 V, while the drain is subjected to V_{in} of 0, 0 and 0.1 V, respectively. **b** The corresponding conductance (G) of the MoS_2 memtransistor measured at each read step, demonstrating random fluctuations in G. **c** Histogram demonstrating that G follows a Gaussian distribution, with μ_G = 3.5 nS and σ_G = 0.9 nS. **d** The quantile-quantile (Q-Q) plot of G, confirming its Gaussian distribution. The quantiles of G (represented using circles) are plotted against the theoretical quantiles from a Gaussian distribution, which follows the straight line expected for a Gaussian distribution. **e** Dependence of μ_G and σ_G and the corresponding **f** coefficient of variation (C_v) on V_P. Here, V_P is changed in the erase-program-read pulse cycle. μ_G and σ_G are seen to be coupled.

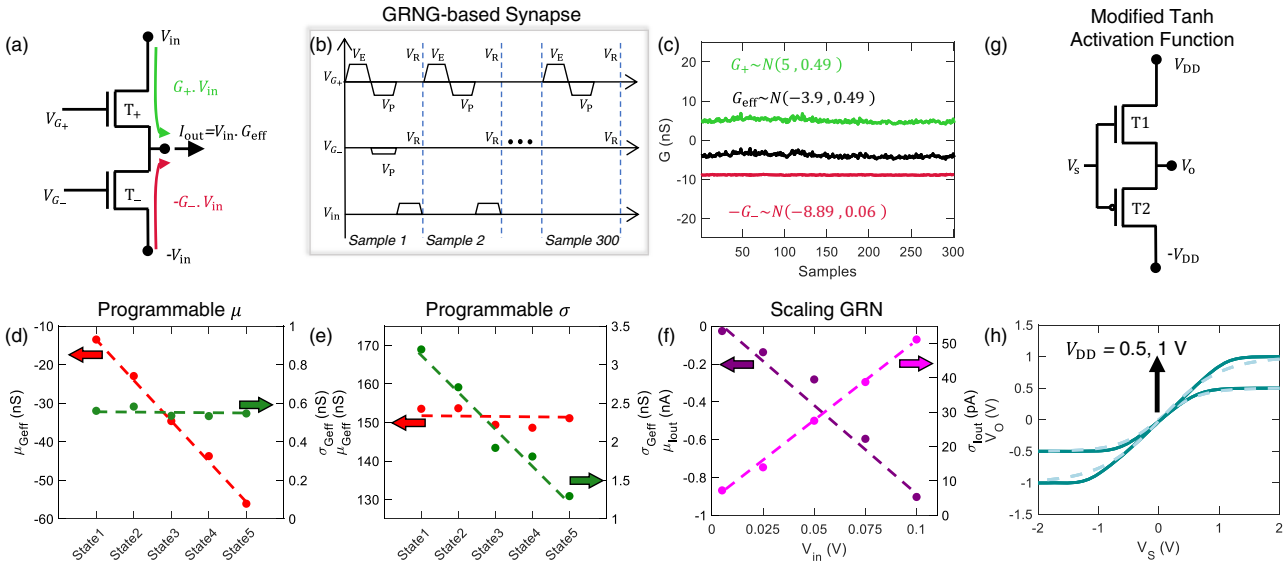


Fig. 4 | GRNG-based synapse and modified tanh activation function. a Schematic of the GRNG-based synapse. The input to the synapse, V_{in} is applied as $+V_{in}$ and $-V_{in}$ to the memtransistors, T_+ and T_- with conductance G_+ and G_- (modulated using V_{G_+} and V_{G_-}), respectively. The effective conductance of this synapse is given by $G_{eff} = G_+ - G_-$, allowing positive and negative conductance. **b** Waveform applied to the synapse to generate GRNs with independent control over μ and σ. **c** G_+, $-G_-$ and G_{eff} of the synapse sampled 300 times, where T_+ controls σ and T_- controls μ. GRNG-based synapse showing **d** independent control of $\mu_{G_{eff}}$ for constant $\sigma_{G_{eff}}$ and **e** independent control of $\sigma_{G_{eff}}$ for constant $\mu_{G_{eff}}$. **f** Linear scaling of the synaptic output (I_{out}) distribution as the function of V_{in}. **g** Schematic of circuit for the modified tanh activation function using a n-type MoS_2 memtransistor (T1) and a V-doped p-type WSe_2 memtransistor (T2), where the input voltage (V_S) is applied to the gate terminal of T1 and T2. **h** The transfer characteristics of the circuit (solid line) i.e., output voltage (V_O) versus V_S, closely models the tanh activation function (dotted line).

using V_{G_+} and V_{G_-} (by applying different V_P), we can obtain both positive and negative G_{eff}. To control $\mu_{G_{eff}}$ and $\sigma_{G_{eff}}$, T_+ is subjected to successive erase-program-read pulse cycles, while T_- is programmed to a given state and subsequently only read, using the waveforms shown in Fig. 4b. This results in G_+ being drawn from a Gaussian distribution, with μ_{G_+} = 5 nS and σ_{G_+} = 0.49 nS i.e., $G_+ \sim N(5, 0.49)$ nS and G_- having a constant value of ≈ 8.89 nS, as shown in Fig. 4c. G_{eff} is expected to be drawn from a distribution with $\sigma_{G_{eff}} = \sigma_{G_+}$ and $\mu_{G_{eff}} = \mu_{G_+} - G_-$. This is confirmed by our measurements as shown in Fig. 4c, $G_{eff} \sim N(-3.9, 0.49)$ nS. Note that, G_- is not perfectly constant due to the presence of random telegraph fluctuations. However, the fluctuations were found to have a standard deviation of 0.06 nS, making its contribution negligible. The histograms and Q-Q plots of G_+, G_-, and G_{eff} are shown in Supplementary Fig. 5. Figure 4d shows the independent control of $\mu_{G_{eff}}$ for constant $\sigma_{G_{eff}}$ using the GRNG-based synapse. Here, T_+ is subjected to the same erase-program-read

cycle, to obtain constant $\sigma_{G_{eff}}$, whereas T_ is programmed to different states (using V_P) to tune $\mu_{G_{eff}}$. Figure 4e shows the independent control of $\sigma_{G_{eff}}$ for constant $\mu_{G_{eff}}$. In order to modulate $\sigma_{G_{eff}}$, σ_{G_+} is changed by applying different erase-program-read cycles (different V_P) to T$_+$. Since this leads to an unfavorable change in μ_{G_+}, T_ is reprogrammed to account for the change in μ_{G_+}, to maintain a constant $\mu_{G_{eff}}$. In a synapse, the distribution of I_{out} is expected to scale linearly with V_{in}, as given by Eq. 2.

$$I_{out} = G_{eff}.V_{in} = N(\mu_{G_{eff}}, \sigma_{G_{eff}}).V_{in} = N(\mu_{G_{eff}}.V_{in}, \sigma_{G_{eff}}.V_{in}) = N(\mu_{I_{out}}, \sigma_{I_{out}}) \quad (2)$$

This is demonstrated in Fig. 4f, where $\mu_{I_{out}}$ and $\sigma_{I_{out}}$ show linear dependence with respect to V_{in}. The output characteristics of a MoS_2 memtransistor is shown in Supplementary Fig. 6 for positive and negative V_{DS}. While the current is highly non-linear and asymmetric for large $\pm V_{DS}$ values, it is seen to be sufficiently linear and symmetric between ± 0.1 V. Hence, we limit the maximum V_{in} to 0.1 V. The low V_{in} allows us to operate the synapse with extremely low currents, as shown in Fig. 4f, offering significant energy efficiency. Overall, we demonstrate independent control over $\mu_{G_{eff}}$ and $\sigma_{G_{eff}}$ to implement a GRNG-based synapse with just two MoS_2 memtransistors resulting in significant area and energy efficiency. Supplementary Fig. 7 demonstrates the stability of analog random number generation for numerous analog states for a total of 20,000 cycles. In addition, it demonstrates that both μ and σ increase as a function of V_{DS} and $|V_P|$, retaining the synaptic behavior for 20,000 cycles. Also, as cycle-to-cycle variation is used for random number generation, Supplementary Fig. 7 also represents that MoS_2 memtransistor has an endurance of at least 20,000 cycles. Higher endurance is desired for BNNs and this can be achieved through further optimization of MoS_2 memtransistors.

Neurons with modified hyperbolic tangent activation function

The hardware for activation function in neural accelerators is generally realized using standard CMOS-based analog and digital components, and hence these implementations do not utilize the advantages offered by emerging materials[49]. Moreover, hyperbolic tangent (tanh) and sigmoid functions are highly non-linear, significantly complicating their hardware demonstration[50]. We demonstrate a circuit for a modified tanh (m-tanh) activation function using a *n*-type MoS_2 memtransistor (T1) and a V-doped *p*-type WSe_2 memtransistor (T2) as shown in Fig. 4g. The transfer function of the circuit i.e., output voltage (V_O) versus input voltage (V_S) closely follows the tanh activation function as shown in Fig. 4h. The maximum of the m-tanh activation function is determined by the drain voltage (V_{DD}). Here, V_S is applied to the gate of T1 and T2. T1 and T2 are programmed to ensure that the m-tanh function passes through the origin. Note that when $V_S = -2$ V, T1 operates in the off-state and T2 operates in the on-state, resulting in $V_O = -V_{DD}$, whereas for $V_S = 2$ V, T1 operates in the on-state and becomes more conductive than T2, which results in $V_O = V_{DD}$. While this output characteristics is well-known[49,51], to the best of our knowledge it has not been used to implement activation functions, as the activation functions are typically implemented using look-up-tables[49]. In addition, modified sigmoid activation function can be realized by applying 0 V to the drain terminal of T1, as shown in Supplementary Fig. 8.

Crossbar array architecture

The crossbar array architecture is routinely used in neural network accelerators to perform the MAC operation of a neuron. Figure 5a shows the circuit used to implement a portion of a BNN shown in Fig. 5b, where $M = 4$ input neurons are connected to $N = 1$ output neuron. Each input neuron is multiplied with their corresponding synaptic weight distributions and the resultants are summed at the output neuron (MAC operation). The resultant of the MAC operation is passed through the m-tanh activation function, to obtain the output. To implement this on circuit, as shown in Fig. 5a, the conductance distribution of a synapse in the i^{th} row, j^{th} column and k^{th} layer ($G_{ij}^{(k)}$), given by the combination of $G_{ij_+}^{(k)}$ and $G_{ij_-}^{(k)}$ is modulated using $V_{Gj_+}^{(k)}$ and $V_{Gj_-}^{(k)}$ lines. Inputs to i^{th} row and k^{th} layer are applied as voltages ($\pm V_i^{(k)}$). The current through the j^{th} column, due to these synapses is then given by the dot product of $V_i^{(k)}$ and $G_{ij}^{(k)}$, according to KCL. To obtain a voltage proportional to this dot product, we use a sense transistor, as shown in Fig. 5a. The voltage-drop ($V_{Sj}^{(k)}$) across this sense transistor is given by Eq. 3.

$$V_{Sj}^{(k)} = \frac{\sum_{i=1}^{M} V_i^{(k)} G_{ij}^{(k)}}{G_{Sj}^{(k)} + \sum_{i=1}^{M} G_{ij}^{(k)}} \quad (3)$$

Here, $G_{Sj}^{(k)}$ is the conductance of the sense transistor, and is modulated using $V_{sense}^{(k)}$. Using $V_{Sj}^{(k)}$ allows us to seamlessly integrate the circuit for m-tanh activation function into the crossbar array as shown in Fig. 5a, to obtain the corresponding output ($V_{Oj}^{(k)}$). There are some non-idealities which are also accounted for. First, the synaptic weight distribution ($W_{ij}^{(k)}$) is mapped to the crossbar array by using a conductance scaling factor (α) to obtained $G_{ij}^{(k)}$. Second, the denominator of $V_{Sj}^{(k)}$ (Eq. 3) presents a non-ideality, which can be expressed as the product of α and a non-ideality factor ($\gamma^{(k)}$). By mapping the input ($x_i^{(k)}$) to $V_i^{(k)}$, using $\gamma^{(k)}$ as the scaling factor, ideal $V_{Sj}^{(k)}$ and $V_{Oj}^{(k)}$ can be obtained as shown in Eq. 4.

$$V_{Sj}^{(k)} = \frac{\sum_{i=1}^{M} V_i^{(k)} G_{ij}^{(k)}}{G_{Sj}^{(k)} + \sum_{i=1}^{M} G_{ij}^{(k)}} = \frac{\sum_{i=1}^{M} \left(x_i^{(k)} \gamma^{(k)}\right) . \left(W_{ij}^{(k)} . \alpha\right)}{\gamma^{(k)} . \alpha} = \sum_{i=1}^{M} x_i^{(k)} W_{ij}^{(k)} \quad (4a)$$

$$V_{Oj}^{(k)} = \tanh\left(V_{Sj}^{(k)}\right) = \tanh\left(\sum_{i=1}^{M} x_i^{(k)} W_{ij}^{(k)}\right) \quad (4b)$$

$G_{Sj}^{(k)}$ is used to make sure that each column of the crossbar array has the same $\gamma^{(k)}$. With this proposed scheme, we can evaluate the dot product between $x_i^{(k)}$ and $W_{ij}^{(k)}$ in the voltage domain and use the m-tanh activation function to obtain the ideal output, $V_{Oj}^{(k)}$. Note that this scheme is not limited to the implementation of a BNN and can be adopted to implement standard ANN crossbar arrays with tanh and sigmoid activation functions.

Neural network evaluation

We evaluate the performance of our BNN implementation using the PIMA Indian Diabetes dataset. This dataset consists of nine parameters such as number of pregnancies, glucose levels, insulin levels, body mass index, age, etc. To classify this dataset, we use a fully connected 8 × 10 × 2 BNN, i.e., it has an input layer with 8 neurons, one hidden layer with 10 neurons, and an output layer with 2 neurons. The dataset with 767 instances is divided into 720 for training and 47 for testing. "Bayes by Backprop" algorithm, with a Gaussian prior is used to train the synaptic weight distributions [12,52] (see the "Methods" for details). The BNN is trained off-chip for 300 epochs as shown in Fig. 5c, to obtain train accuracy of 75.41% and test accuracy of 80.85%. Similar accuracy numbers have been reported in prior works[53–55].

Figure 5d shows the circuit of the BNN used to classify the PIMA Indians dataset. Here, the synapses are arranged in the crossbar array architecture. The BNN circuit to classify PIMA Indian diabetes dataset is evaluated using LTSpice simulations (See "Methods" for details). Note that, at the output layer we do not use the tanh activation function. Instead, following Eq. 6, the output of the BNN, $V_{Sj}^{(k)}$ is sampled $Z = 100$ times to obtain a distribution, and its mean is used to make the classification. Using the LTSpice simulations, we are able replicate the

https://doi.org/10.1038/s41467-022-33699-7

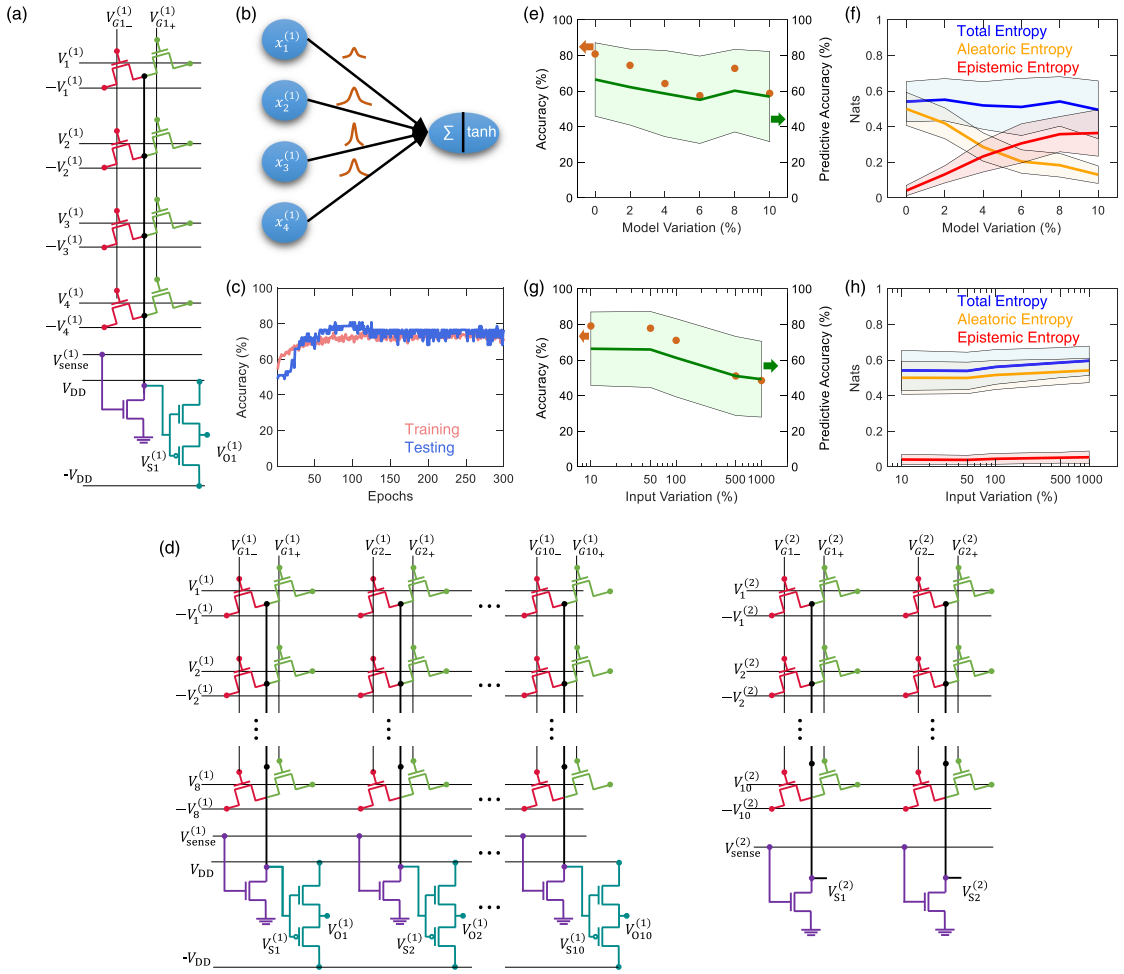


Fig. 5 | Crossbar array architecture to implement the BNN. a Schematic of a portion of a BNN, where multiple input neurons are connected to one output neuron through multiple synapses. **b** The corresponding circuit implementation using the crossbar array architecture. Here, a sense transistor is used to obtain a voltage ($V_{Sj}^{(k)}$) proportional to the dot product between the inputs ($V_i^{(k)}$) and synaptic distributions ($G_{ij}^{(k)}$), where i, j, and k represents the row, column, and layer, respectively. $V_{Sj}^{(k)}$ is applied to the tanh activation circuit to obtain the output ($V_{Oj}^{(k)}$). **c** Training and testing curves for 300 epochs of the BNN constructed to classify the PIMA Indians dataset. **d** Circuit implementation of the BNN to perform inference on-chip. **e** Accuracy and predicative accuracy as function of model variation. Here, the effect of variation in synaptic devices, sense resistors, and activation function is demonstrated. **f** Total entropy, aleatoric entropy and epistemic entropy as a function of model variation. **g** Accuracy and predictive accuracy as a function of input variation. **h** Total entropy, aleatoric entropy and epistemic entropy as a function of input variation.

test accuracy of 80.85%. The BNN circuit consumes a miniscule 18.37 nJ/test sample in order to make the classification (see "Methods" for details).

While the cycle-to-cycle variation in memtransistors is useful for the generation of Gaussian random numbers, the device-to-device variation observed along with it is undesirable for the operation of a neural network. Supplementary Note 1 and Supplementary Fig. 9 discusses a method to model and reduce the device-to-device variation. Since, some variation is still expected to exist, we evaluate its effect on the performance of the circuit for the Bayesian neural network. We implement up to 10 % variation for the parameters of the synapse: μ_{G_+}, σ_{G_+}, and G_-, which results in a corresponding variation in G_{eff}, the parameters of the tanh function: $V_{\text{TH},n}$ and $V_{\text{TH},p}$, and for the n-type FET and the p-type FET, and the conductance of the sense resistor G_S. These parameters are drawn from Gaussian distributions where the mean is given by expected parameter value and standard deviation of up to 10% is considered. Figure 5e shows the effect of device-to-device (model) variation on the test accuracy and predictive accuracy. The predictive accuracy demonstrates how well the predictions of the expected correct classes are made. Here, the BNN is simulated and averaged over 5 runs. While, we observe a decrease in the test accuracy, it is not seen to significantly impact the operation of the BNN and an accuracy of ≈60% is maintained for 10% variation.

In a BNN, we can use entropy estimation and entropy decomposition to quantify uncertainty and to find its source, respectively (see the "Methods" section for details). The distribution of $V_{Sj}^{(k)}$ at the output layer can be used to calculate the uncertainty in classification[56,57]. Figure 5f shows the total entropy, aleatoric entropy, and epistemic entropy as a function of model variation, calculated using Eqs. 8 and 9. In model variation, it was determined that the variation in synapses is more determinantal to the performance of BNN compared to variations in the neuron, due to the binary nature of our classification problem. For a multiclass problem, variations in the neuron are expected to have a larger impact on the accuracy. In Fig. 5f, we would expect the aleatoric uncertainty to remain unchanged and total entropy to increase. However, their extraction is impacted by the increased model variation. Nevertheless, as expected, the epistemic entropy increases as the model variation increases. As shown in Fig. 5g, h, an increase in the input variation results in the degradation of accuracy, along with an increase in the total and aleatoric entropy, while a constant epistemic entropy is maintained, as expected. Hence, in addition to the estimation of total entropy

using a BNN, with uncertainty decomposition various sources of entropy can be identified.

Discussion

This work demonstrates the development of computational primitives needed for a BNN accelerator, using 2D memtransistors. The cycle-to-cycle variation in the programming of the MoS_2 memtransistor is exploited as a source of randomness and a circuit comprising of two such memtransistors is used to obtain an ultra-low-power and stochastic synapse, which allows sampling of both positive and negative weights from a Gaussian distribution with reconfigurable mean and standard deviation. We also developed circuits to implement the modified hyperbolic tangent and sigmoid activation functions based on the integration of MoS_2 and WSe_2 memtransistors. In addition, we integrate these components into a crossbar array architecture to perform efficient MAC operations and we develop a BNN circuit to perform on-chip inference to classify the PIMA Indians dataset and evaluate its performance using circuit simulations. Finally, we have also performed uncertainty decomposition in order to identify the various sources of uncertainty.

Methods

Device fabrication

MoS_2 memtransistors are fabricated using photolithography and e-beam lithography. Photo-lithography is used to define the back-gate islands. A p^{++} Si substrate is first spin coated with LOR 5 A and baked at 180 °C for 120 s, and subsequently spin coated with SPR 3012 and baked at 95 °C for 60 s. Using Heidelberg MLA 150, the desired regions are exposed to 405 nm light. The exposed regions are developed using 1:1 CD 26 and DI water. To form the back gate islands, 20 nm TiN followed by 50 nm Pt is deposited through sputtering. 50 nm Al_2O_3 gate dielectric is deposited using atomic layer deposition. Al_2O_3 is etched from back gate contact regions using BCl_3 etch, where the etch region was defined by photolithography. Following this MOCVD MoS_2 is transferred onto this substrate and the MoS_2 transistors are fabricated as discussed in our previous reports[28,51]. The fabrication of WSe_2 memtransistors have been discussed in our previous report[30].

Electrical characterization

Lake Shore CRX-VF probe station and Keysight B1500A parameter analyzer were used to perform the electrical characterization at room temperature. The device-to-device variation measurements were performed using the FormFactor Cascade Summit 12000 semi-automated probe station.

BNN training and inference

According to the Bayes theorem, the weights of a BNN are given by posterior probability distributions given by Eq. 5.

$$p(W|D) = \frac{p(D|W).\,p(W)}{p(D)} \tag{5}$$

Here D is the training data, $p(W|D)$ is the posterior distribution, $p(D|W)$ is the likelihood, $p(W)$ is the prior, and $p(D)$ is the evidence. The true posterior distribution is untraceable in BNNs and hence, methods such as variational inference[12] and Markov chain Monte Carlo (MCMC) sampling[58] are used to approximate the posterior distribution. Variational inference is typically preferred due to better convergence and scalability compared to MCMC[20]. In the variational inference method, $p(D|W)$ is estimated using a family of variational posterior distributions (typically Gaussian distributions), $q(W;\theta)$, where θ represents the variation parameters. For a Gaussian distribution, the variation parameters are its mean and standard deviation, i.e., $\theta = \mu,\sigma$. The estimation is performed by minimizing the Kullback-Leibler divergence between $p(D|W)$ and $q(W;\theta)$. In the training phase, for each synapse, μ and σ are learned using the traditional backpropagation method[12]. Here, σ represents the uncertainty introduced by each synapse. To perform inference using a BNN, multiple forward passes of the trained network is evaluated. During each forward pass, each of the Gaussian weight distributions are sampled once. The output of the network or the predictive distribution is obtained by averaging the outputs of these forward passes obtained by sampling the weight distributions. It can be approximated by drawing Z Monte Carlo samples and finding its mean given by Eq. 6.

$$p(y^*|x^*,D) \approx \frac{1}{Z}\sum_{z=1}^{Z} p\left(y^*|x^*,W^z\right) \tag{6}$$

Here, x^* and y^* are the test input and output, respectively, D is the training data, and W^z represents the z^{th} Monte Carlo weight sample.

LTSpice Circuit Implementation

The synapses consisting of two memtransistors, T_+ and T_- are modeled using resistors. During the read step, V_R or of 0 V is applied to prevent read-disturb and V_{in} is applied such that the transistors are operating in the linear part of the triode region. Hence, these memtransistors are modeled using resistors. Here, T_+ is modeled as a resistor which draws from Gaussian conductance distribution given by $G_+ \sim N(\mu_{G_+},\sigma_{G_+})$ and T_- is modeled as a resistor set to a constant conductance of G_-. The weight distributions are mapped to conductance distributions with $\alpha = 10^{-9}$. For the actual circuit implementation using memtransistors, in order to program and reprogram the crossbar array, we propose to use a scheme where all devices are erased with a high V_E of 13 V. Following that, each device must be programmed with the separate V_P values to set them to their expected states, requiring additional storge to store V_P values. The sense transistor is modeled using a resistor, ensuring that each column of the crossbar array has the same $\gamma^{(k)}$. $\gamma^{(k)}$ is determined for each layer and multiplied with $x_i^{(k)}$ to obtain $V_i^{(k)}$. The tanh activation function is implemented using the combination of an n-type FET and a p-type FET. V_{DD} of 1 V is used for the m-tanh activation circuit.

Energy analysis

The total energy consumption of the BNN accelerator is given by Eq. 7.

$$E_{\text{Total}} = E_{\text{syn}} + E_{\text{sense}} + E_{\text{tanh}} + E_P + E_E$$

$$= \sum_N \left[\left(I_{T_+} + I_{T_-}\right)V_{in}t\right]_{\text{syn}} + \left[I_S V_S t\right]_{\text{sense}} + \left[\left(I_{T1} + I_{T2}\right)V_{DD}t\right]_{\text{tanh}} + E_P + E_E \tag{7a}$$

$$E_P + E_E = \sum_{N-100} t\left(I_P V_P + I_E V_E\right) + \sum_{100}\left[Qt(I_P V_P + I_E V_E)\right]_{P,E} \tag{7b}$$

Here, E_{Total} represents the total energy consumption per test sample. E_{syn}, E_{sense}, and E_{tanh} represents the total energy consumption by the synapses, sense resistors, and the tanh circuit, respectively, during the inference step. I_S represents the current through the sense resistor and N is the total number of devices. E_P and E_E represents the energy consumption for programming and erasing operation on these components. I_P and I_E respectively, are the gate currents associated with programming and erase operations. Here, only half of the synaptic devices (100 devices) needs to be erased and programmed for $Z = 100$ Monte Carlo samples. The rest of the synaptic devices, sense transistors, and tanh circuit components are erased and programmed once and subsequently only used for inference. For E_P and E_E evaluations, maximum I_P, V_P, I_E, V_E values expected are used. Using Eq. 7, E_{syn} of 18 nJ, E_{sense} of 0.025 nJ, and E_{tanh} of 0.012 nJ are obtained from LTSpice simulations and $E_P + E_E$ of 0.34 nJ is estimated for a total of $Z = 100$

Monte Carlo samples, resulting in an E_{Total} of 18.37 nJ. Note that while the program/erase voltages are high compared to other emerging memory technologies such as memristors and phase-change memory, the energy consumption associated with memtransistors is low as the program (gate) currents are on the order of pA and since the transistors are biased with a drain voltage of 0 V during programming/erasing, there is no drain current (See Supplementary Table 1 for energy analysis of various emerging memories)[59].

Entropy calculation and decomposition

The softmax of predictive distribution can be used to calculate the uncertainty in classification or entropy given by Eq. 8.[56,57].

$$H\left(p(\hat{y}^*|x^*,D)\right) = -\sum_{j=1}^{J} p_j\left(\hat{y}^*|x^*,D\right)^* \log\left(p_j\left(\hat{y}^*|x^*,D\right)\right) \quad (8)$$

Here, $\hat{y}^*$ is the softmax output and J is the number of output classes. The entropy can be decomposed into epistemic entropy, i.e., uncertainty in model and aleatoric entropy, i.e., uncertainty in data as shown in Eq. 9.

$$H\left(p(\hat{y}^*|x^*,D)\right) = \prod\left(p\left(\hat{y}^*|x^*,D\right)\right) + E_{W\sim q(W;\theta)}\left[H\left(p(\hat{y}^*|x^*,W)\right)\right] \quad (9)$$

Here, on the right-hand side, the first term represents epistemic entropy and the second term represents aleatoric entropy. Aleatoric entropy is the average entropy for fixed weights and hence the uncertainty arises from the data. Epistemic entropy can be obtained by subtracting aleatoric entropy from total entropy, following Eq. 9.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding authors on reasonable request.

Code availability

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

References

1. Chauhan, R., Ghanshala, K. K. & Joshi, R. C. Convolutional neural network (CNN) for image detection and recognition. In *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)* pp. 278–282 (IEEE, 2018).
2. Fente, D. N. & Kumar Singh, D. Weather forecasting using artificial neural network. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)* pp. 1757–1761 (IEEE, 2018).
3. Guresen, E., Kayakutlu, G. & Daim, T. U. Using artificial neural network models in stock market index prediction. *Expert Syst. Appl.* **38**, 10389–10397 (2011). pp.
4. Le, T. H. Applying artificial neural networks for face recognition. *Adv. Artif. Neural Syst.* **2011**, 1–16 (2011).
5. Nirkhi, S. Potential use of artificial neural network in data mining. In *The 2nd International Conference on Computer and Automation Engineering (ICCAE)* pp. 339–343 (IEEE, 2010).
6. Nusrat, I. & Jang, S.-B. A comparison of regularization techniques in deep neural networks. *Symmetry* **10**, 648 (2018).
7. Ding, L. et al. Capture Uncertainties in Deep Neural Networks for Safe Operation of Autonomous Driving Vehicles, *2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*, pp. 826–835 (2021).
8. Chakraborty, S. & Ghosh, M. Applications of Bayesian Neural Networks in Prostate Cancer Study. **28**, *Handbook Stat*. 241–262 (2012).
9. Ticknor, J. L. A Bayesian regularized artificial neural network for stock market forecasting. *Expert Syst. Appl.* **40**, 5501–5506 (2013).
10. Shi, J., Zhu, Y., Khan, F. & Chen, G. Application of Bayesian Regularization Artificial Neural Network in explosion risk analysis of fixed offshore platform. *J. Loss Prev. Process Industries* **57**, 131–141 (2019).
11. Kayri, M. Predictive Abilities of Bayesian Regularization and Levenberg–Marquardt Algorithms in Artificial Neural Networks: A Comparative Empirical Study on Social Data. *Math. Comput. Appl.* **21**, 20 (2016).
12. Blundell, C., Cornebise, J., Kavukcuoglu, K. & Wierstra, D. Weight uncertainty in neural networks. *ArXiv*, abs/1505.05424 (2015).
13. MacKay, D. J. C. A practical Bayesian framework for backpropagation networks. *Neural Comput.* **4**, 448–472 (1992).
14. Gal, Y. & Ghahramani, Z. Bayesian convolutional neural networks with bernoulli approximate variational inference. *ArXiv* abs/1506.02158 (2015).
15. Chen, W. et al. Quantization of deep neural networks for accurate edge computing. *ACM J. Emerg. Technol. Comput. Syst.* **17**, 1–11 (2021).
16. Marchisio, A. et al. Deep learning for edge computing: current trends, cross-layer optimizations, and open research challenges. *2019 IEEE Computer Society Annual Symposium on VLSI (IVLSI)*, pp. 553–559 (2019).
17. Ardakani, A., Condo, C. & Gross, W. J. Fast and efficient convolutional accelerator for edge computing. *IEEE Trans. Computers* **69**, 138–152 (2020).
18. Hirayama, Y., Asai, T., Motomura, M. & Takamaeda, S. A hardware-efficient weight sampling circuit for Bayesian neural networks. *2020* **10**, 10 (2020).
19. Cai, R., Ren, A., Wangy, L., Pedramy, M. & Wang, Y. Hardware Acceleration of Bayesian Neural Networks Using RAM Based Linear Feedback Gaussian Random Number Generators. In *2017 IEEE International Conference on Computer Design (ICCD)* pp. 289–296 (IEEE, 2017).
20. Cai, R. et al. VIBNN: Hardware Acceleration of Bayesian Neural Networks. In *Proc. Twenty-Third International Conference on Architectural Support for Programming Languages and Operating Systems* (2018).
21. Malhotra, A., Lu, S., Yang, K. & Sengupta, A. Exploiting oxide based resistive RAM variability for Bayesian neural network hardware design. *IEEE Trans. Nanotechnol.* **19**, 328–331 (2020).
22. Dalgaty, T., Esmanhotto, E., Castellani, N., Querlioz, D. & Vianello, E. Ex situ transfer of bayesian neural networks to resistive memory-based inference hardware. *Adv. Intell. Syst.* **3**, 2000103 (2021).
23. Yang, K., Malhotra, A., Lu, S. & Sengupta, A. All-spin Bayesian neural networks. *IEEE Trans. Electron Devices* **67**, 1340–1347 (2020).
24. Pendurthi, R. et al. Heterogeneous integration of atomically thin semiconductors for non-von Neumann CMOS. *Small* 2202590 (2022).
25. Dodda, A., Trainor, N., Redwing, J. & Das, S. All-in-one, bio-inspired, and low-power crypto engines for near-sensor security based on two-dimensional memtransistors. *Nat. Commun.* **13**, 1–12 (2022).
26. Zheng, Y. et al. Hardware implementation of Bayesian network based on two-dimensional memtransistors. *Nat. Commun.* **13**, 5578 (2022).
27. Radhakrishnan, S. S. et al. A sparse and spike-timing-based adaptive photo encoder for augmenting machine vision for spiking neural networks. *Adv. Mater.* 202535 (2022).
28. Sebastian, A., Pendurthi, R., Choudhury, T. H., Redwing, J. M. & Das, S. Benchmarking monolayer MoS_2 and WS_2 field-effect transistors. *Nat. Commun.* **12**, 693 (2021).
29. Dodda, A. et al. Stochastic resonance in MoS_2 photodetector. *Nat. Commun.* **11**, 4406 (2020).

30. Kozhakhmetov, A. et al. Controllable p-type doping of 2D WSe2 via vanadium substitution. *Adv. Funct. Mater.* **31**, 2105252 (2021).
31. Das, S. et al. Transistors based on two-dimensional materials for future integrated circuits. *Nat. Electron.* **4**, 786–799 (2021).
32. Zhu, K. et al. The development of integrated circuits based on two-dimensional materials. *Nat. Electron.* **4**, 775–785 (2021).
33. Sebastian, A. et al. Gaussian synapses for probabilistic neural networks. *Nat. Commun.* **10**, 4199 (2019).
34. Subbulakshmi Radhakrishnan, S., Sebastian, A., Oberoi, A., Das, S. & Das, S. A biomimetic neural encoder for spiking neural network. *Nat. Commun.* **12**, 2143 (2021).
35. Jayachandran, D. et al. A low-power biomimetic collision detector based on an in-memory molybdenum disulfide photodetector. *Nat. Electron.* **3**, 646–655 (2020).
36. Das, S., Dodda, A. & Das, S. A biomimetic 2D transistor for audiomorphic computing. *Nat. Commun.* **10**, 3450 (2019).
37. Arnold, A. J. et al. Mimicking neurotransmitter release in chemical synapses via hysteresis engineering in MoS2 transistors. *ACS nano* **11**, 3110–3118 (2017).
38. Ko, T. J. et al. Two-dimensional near-atom-thickness materials for emerging neuromorphic devices and applications. *iScience* **23**, 101676 (2020).
39. Kwon, K. C., Baek, J. H., Hong, K., Kim, S. Y. & Jang, H. W. Memristive devices based on two-dimensional transition metal chalcogenides for neuromorphic computing. *Nanomicro Lett.* **14**, 58 (2022).
40. Bian, J., Cao, Z. & Zhou, P. Neuromorphic computing: devices, hardware, and system application facilitated by two-dimensional materials. *Appl. Phys. Rev.* **8**, 041313 (2021).
41. Mennel, L. et al. Ultrafast machine vision with 2D material neural network image sensors. *Nature* **579**, 62–66 (2020).
42. Wu, P. et al. Next-generation machine vision systems incorporating two-dimensional materials: progress and perspectives. *InfoMat*, **4**, e12275 (2021).
43. Dalgaty, T. et al. In situ learning using intrinsic memristor variability via Markov chain Monte Carlo sampling. *Nat. Electron.* **4**, 151–161 (2021).
44. Wali, A., Ravichandran, H. & Das, S. A machine learning attack resilient true random number generator based on stochastic programming of atomically thin transistors. *ACS Nano* **15**, 17804–17812 (2021).
45. Danial, L., Gupta, V. Pikhay, E. Roizin, Y. & Kvatinsky, S. Modeling a floating-gate memristive device for computer aided design of neuromorphic computing. pp. 472–477, (2020).
46. Monzio Compagnoni, C., Gusmeroli, R., Spinelli, A. S. & Visconti, A. Analytical model for the electron-injection statistics during programming of nanoscale nand flash memories. *IEEE Trans. Electron Devices* **55**, 3192–3199 (2008).
47. Han, J. K. et al. Cointegration of single-transistor neurons and synapses by nanoscale CMOS fabrication for highly scalable neuromorphic hardware. *Sci. Adv.* **7**, eabg8836 (2021).
48. Hwang, S. et al. Capacitor-based synaptic devices for hardware spiking neural networks. *IEEE Electron Device Lett.* **43**, 549–552 (2022).
49. Xiao, T. P., Bennett, C. H., Feinberg, B., Agarwal, S. & Marinella, M. J. Analog architectures for neural network acceleration based on non-volatile memory. *Appl. Phys. Rev.* **7**, 031301 (2020).
50. Chang, C.-H., Kao, H.-Y. & Huang, S.-H. Hardware implementation for multiple activation functions. pp. 1–2 (2019).
51. Sebastian, A., Das, S. & Das, S. An annealing accelerator for ising spin systems based on in-memory complementary 2D FETs. *Adv. Mater.* **34**, e2107076 (2022).
52. Antoran, J. Bayesian-Neural-Networks. https://github.com/JavierAntoran/Bayesian-Neural-Networks (2019).
53. Zou, Q. et al. Predicting diabetes mellitus with machine learning techniques. *Front. Genet.* **9**, 515 (2018).
54. Vaishali, R., Sasikala, R. Ramasubbareddy, S. Remya, S. & Nalluri, S. Genetic algorithm based feature selection and MOE Fuzzy classification algorithm on Pima Indians Diabetes dataset. pp. 1–5 (2017).
55. Zolfaghari. R. Diagnosis of Diabetes in Female Population of Pima Indian Heritage with Ensemble of BP Neural Network and SVM. *IJCEM International Journal of Computational Engineering & Management* **15**, 2230–7893 (2012).
56. Kwon, Y., Won, J.-H., Kim, B. J. & Paik, M. C. Uncertainty quantification using Bayesian neural networks in classification: application to biomedical image segmentation. *Comput. Stat. Data Anal.* **142**, 106816 (2020).
57. Kendall, A. & Gal, Y. What uncertainties do we need in Bayesian deep learning for computer vision?. In *Proc. 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA, 2017).
58. Andrieu, C., de Freitas, N., Doucet, A. & Jordan, M. I. An introduction to MCMC for machine learning. *Mach. Learn.* **50**, 5–43 (2003).
59. Kavehei, O. Memristive devices and circuits for computing, memory, and neuromorphic applications. Thesis (Ph.D.) University of Adelaide, School of Electrical and Electronic Engineering. (2012).

Acknowledgements

The work was supported by Army Research Office (ARO) through Contract Number W911NF1920338. Authors also acknowledge Mr. Shiva Subbulakshmi Radhakrishnan for help with m-tanh measurements. 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-2039351. A.K. and J.A.R. acknowledge funding support from Intel, Inc. through the Semiconductor Research Corporation Task 2746.

Author contributions

A.S. conceived the idea and designed the experiments. A.S. performed the experiments, analyzed the data. R.P. performed the fabrication and characterization of WSe_2 memtransistors. A.K. and N.T. synthesized the 2D materials under the supervision of J.A.R. and J.M.R., respectively. A.S. and S.D. discussed the results, agreed on their implications. All authors contributed to the preparation of the manuscript.

Competing interests

The authors declare no competing interests

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-33699-7.

Correspondence and requests for materials should be addressed to Amritanand Sebastian or Saptarshi Das.

Peer review information *Nature Communications* thanks the anonymous reviewer for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022

nature communications

Article

https://doi.org/10.1038/s41467-022-33053-x

Hardware implementation of Bayesian network based on two-dimensional memtransistors

Received: 13 January 2022

Accepted: 31 August 2022

Published online: 23 September 2022

Check for updates

Yikai Zheng[1], Harikrishnan Ravichandran[1], Thomas F. Schranghamer[1], Nicholas Trainor[2,3], Joan M. Redwing[2,3] & Saptarshi Das[1,2,3,4]

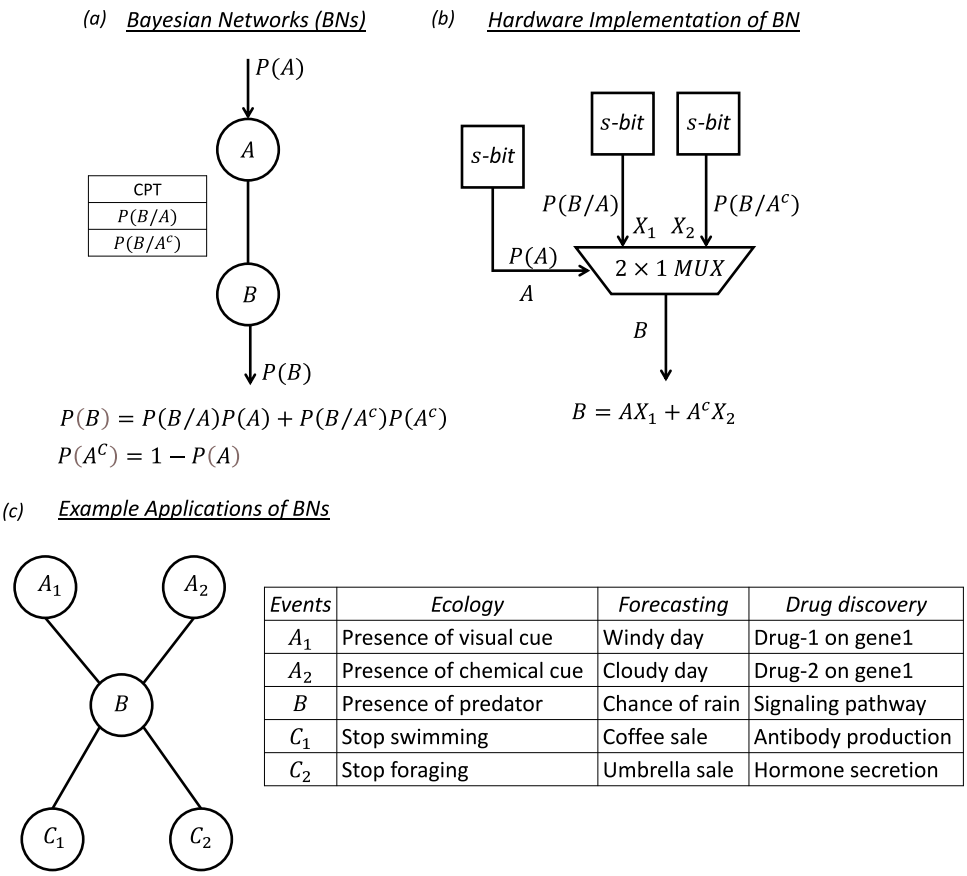
Bayesian networks (BNs) find widespread application in many real-world probabilistic problems including diagnostics, forecasting, computer vision, etc. The basic computing primitive for BNs is a stochastic bit (s-bit) generator that can control the probability of obtaining '1' in a binary bit-stream. While silicon-based complementary metal-oxide-semiconductor (CMOS) technology can be used for hardware implementation of BNs, the lack of inherent stochasticity makes it area and energy inefficient. On the other hand, memristors and spintronic devices offer inherent stochasticity but lack computing ability beyond simple vector matrix multiplication due to their two-terminal nature and rely on extensive CMOS peripherals for BN implementation, which limits area and energy efficiency. Here, we circumvent these challenges by introducing a hardware platform based on 2D memtransistors. First, we experimentally demonstrate a low-power and compact s-bit generator circuit that exploits cycle-to-cycle fluctuation in the post-programmed conductance state of 2D memtransistors. Next, the s-bit generators are monolithically integrated with 2D memtransistor-based logic gates to implement BNs. Our findings highlight the potential for 2D memtransistor-based integrated circuits for non-von Neumann computing applications.

The concept of a Bayesian network (BN) is deep rooted within natural intelligence. Animals gather information from their surroundings with the help of their sensory organs and process this information using their brain to make decisions, enabling their survival. However, gathering accurate information is often very difficult in practice either due to the limitations of sensory organs or due to noisy environment. For example, visual cues are an unreliable source of information for freshwater fish like the rainbow trout to identify the presence of a predator. In contrast, chemical cues released into the water from an injured fish are more reliable indicators of a predatory event[1]. The decision to invoke an alarm response, therefore, depends on how the brain processes the visual and chemical cues based on their relative probability of success from prior experiences. While the neural basis of such computations is relatively unknown, the mathematical construct is represented using a BN with theoretical foundation in Bayes' theorem.

A BN is a probabilistic graphic network used to estimate and infer the probability of interdependent events[2]. Figure 1a shows the basic building block of a BN, comprising a parent node, A, a child node, B, and an edge connecting the two. Each node represents an event, e.g., the presence of a chemical cue (A) and the presence of a predator (B), and the connection represents how two events are mutually dependent. The dependence is provided in a conditional probability table (CPT) which contains the conditional probability (likelihood) values $P(B/A)$ and $P(B/A^c)$, where A^c is the complement of the event A. In the present example, these represent the likelihood of the presence of a

[1]Engineering Science and Mechanics, Penn State University, University Park 16802 PA, USA. [2]Materials Science and Engineering, Penn State University, University Park 16802 PA, USA. [3]Materials Research Institute, Penn State University, University Park 16802 PA, USA. [4]Electrical Engineering and Computer Science, Penn State University, University Park 16802 PA, USA. e-mail: sud70@psu.edu

https://doi.org/10.1038/s41467-022-33053-x



Events	Ecology	Forecasting	Drug discovery
A_1	Presence of visual cue	Windy day	Drug-1 on gene1
A_2	Presence of chemical cue	Cloudy day	Drug-2 on gene1
B	Presence of predator	Chance of rain	Signaling pathway
C_1	Stop swimming	Coffee sale	Antibody production
C_2	Stop foraging	Umbrella sale	Hormone secretion

Fig. 1 | Bayesian networks (BNs). a Schematic of the basic building block of a BN, comprising a parent node, A, a child node, B, and an edge connecting the two. Each node represents an event, and the connection represents how two events are mutually dependent. The dependence is provided in a conditional probability table (CPT), which contains the conditional probability (likelihood) values $P(B/A)$ and $P(B/A^c)$, where A^c is the complement of the event. Knowing the probability of occurrence for event A, i.e., $P(A)$, the marginal probability of occurrence of event B, i.e., $P(B)$, can be evaluated using Bayes' theorem. **b** Hardware implementation of the 2-node BN in (**a**) using three stochastic bit (s-bit) generators and one 2×1 multiplexer (MUX) circuit. **c** Examples of BN architecture that represent real-life situations from ecology to forecasting and drug discovery, highlighting its usefulness in decision making.

predator when a chemical cue is present (A) or absent (A^c), respectively. When the probability of occurrence for event A, i.e., $P(A)$, is known, the marginal probability of occurrence of event B, i.e., $P(B)$, can be evaluated using Bayes' theorem following Eq. 1.

$$P(B) = P(B/A)P(A) + P(B/A^c)P(A^c) = P(B/A)P(A) + P(B/A^c)[1 - P(A)] \quad (1)$$

$$P(A) + P\left(A^C\right) = 1 \quad (2)$$

In a generic BN, a child node can have multiple parent nodes, and a parent node can have multiple children. For example, Supplementary Fig. S1a shows a BN where the child node, B, is connected to 2 parent nodes, A_1 and A_2. Note that the CPT in this instance contains $N = 4$ entries, which are the conditional probability (likelihood) for the occurrence of event B under all possible combinations of the occurrence of events A_1 and A_2. Similarly, Supplementary Fig. S1b shows a BN where the parent node, A, is connected to 2 children, B_1 and B_2. In this case, there are 2 CPTs with $N = 2$ entries each.

Note that the probability estimation for a child node requires multiple arithmetic operations such as multiplication, subtraction, and addition. This makes hardware implementation of a BN using conventional silicon complementary metal-oxide-semiconductor (CMOS) technology[3, 4] less attractive because 1) arithmetic operations require circuits consisting of hundreds of transistors, which have large footprints and consume a significant amount of energy, and 2) the von Neumann bottleneck necessitates storing of the CPT in the memory, which is physically separated from the arithmetic core and therefore requires frequent data shuttling between the two, further aggravating the energy burden. In contrast, even the tiniest brains with very limited numbers of neurons can perform such apparently complex computational tasks with miniscule energy expenditure. The success of biological brains in implementing BNs could lie in the inherently stochastic nature of neural computation.

Drawing inspiration from biology, stochastic computing (SC) has been explored for the hardware implementation of BNs[5]. The key difference from classical computing, where information in presented in the form of binary values (1's and 0's), is that SC encodes information using stochastic bits (s-bits) that are interpreted as probabilities that fall in the interval [0,1]. For instance, the bit-stream S = [1 0 0 1 0 1 0 0] encodes the value $P(S) = 3/8$, i.e., the probability of finding '1' in the bit-stream S. An attractive feature of SC is that arithmetic operations can be performed using simple logic gates[6, 7]. For example, the 2-node BN in Fig. 1a can be realized using a multiplexer (MUX) circuit as shown in Fig. 1b. The output, B, of a MUX with two input variables, X_1 and X_2, and a select line, A, is given by Eq. (3).

$$B = AX_1 + A^cX_2 \quad (3)$$

If, instead of being digital variables, X_1, X_2, and A represent stochastic variables with $P(X_1)$, $P(X_2)$, and $P(A)$ being the respective probability of obtaining '1' in their bit-streams, then B also transforms into a

random variable whose probability is given by Eq. (4).

$$P(B) = P(A)P(X_1) + P(A^c)P(X_2) \quad (4)$$

Note that, if $P(X_1) = P(B/A)$ and $P(X_2) = P\left(B/A^C\right)$, then Eq. (4) transforms into Eq. (1). Therefore, hardware implementation of a child node with a single parent can be accomplished by using 3 s-bit generators and a 2 × 1 *MUX*. Interestingly, the *MUX* architecture can be scaled to implement any BN. For example, hardware implementation of the BN in Fig. 1a can be achieved by using 2 s-bit generators to obtain A_1 and A_2, another 4 s-bit generators to obtain the CPT, and one 4 × 1 *MUX* with 2 select lines as shown in Supplementary Fig. S1c. Similarly, Supplementary Fig. 1d shows the hardware architecture for the BN in Supplementary Fig. S1b, consisting of 1 s-bit generator to obtain *A*, another 4 s-bit generators to obtain the 2 CPTs, and 2 2 × 1 *MUX*s.

Note that BN architecture can be used to represent many real-life situations, as shown in Fig. 1c. For example, in the case of the rainbow trout, events A_1 and A_2 represent the presence of independent visual and chemical cues and event *B* represents the presence of a predator. Events C_1 and C_2, meanwhile, represent the decision taken by the rainbow trout to stop swimming and stop foraging, respectively, which are also independent of each other but depend on *B*. Similarly, in forecasting, events A_1 and A_2 represent the probability of a day being cloudy and windy, respectively, event *B* represents the probability of rain, and events C_1 and C_2 may represent the decision to purchase an umbrella or drink coffee, respectively. Finally, a third example is derived from genetics and drug discovery, where events A_1 and A_2 may represent the probability of expressing *gene* 1 and *gene* 2 when intervening with a specific drug, respectively, event *B* represents the activation of a critical signaling pathway, and events C_1 and C_2 represent production of specific hormones or antibodies, respectively. The above discussion exemplifies the usefulness of BNs in depicting causal relationships using acyclic graphs, which can subsequently be used to predict outcomes based on prior knowledge and likelihood. For example, to predict the relative effectiveness between drug-1 and drug-2 that influence expression for *gene* 1 and *gene* 2, respectively, the only experiments that one needs to do is to obtain respective prior results, i.e., $P(A_1)$ and $P(A_2)$. A BN can then be used to obtain marginal likelihoods, i.e., $P(C_1)$ and/or $P(C_2)$, to assess the relative effectiveness of the two drugs.

The fundamental computing primitive for the stochastic computing implementation of a BN is an s-bit generator, which allows control of the output probability of obtaining '1' in a given bit-stream. So far, probabilistic CMOS[8], field-programmable gate arrays (FPGAs)[9–11], memristors[12–14], and spintronic devices[15–21] have been successfully used for BN implementation. However, CMOS- and FPGA-based BN architectures require hundreds of transistors to generate s-bits, which limits their area and energy efficiency[22–27]. In contrast, memristors offer inherent stochasticity in their switching dynamics, which can be exploited to obtain random bits. However, memristor-based BN architectures heavily rely on CMOS peripherals to translate random bits into s-bits and for subsequent logic operations using those s-bits. Recently, spintronic devices such as magnetic random access memory (MRAM)[28] and magnetic tunnel junctions (MTJs)[29–31] have shown potential for BN implementation since s-bits can be obtained by controlling the probability of spin-flip through externally driven current. However, temperature and supply voltage fluctuations can impact the spin-flip probability, which necessitates additional CMOS-based peripheral circuits to remove the bit-bias. In addition, spin-based devices still require CMOS-based logic circuits for BN implementation.

In this work, we demonstrate hardware implementation of a BN using a monolithic memtransistor technology based on two-dimensional (2D) semiconductors such as monolayer MoS_2. Memtransistors are three-terminal devices in which the gate terminal allows non-volatile and analog programming of the conductance states, which can then be readout by applying a source-to-drain bias. Our main contributions in this work are 1) the design of an area and energy efficient s-bit generator circuit composed of six memtransistors, allowing it to achieve a tunable probability of obtaining '1' in the bit-stream over the range [0,1], and 2) integration of s-bit generators with a 2D memtransistor-based 2×1 *MUX* that consists of three *NAND* gates and one *NOT* gate for BN implementation. In brief, we exploit the inherent stochasticity of the charge trapping and detrapping processes in the gate dielectric of the memtransistor as the source of randomness. Our in-memory computing approach based on three-terminal 2D memtransistors not only overcomes the von Neumann limitations of conventional digital CMOS, but also eliminates the need for peripherals, which is inescapable for emerging memristor- and spin-based 2-terminal stochastic devices for BN implementation.

Our choice of monolayer MoS_2 is motivated by the fact that atomically thin 2D materials are being considered for advanced technology nodes[32]. It is widely accepted that scaling silicon thickness beyond ~3–4 nm is challenging. Yet, the gate electrostatics demand aggressive reduction in the channel thickness to preserve the desired device performance for sub-10 nm technology nodes[33]. The ultimate channel thickness for a field-effect transistor (FET) would be in the sub-1 nm range, which is difficult to realize using bulk semiconductors[34], making 2D materials a natural choice for ultra-scaled FETs[35–41]. In fact, recent years have witnessed many experimental breakthroughs in the development of high-performance 2D FETs[42–45], neurosynaptic devices[46–50], and very large scale integrated (VLSI) circuits[51–54]. Similarly, theoretical calculations and quantum mechanical simulation have found that the 2D FETs can outperform CMOS HP (high performance) in both energy and delay[55–58].

Results

Fabrication and characterization of 2D memtransistors

Figure 2a, b, respectively, show the 2D schematic and optical image of a representative 2D memtransistor based on monolayer MoS_2, which is locally back-gated with sputter-deposited 40/30 nm Pt/TiN serving as the back-gate electrode with atomic layer deposition (ALD) grown 50 nm Al_2O_3 as the gate dielectric. All back-gate islands were placed on a commercially purchased SiO_2/p^{++}-Si substrate. As we will discuss later, the analog, non-volatile, and stochastic programming capability offered by the Al_2O_3/Pt/TiN gate stack is central to our BN architecture. The monolayer MoS_2 used in this work was grown using a metal-organic chemical vapor deposition (MOCVD) technique on a sapphire substrate at 950 °C[45, 59]. Use of an epitaxial substrate and elevated growth temperature ensured a uniform and high quality 2D film, which is critical for the successful demonstration of our BN architecture that involves many 2D memtransistors. For subsequent 2D memtransistor fabrication, the monolayer MoS_2 film was transferred from the growth substrate to the SiO_2/p^{++}-Si substrate with predefined islands of Al_2O_3/Pt/TiN. Details on monolayer MoS_2 synthesis, film transfer, and fabrication of the local back-gate gate islands, MoS_2 memtransistors, and BN architecture can be found in the "Methods" section as well as in the Methods sections of our recent works [45, 60–63].

The film quality and device performance were assessed using optical and electrical measurements. The Raman spectra (Supplementary Fig. S2a) obtained for a representative 2D memtransistor shows two characteristic monolayer MoS_2 peaks at 383 cm^{-1} and 404 cm^{-1} corresponding to the in-plane E_{2g} and out-of-plane A_{1g} modes, respectively, with the expected peak separation of ~20 cm^{-1} for monolayer MoS_2[64]. Similarly, the photoluminescence (PL) spectra (Supplementary Fig. S2b) shows a peak at 1.83 eV corresponding to the direct bandgap of monolayer MoS_2. The transfer characteristics, i.e., source-to-drain current (I_{DS}) versus local back-gate voltage (V_{BG}), measured using a source-to-drain bias (V_{DS}) of 1 V are shown in Fig. 2c in both linear and logarithmic scale for a representative MoS_2

https://doi.org/10.1038/s41467-022-33053-x

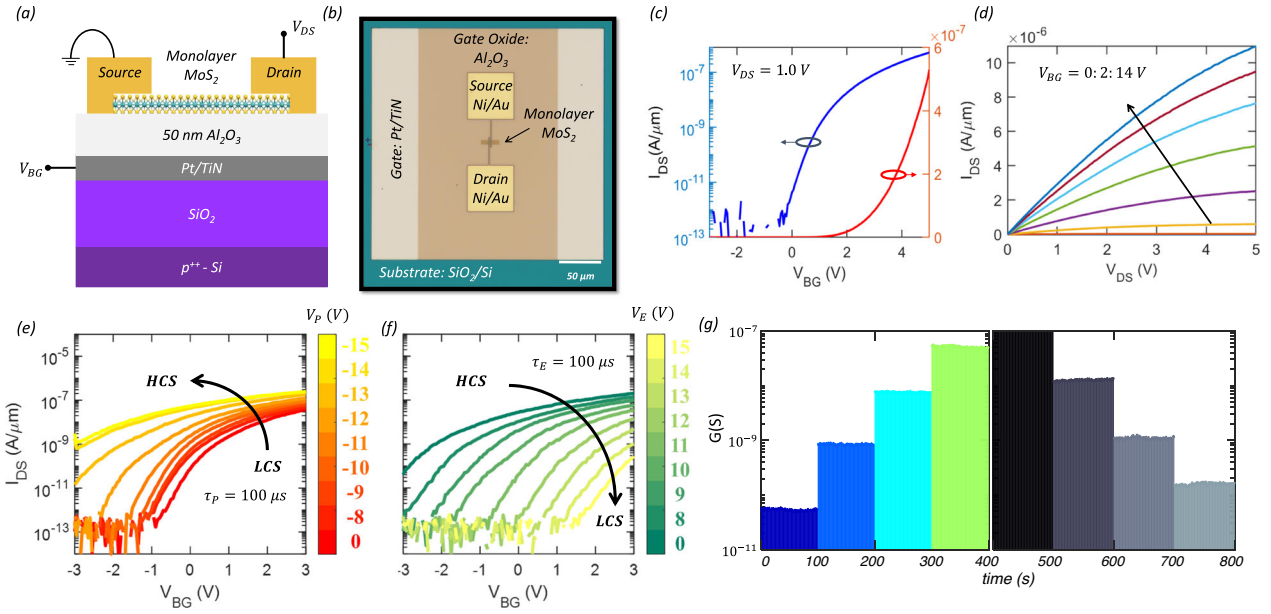


Fig. 2 | 2D memtransistors. a 2D schematic and **b** optical image of a representative 2D memtransistor based on monolayer MoS_2, which is locally back-gated with sputter-deposited 40/30 nm Pt/TiN serving as the back-gate electrode with atomic layer deposition (ALD) grown 50 nm Al_2O_3 as the gate dielectric. All back-gate islands were placed on a commercially purchased SiO_2/p^{++}-Si substrate. **c** Transfer characteristics, i.e., source-to-drain current (I_{DS}) versus local back-gate voltage (V_{BG}), measured at a source-to-drain bias (V_{DS}) of 1 V, in linear and logarithmic scale for a representative MoS_2 memtransistor with channel length $L = 1\,\mu m$ and channel width $W = 5\,\mu m$. **d** Output characteristics, i.e., I_{DS} versus V_{DS}, at different V_{BG} for the same MoS_2 memtransistor. **e** Post-programmed and **f** post-erased transfer characteristics of a representative 2D memtransistor after being subjected to negative "Write" (V_P) and positive "Erase" (V_E) voltage pulses of different amplitudes applied to the local back-gate electrode, each for a duration of $\tau_{P/E} = 100\,\mu s$ **g** Non-volatile retention for 4 representative post-programmed and post-erased conductance states (G_{MT}) over 100 s.

memtransistor with a channel length (L) of 1 μm and a channel width (W) of 5 μm. As expected, n-type transport is observed in MoS_2, which is attributed to the pinning of the metal Fermi level near the conduction band[65–67]. Nevertheless, the MoS_2 memtransistor exhibits excellent electrostatic gate control with a current on/off ratio ($r_{ON/OFF}$) > 10^5, a subthreshold slope (SS) < 400 mV/decade averaged over 3 orders of magnitude change in I_{DS}, minimal gate hysteresis when measured in air, and low gate leakage current. The threshold voltage (V_{TH}) was found to be ~1.75 V extracted at an iso-current of 10 nA/μm and the electron field effect mobility (μ_{FE}) extracted from the peak transconductance was found to be 5 cm^2/V-s. Figure 2d shows the output characteristics, i.e., I_{DS} versus V_{DS}, at different V_{BG} for the same representative MoS_2 memtransistor. The on-current (I_{ON}) reached as high as ~ 11 μA/μm for an inversion carrier density of ~1×10^{12}/cm^2 at V_{DS} = 5 V. These results suggest that the monolayer MoS_2 film grown using MOCVD is of reasonably good quality, and that the memtransistor fabrication processes including the film transfer are clean and damage-free.

The post-programmed and post-erased transfer characteristics of a representative 2D memtransistor after being subjected to negative "Write" (V_P) and positive "Erase" (V_E) voltage pulses applied to the local back-gate electrode of varying amplitudes, each for a duration of $\tau_{P/E}$ = 100 μs, are shown in Fig. 2e, f, respectively. The negative and positive shift in the respective transfer characteristics can be ascribed to electron trapping and detrapping at and near the MoS_2/Al_2O_3 interface, respectively. Note that trap states can originate from defects/imperfections in the dielectric and/or adsorbed species at the 2D/dielectric interface as reported in various earlier studies[68–70]. These states can also be engineered at desired energetic locations by introducing intentional defects in the 2D channel material[51, 71]. Carrier occupancy in these trap states follow Fermi-Dirac distribution. As illustrated using the energy band diagrams in Supplementary Fig. S3, at equilibrium, i.e., in the absence of any gate bias, the trap states with energy levels above the Fermi energy (E_F) are empty, whereas the ones below E_F are filled. When the memtransistor is subjected to a negative "Write" (V_P) voltage pulse, electrons are released (detrapped) from these trap states leaving them positively charged. This leads to screening of the back-gate bias, which is reflected as shift in the threshold voltage ($\triangle V_{TH}$). Similarly, when the memtransistor is subjected to a positive "Erase" (V_E) voltage pulse, electrons are captured back (trapped) into the trap states, restoring the V_{TH}. Note that the number of electrons getting trapped/detrapped can be controlled by both the magnitude and duration of V_P and V_E, which allow us to have an analog control of the $\triangle V_{TH}$ and of the conductance state of the memtransistor.

The minimum program/erase pulse width is determined by the trapping/detrapping time constants. Supplementary Fig. S4a–d show the post-programmed and post-erased transfer characteristics of a 2D memtransistor subjected to V_P and V_E voltage pulses of different amplitudes ranging from 8 V to 15 V applied to the local back-gate electrode, each for a duration of $\tau_{P/E}$ = 100 μs, 10 μs, 1 μs, and 100 ns, respectively. Clearly, the charge trapping and detrapping processes can occur as fast as 100 ns, which is the limit set by our measurement tools, allowing further improvement in the programming speed[72, 73]. Supplementary Fig. S4e, f show the extracted shift in the threshold voltage ($\triangle V_{TH}$) as a function of $V_{P/E}$ for $\tau_{P/E}$ = 100 μs and $\tau_{P/E}$ = 100 ns, respectively. From these results, we can conclude that, for any given pulse magnitude $V_{P/E}$, $\triangle V_{TH}$ becomes smaller as $\tau_{P/E}$ becomes shorter. To retain similar $\triangle V_{TH}$ for smaller $\tau_{P/E}$, larger $V_{P/E}$ is required, which will increase the energy expenditure. Therefore, one needs to strike a balance between fast programmability and energy consumption based on the application needs.

The trapping and detrapping processes were found to be non-volatile, as shown in Fig. 2g for 4 representative post-programmed and post-erased conductance states (G_{MT}) over 100 s. We also examined long-term memory retention for the 2D memtransistors and found that states remain distinguishable even after 3 hrs. Memory retention is important to store the CPT and the memtransistors

demonstrate adequate memory performance for the hardware implementation of BNs using SC. The program/erase endurance is also important for the 2D memtransistor. Supplementary Fig. S5 shows the post-programmed and post-erased conductance states of a representative memtransistor, achieved with $V_P = -7$ V and $V_E = 10$ V using $\tau_{P/E} = 100$ ns and measured at $V_{BG} = 0$ V for up to 10^9 endurance cycles. Clearly, there is no significant change in the two states. While it is desirable to demonstrate endurance for an even higher number of cycles, note that, for the many edge applications, the current endurance results can be sufficient. For example, in weather forecasting, the BN will be used every minute rather than every microsecond; similarly, in medical diagnostics, the BN will be only used several thousand times a day to assess patients.

Programming stochasticity in 2D memtransistors and design of s-bit generator

Design of hardware for high-quality random bit generation is central to the hardware implementation of BNs. Here, we exploit the cycle-to-cycle variation in the post-programmed and post-erased conductance states (G_{MT}) of 2D memtransistors as a source of true randomness. Figure 3a shows the transfer characteristics of a representative MoS_2 memtransistor, which is measured each time after applying $V_P = -10$ V and $V_E = 10$ V for $\tau_s = 100$ µs, for a total of 100 cycles and Fig. 3b, c, respectively, show the histograms of post-programmed and post-erased G_{MT} values extracted at $V_{BG} = 0$ V. Clearly, the G_{MT} values follow Gaussian random distributions. The cycle-to-cycle variation in program/erase processes is a direct consequence of the stochastic nature of charge trapping and detrapping observed in most semiconductor/dielectric interfaces[74–76]. In the simple two-state model, a trap state can be electrically neutral or charged, and it can transition between the two states even under equilibrium condition with transition times exponentially distributed. In other words, the state transition dynamics for traps follows the classic Markovian process[77,78]. In ultra-scaled metal-oxide-semiconductor field effect transistors (MOSFETs) such stochastic state transitions lead to random telegraph noise (RTN). Metastable states are also often involved in the trapping/detrapping processes, making the transition dynamic more complex, rich, and, at the same time, introducing an additional source of randomness[79]. While RTN is not observed in our relatively large area memtransistors, the stochasticity of trapping/detrapping processes manifest during the program/erase operations, thus leading to the cycle-to-cycle variation in $\triangle V_{TH}$.

To translate the stochastic conductance fluctuation into s-bits, we deploy a circuit consisting of six memtransistors ($MT1$, $MT2$, $MT3$, $MT4$, $MT5$, and $MT6$), as shown using the circuit diagram and corresponding optical image in Fig. 3d, e, respectively. The voltage waveforms applied to the nodes $N1$ and $N2$, i.e., V_{N1} and V_{N2}, respectively, are shown in Fig. 3f. Note that during each clock cycle (τ_{clk}), V_{N1} switches between 0 V, 0 V, and 2 V and V_{N2} switches between $V_P = -7$ V, $V_E = 10$ V, and $V_R = 1$ V. Voltages applied to nodes $N3$ and $N4$, i.e., V_{N3} and V_{N4}, are held constant at 1 V and 0 V, respectively. This allows programming and erasing of $MT1$ during each τ_{clk}. The voltage readout at node $N5$, i.e., V_{N5}, is shown in Fig. 3g and exhibits stochastic fluctuation. Note that the series connection of memtransistors $MT1$ and $MT2$ represents a voltage divider circuit, and hence V_{N5} is determined by their respective conductance values, i.e., G_{MT1} and G_{MT2}. Since G_{MT1} fluctuates from cycle-to-cycle owing to the programming and erasing voltages applied to its local back-gate terminal, i.e., $N2$, so does V_{N5}. In other words, the voltage divider translates conductance fluctuations into voltage fluctuations. Figure 3h shows the histogram of V_{N5}, which, as expected, follows a random Gaussian distribution with a mean (μ_{VN5}) of 0.40 V and standard deviation (σ_{VN5}) of 0.02 V.

Next, the Gaussian distribution is broadened by using an inverting amplifier constructed using $MT3$ and $MT4$. Note that the local back-gate of $MT3$ is shorted to its source at node N_6. This ensures that $MT3$ operates as a depletion mode (normally on) transistor or as a load resistor. Figure 3i shows the output, V_{N6}, as a function of the input, V_{N5}. The slope of the curve is referred to as the gain of the amplifier, and the higher the gain, the wider the broadening of the Gaussian. We achieved a gain of ~24. The gain can be increased further by cascading multiple amplifiers; however, this adds area and energy overhead. Figure 3j shows V_{N6} corresponding to V_{N5} obtained in Fig. 3g. Clearly, the histogram of V_{N6} shown in Fig. 3k exhibits a Gaussian distribution with a mean (μ_{VN6}) of 0.99 V and an increased standard deviation (σ_{VN6}) of 0.41 V.

Finally, to transform the analog fluctuations seen in V_{N6} into s-bits, a thresholding inverter with a programmable inversion threshold, V_{IT}, is constructed using $MT5$ and $MT6$. Figure 3l shows the output, V_{N7}, as a function of the input, V_{N6}, for different V_{IT}. Note that V_{IT} is the magnitude of V_{N6} for which V_{N7} reaches $V_{DD}/2$, i.e., 1 V in the present case. The programmability of V_{IT} is a critical feature that distinguishes 2D memtransistor-based inverters from conventional CMOS-based inverters and allows us to seamlessly obtain the s-bits. Figure 3m shows V_{N7} corresponding to V_{N6} obtained in Fig. 3j for different V_{IT} and Fig. 3n shows the corresponding probability of obtaining '1' in the bit-stream, i.e., p_s as a function of V_{IT}. As expected, if V_{IT} is too low, then almost all V_{N6} values translate into $V_{N7} \approx 0$ V, which is reflected as near zero p_s. Similarly, if V_{IT} is too high, then almost all V_{N6} values translate into $V_{N7} \approx 2$ V, leading to $p_s = 1$. Between these two extremes, p_s increases monotonically with V_{IT}. This clearly shows that we are able to convert the cycle-to-cycle random conductance fluctuations in 2D memtransistor into s-bits with reconfigurable p_s values that lie between [0,1] using the described circuit.

Note that the cycle-to-cycle variation in the programming of 2D memtransistors will lead to fluctuations in the threshold voltage (V_{TH}) of MT_6 and hence in V_{IT} of the thresholding inverter and p_s for the s-bit-stream. Supplementary Fig. S6a-b, respectively, show the distribution of V_{TH} and V_{IT} when MT_6 is subjected to 50 program/erase/read cycles with $V_P = -7$ V, $V_E = 10$ V, and $\tau_{P/E} = 100$ µs. The means and standard deviations were found to be −0.04 V and 0.08 V for V_{TH}, respectively, and 0.14 V and 0.08 V for V_{IT}, respectively. Therefore, p_s will not be perfectly deterministic; instead there will be a small uncertainty in its value, which is represented using the uncertainty band in Fig. 3n. Next, to assess randomness, we utilized the s-bit generator to generate 10^4 random bits using the same programming and erasing voltage pulses of $V_E = 10$ V and $V_P = -7$ V, respectively, at $\tau_{P/E} = 100$ µs. Supplementary Fig. S7 shows the results of eight of the statistical tests developed by the National Institute of Standards and Technology (NIST) performed on these 10^4 bits. According to the test protocol, the bit-streams are considered random only if the p-value is greater than 0.01 with the null hypothesis that the sequence is random with 99% confidence level. The NIST test results confirm that the s-bits generated are truly random.

The rough estimate of the energy expenditure for s-bit generation (E_{s-bit}) was calculated using Eq. (5).

$$E_{s-bit} = C_G\left(V_P^2 + V_E^2 + V_R^2 + V_{DD}^2\right) + \langle I_{N1N4}\rangle V_{DD}\tau_{clk} \tag{5}$$

$$\langle I_{N1N4}\rangle = \frac{1}{n}\sum_{i=1}^{n} I_{N1N4,i} \tag{6}$$

$$C_G = \varepsilon_0\varepsilon_{ox}WL/t_{ox} \tag{7}$$

In Eq. (5), V_P, V_E, V_R, and V_{DD} are the program, erase, read, and supply voltages, respectively. $C_G \approx 10^{-14}$ F is the gate capacitance, $\varepsilon_0 = 8.85 \times 10^{-12} F/m$ is the vacuum permittivity, and $\varepsilon_{ox} = 10$ and $t_{ox} = 50\,nm$ are, respectively, the relative permittivity and thickness of

https://doi.org/10.1038/s41467-022-33053-x

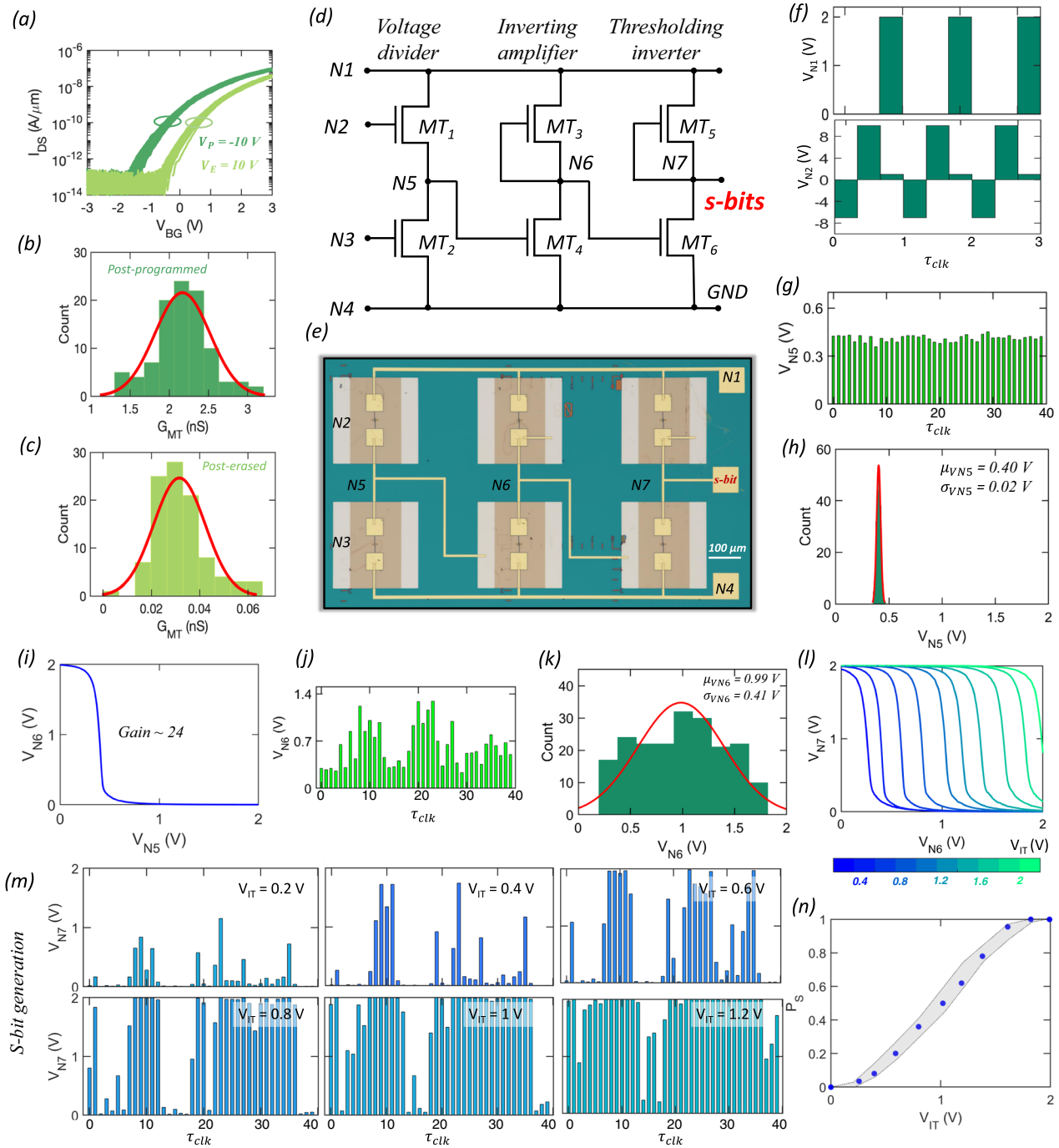


Fig. 3 | 2D memtransistor-based s-bit generator. a Transfer characteristics of a representative 2D memtransistor measured after the application of 100 cycles of $V_P = -10$ V (dark green) and $V_E = 10$ V (light green) pulses, each for $\tau_s = 100$ μs. Distribution of **b** post-programmed and **c** post-erased conductance states (G_{MT}) measured using $V_{BG} = 0$ V. **d** Circuit diagram and **e** corresponding optical image for the proposed s-bit generator consisting of six memtransistors (*MT*1, *MT*2, *MT*3, *MT*4, *MT*5, *MT*6). **f** Voltage waveforms applied to nodes *N*1 and *N*2, i.e., V_{N1} and V_{N2}. During each clock cycle (τ_{clk}), V_{N1} toggles between 0 V, 0 V, and $V_{DD} = 2$ V and V_{N2} toggles between $V_P = -7$ V, $V_E = 10$ V, and $V_R = 1$ V. Voltages applied to nodes *N*3 and *N*4, i.e., V_{N3} and V_{N4}, are held constant at 1 V and 0 V, respectively. **g** Voltage readout at node *N*5, i.e., V_{N5}. **h** Distribution of V_{N5} over 200 τ_{clk} follows a random Gaussian distribution with a mean (μ_{VN5}) of 0.40 V and a standard deviation (σ_{VN5}) of 0.02 V. **i** Output, V_{N6}, of an inverting amplifier constructed using *MT*3 and *MT*4 as a function of the input, V_{N5}, with a gain of ~24. **j** V_{N6} corresponding to V_{N5} shown in **g**. **k** Distribution of V_{N6} follows a random Gaussian distribution with a mean (μ_{VN5}) of 0.99 V and a standard deviation (σ_{VN5}) of 0.41 V. **l** Output, V_{N7}, of a thresholding inverter constructed using *MT*5 and *MT*6 as a function of the input, V_{N6}, for different inversion threshold, V_{IT}. **m** V_{N7} corresponding to V_{N6} shown in **i** for different V_{IT}. **n** Probability of obtaining '1' in the bit-stream (p_s) as a function of V_{IT}.

Al_2O_3; $W = 5$ μm and $L = 1$ μm are, respectively, the channel width and length of the 2D-memtransistor. $\langle I_{N1N4} \rangle$ is the average current flowing through the s-bit generator circuit, i.e., the total current through the voltage divider, inverting amplifier, and threshold inverter during each τ_{clk}. We have used $n = 200$ to calculate the average current per $\tau_{clk} = 100$ μs based on the experimental measurements. Since most of the memtransistors operate in their respective subthreshold regimes, the extracted $\langle I_{N1N4} \rangle$ is ~1.5 nA as shown in Supplementary Fig. S8. As such, the second term in Eq. (5) accounts for ~0.3 pJ, whereas the first term in Eq. (5) accounts for ~2 pJ. This results in $E_{s-bit} \approx 2$ pJ/clock-cycle,

https://doi.org/10.1038/s41467-022-33053-x

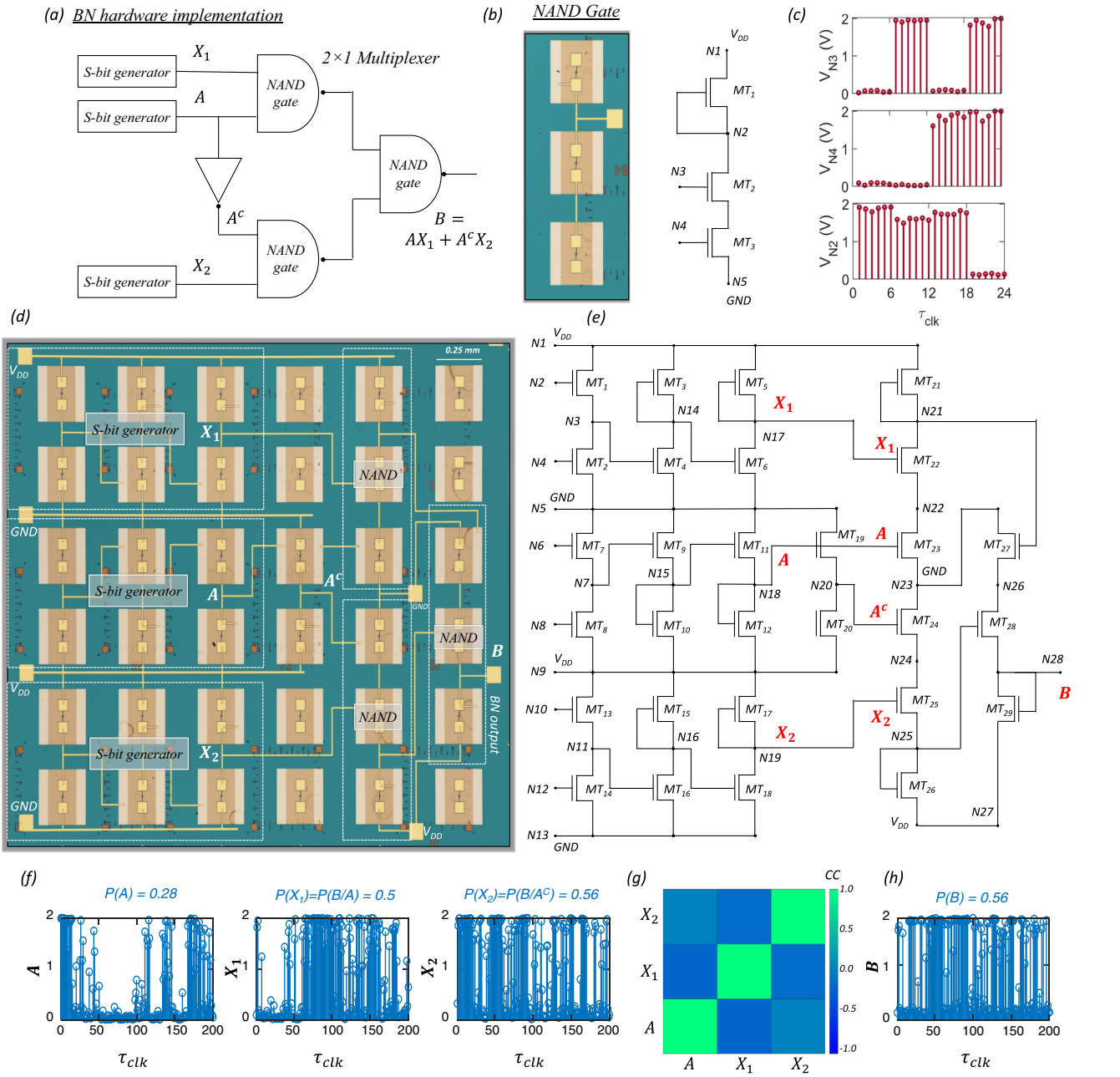


Fig. 4 | Hardware implementation of BN. a Circuit schematic for hardware implementation of a BN using three s-bit generators and one 2×1 *MUX*. The *MUX* consists of one inverter and three 2-input *NAND* gates. **b** Optical image and corresponding circuit configuration of a 2-input *NAND* gate comprising 3 memtransistors (*MT*1, *MT*2, and *MT*3) connected in series, with *MT*1 serving as the depletion load. **c** Input waveforms, V_{N3} and V_{N4}, which are applied to the local back-gate terminals of *MT*2 and *MT*3 at nodes N_3 and N_4, respectively, and the corresponding output waveform, V_{N2}, which is obtained at node N_2. **d** Optical image and **e** corresponding circuit configuration for hardware implementation of a 2-node BN consisting of 3 s-bit generators and a 2×1 *MUX* for a total of 29 memtransistors. The V_{IT} values for the s-bit generators for X_1 and X_2 can be pre-programmed using the CPT for the nodes A and B. **f** Representative stochastic bit-streams for the random variables A, X_1, and X_2 with $P(A) = 0.28$, $P(X_1) = P(B/A) = 0.50$, and $P(X_2) = P\left(B/A^C\right) = 0.56$. **g** Correlation coefficient (*CC*) values between A, X_1, and X_2 confirm mutual independence of the s-bit generator modules. **h** Stochastic bit-streams obtained at the output node, B. The measured and expected values for $P(B)$ are 0.56 and 0.54, respectively.

which supports our claim of energy efficient s-bit generation. Also note that since each memtransistor has an active device area of ~5 μm², excluding the large contact pads used for probing, the active footprint for the s-bit generator is ~30 μm². Since monolayer 2D materials offer aggressive dimensional scalability, it is possible to reduce the footprint of s-bit generators even further. Nevertheless, the use of only 6 memtransistors is the key towards the realization of area and energy efficient s-bit generator circuits.

2D memtransistor-based digital circuits and BN implementation

As described earlier, stochastic multiplexers (*MUX*s) can be used for computing the marginal probability values at any BN node. Figure 4a shows the circuit configuration of a 2×1 *MUX* which consists of one inverter and three 2-input *NAND* gates. Figure 4b shows the optical image and corresponding circuit configuration of a 2-input *NAND* gate comprising 3 memtransistors (*MT*1, *MT*2, and *MT*3) connected in series, with *MT*1 serving as the depletion load. The supply voltage,

V_{DD} = 2 V, is applied to the drain terminal of $MT1$ at node N_1, whereas the source terminal of $MT3$, i.e., node N_5, is kept grounded. Figure 4c shows the input waveforms, V_{N3} and V_{N4}, which are applied to the local back-gate terminals of $MT2$ and $MT3$ at nodes N_3 and N_4, respectively, and the corresponding output waveform, V_{N2}, which is obtained at node N_2. Clearly, the circuit operates as a *NAND* gate.

Figure 4d, e, respectively, show the optical image and corresponding circuit configuration for hardware implementation of a 2-node BN consisting of 3 s-bit generators and a 2 × 1 *MUX* for a total of 29 memtransistors. The V_{IT} values for the s-bit generators generating X_1 and X_2 can be pre-programmed corresponding to the CPT for the nodes A and B of the 2-node BN. Figure 4f shows the representative stochastic bit-streams for the random variables A, X_1, and X_2 with $P(A)$ = 0.28, $P(X_1) = P(B/A) = 0.50$, and $P(X_2) = P\left(B/A^C\right) = 0.56$. Note that accurate estimation of $P(B)$ requires that the stochastic input variables to the *MUX*, i.e., A, X_1, and X_2, must be mutually independent. Figure 4g shows the correlation coefficient (*CC*) between these three variables. The *CC* values were found to be close to zero, which confirms mutual independence of the s-bit generator modules. Figure 4h shows the stochastic bit-streams obtained at the output node, B. The measured and expected values for $P(B)$ are 0.56 and 0.54, respectively. Supplementary Fig. S9 shows the results for three more sets of measurements. In all instances, we found that our 29 memtransistor module is able to demonstrate a 2-node BN with relatively high accuracy. The rough estimate of the energy expenditure for our hardware BN implementation is miniscule at ~1.2 nJ when 200 τ_{clk} are used. Certainly, the energy expense can be further reduced by shortening the length of the s-bit streams at the cost of reduced precision. Supplementary Fig. S10 shows the numerical simulation of the error in expected values for $P(B)$ as a function of the bit-length of the s-bit stream for the inputs $P(A)$, $P(B/A)$, and $P(B/A^C)$. The percentage error increases significantly with the reduction in bit-length of the s-bit streams.

While we have experimentally demonstrated that the distribution of the output voltage (V_{N6}) from the inverting amplifier follows a Gaussian profile, it is possible that the distribution may deviate from a perfect Gaussian distribution due to many operational reasons. This will definitely lead to computation error. To assess the impact of a skewed distribution on the precision of the BN, we have performed numerical simulations assuming that V_{N6} follows the Pearson random distribution function. Supplementary Fig. S11a shows the distribution of V_{N6} for different values of skewness from −1 to 1 in steps of 0.5. Note that a skewness of −1 or 1 will be a rare occurrence under most practical circumstances. Supplementary Fig. S11b shows the corresponding p_s as a function of V_{IT}. As the skewness increases, the deviation of p_s from its expected value also increases. Supplementary Fig. S11c shows the colormap of the percentage error in estimating $P(B)$ using the BN hardware for different skewness in the stochastic input variables X_1 and X_2 that represent $P(B/A)$ and $P\left(B/A^C\right)$, respectively. As expected, the percentage error increases with increasing skewness. Furthermore, we have experimentally demonstrated that the distribution of the inverting threshold voltage (V_{IT}) exhibits a Gaussian distribution after $MT6$ is subjected to 50 program/erase/read cycles with V_P = −7 V, V_E = 10 V, and $\tau_{P/E}$ = 100 µs. This V_{IT} distribution leads to a small uncertainty ($\triangle P$) in probability of output voltages (V_{N7}), as shown in Fig. 3n. We have used numerical simulations to assess the impact of uncertainty in obtained probabilities on the precision of the BN, where the probability of the select line, A, remains as a constant while the probability of both X_1 and X_2 are inflicted with $\triangle P$ due to cycle-to-cycle variation in the programmed probability. Supplementary Fig. S12 shows the colormap of the percentage error in estimating $P(B)$ using the BN hardware for uncertainty in the stochastic input variables X_1 and X_2 that represent $P(B/A) = 0.50$ and $P\left(B/A^C\right) = 0.56$, respectively, while $P(A) = 0.28$ and $\triangle P \approx 0.065$. From this colormap, we can conclude that even if the V_{IT} of the thresholding inverter ($MT6$) is inflicted with cycle-to-cycle variation from device programming, the inaccuracy of the 2-node Bayesian network ($B = AX_1 + A^C X_2$) is less than 15%. This simulation result shows decent accuracy in hardware implementation of the BN.

Finally, the impact of device-to-device variation on the operation of BN is examined. Supplementary Fig. S13a shows the transfer characteristics of 10 MoS_2 memtransistors and Supplementary Fig. S13b shows the transfer characteristics for these 10 devices after one programming/erasing clock cycle (V_P = −7 V, V_E = 10 V, and $\tau_{P/E}$ = 100 µs.). The device-to-device variation translates into error in $\triangle P$ and impacts the accuracy at the output of the BN. Supplementary Fig. S14 shows the colormap of error in $P(B)$ for $P(X_1) = 0.5$, $P(X_2) = 0.56$, and $P(A) = 0.28$. We have used $\triangle P$ =0.046 for both X_1 and X_2 inferred from Supplementary Fig. S13b. From the error map, it is evident that the variation in the programmed probability inflicted by the device-to-device programming variation of the memtransistors resulted in a maximum error of 8% at the output of the BN.

Discussion

In conclusion, we have exploited cycle-to-cycle variability in the programmed conductance of 2D memtransistors and transcribed the same into s-bits with reconfigurable probability of obtaining '1' in the bit-stream using a circuit that comprises only 6 memtransistors and spends < 2 pJ per s-bit. We subsequently combined the s-bit generator with a 2D memtransistor-based 2 × 1 *MUX* to demonstrate hardware implementation of a BN. The BN architecture comprises 29 memtransistors and requires ~ 1.2 nJ of energy for precise computation. Our demonstration of a memtransistor-based standalone in-memory compute fabric shows the potential for emerging 2D materials and devices.

Methods

Fabrication of local back-gate islands

To define the back-gate island regions, a commercially-purchased substrate (285 nm SiO_2 on p^{++}-Si) was spin coated (4000 RPM for 45 s) with bilayer photoresist consisting of Lift-Off-Resist (LOR 5 A) and Series Photoresist (SPR 3012) and baked at 185 °C for 120 s and 95 °C for 60 s, respectively. The bilayer photoresist was then exposed using a Heidelburg 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. An atomic layer deposition (ALD) process was then implemented to grow 50 nm Al_2O_3 across 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 5 A and SPR 3012. The bilayer photoresist was then exposed to MLA 150 and developed using MF CD26 microposit. The 50 nm Al_2O_3 was subsequently dry etched using a BCl_3 reactive ion etch (RIE) chemistry at 5 °C for 20 s, which was repeated four times to minimize heating in the substrate. Finally, the photoresist was removed to give access to the individual Pt electrodes.

Large-area monolayer MoS_2 film growth

Monolayer MoS_2 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[80]. Molybdenum hexacarbonyl ($Mo(CO)_6$) and hydrogen sulfide (H_2S) were used as precursors. $Mo(CO)_6$ maintained at 10 °C and 650 Torr in a stainless-steel bubbler was used to deliver 1.1×10^{-3} sccm of the metal precursor for the growth, while 400 sccm of H_2S was used for the process. MoS_2 deposition was carried out at

1000 °C and 50 Torr in H_2 ambient, where monolayer growth was achieved in 18 min. The substrate was first heated to 1000 °C in H_2 and maintained for 10 min before the growth was initiated. After growth, the substrate was cooled in H_2S to 300 °C to inhibit decomposition of the MoS_2 films. More details can be found in our earlier work[45, 48, 81].

MoS_2 film transfer to local back-gate islands

To fabricate the 2D memtransistors, the MOCVD-grown monolayer MoS_2 film was transferred from the sapphire growth substrate to the SiO_2/p^{++}-Si substrate with local back-gate islands using a PMMA (polymethyl-methacrylate) assisted wet transfer process. First, growth substrate was spin coated with PMMA and left to sit for 24 h to promote PMMA/MoS_2 adhesion. 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 caused the NaOH to be drawn into the substrate/film interface, separating the PMMA/MoS_2 film from the sapphire substrate. The separated film was rinsed three times inside separate water baths and fished-out using the SiO_2/p^{++}-Si substrate with local back-gate islands. The substrate was then baked at 50 °C and 70 °C for 10 min each to remove moisture and promote adhesion. An acetone bath was usd to remove the PMMA supporting layer, with a subsequent IPA bath to remove residue.

Fabrication of 2D memtransistors

To define the channel regions for the memtransistors, the substrate was spin-coated with PMMA and baked at 180 °C for 90 s. The resist was then patterned using electron beam (e-beam) lithography and developed using a 1:1 mixture of 4-methyl-2-pentanone (MIBK) and 2 propanol (IPA), with a subsequent IPA rinse. The monolayer MoS_2 film was then etched using a sulfur hexafluoride (SF_6) RIE chemistry at 5 °C for 30 s. Next, the sample was rinsed in acetone and IPA to remove PMMA. To define the source and drain contacts, sample was then spin coated with methyl methacrylate (MMA) followed by PMMA. E-beam lithography was used to pattern the source and drain contacts and 1:1 MIBK/ IPA was again used for development. 40 nm of nickel (Ni) and 30 nm of gold (Au) were deposited using e-beam evaporation. Finally, a lift-off process was performed to remove the excess Ni/Au and resist by immersing the sample in acetone for 30 min followed by IPA for another 30 mins. Each island contains one memtransistor to allow for individual gate control.

Monolithic integration

To define the connections between respective memtransistors, the substrate was spin coated with MMA and PMMA, followed by e-beam lithography and development using a 1:1 mixture of MIBK/IPA. E-beam evaporation of was used to deposit 60 nm of Ni and 30 nm of Au to form the connections. Finally, the e-beam resist was rinsed away by the same acetone and IPA lift-off process used previously.

Electrical characterization

Electrical characterization of the fabricated devices was performed using a Lake Shore CRX-VF probe station under atmospheric conditions and with 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. Brown, G. E. & Smith, R. J. F. Conspecific skin extracts elicit anti-predator responses in juvenile rainbow trout (Oncorhynchus mykiss). *Can. J. Zool.* **75**, 1916–1922 (1997).
2. Puga, J. L., Krzywinski, M. & Altman, N. Bayesian networks. *Nat. Methods.* **12**, 799–800 (2015).
3. Smithson, S. C., Onizawa, N., Meyer, B. H., Gross, W. J. & Hanyu, T. Efficient CMOS Invertible Logic Using Stochastic Computing. *IEEE Trans. Circuits Syst. I: Regul. Pap.* **66**, 2263–2274 (2019).
4. Ardakani, A., Leduc-Primeau, F., Onizawa, N., Hanyu, T. & Gross, W. J. VLSI Implementation of Deep Neural Network Using Integral Stochastic Computing. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **25**, 2688–2699 (2017).
5. Alaghi, A. & Hayes, J. P. Survey of stochastic computing. *ACM Trans. Embedded Comput. Syst. (TECS)* **12**, 1–19 (2013).
6. Gaines, B. R. In *Proceedings of the April 18–20, 1967, spring joint computer conference*. 149–156.
7. Poppelbaum, W., Afuso, C. & Esch, J. In *Proceedings of the November 14–16, 1967, fall joint computer conference*. 635–644.
8. Weijia, Z., Ling, G. W. & Seng, Y. K. In *2007 IEEE Conference on Electron Devices and Solid-State Circuits*. 337–340.
9. Cai, R. et al. in *Proceedings of the Twenty-Third International Conference on Architectural Support for Programming Languages and Operating Systems* 476–488 (Association for Computing Machinery, Williamsburg, VA, USA, 2018).
10. Kulesza, Z. & Tylman, W. In *Proceedings of the International Conference Mixed Design of Integrated Circuits and System, 2006. MIXDES 2006*. 711-715.
11. Zermani, S., Dezan, C., Chenini, H., Diguet, J. & Euler, R. In *2015 IEEE Conference on Prognostics and Health Management (PHM)*. 1-10.
12. Knag, P., Lu, W. & Zhang, Z. A native stochastic computing architecture enabled by memristors. *IEEE Trans. Nanotechnol.* **13**, 283–293 (2014).
13. Gaba, S., Knag, P., Zhang, Z. & Lu, W. In *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*. 2592-2595 (IEEE).
14. Gaba, S., Sheridan, P., Zhou, J., Choi, S. & Lu, W. Stochastic memristive devices for computing and neuromorphic applications. *Nanoscale* **5**, 5872–5878 (2013).
15. Debashis, P. *et al*. Hardware implementation of Bayesian network building blocks with stochastic spintronic devices. *Sci. Reports.* **10**, https://doi.org/10.1038/s41598-020-72842-6 (2020).
16. Faria, R., Kaiser, J., Camsari, K. Y. & Datta, S. Hardware Design for Autonomous Bayesian Networks. *Front. Computational Neurosci.* **15**, https://doi.org/10.3389/fncom.2021.584797 (2021).
17. Shim, Y., Chen, S., Sengupta, A. & Roy, K. Stochastic Spin-Orbit Torque Devices as Elements for Bayesian Inference. *Sci. Rep.* **7**, 14101 (2017).
18. Faria, R., Camsari, K. Y. & Datta, S. Implementing Bayesian networks with embedded stochastic MRAM. *AIP Adv.* **8**, 045101 (2018).
19. Venkatesan, R., Venkataramani, S., Fong, X., Roy, K. & Raghunathan, A. in *2015 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. 1575–1578 (IEEE).
20. Finocchio, G. et al. The promise of spintronics for unconventional computing. *J. Magn. Magn. Mater.* **521**, 167506 (2021).
21. Hu, J., Li, B., Ma, C., Lilja, D. & Koester, S. J. Spin-Hall-Effect-Based Stochastic Number Generator for Parallel Stochastic Computing. *IEEE Trans. Electron Devices.* **66**, 3620–3627 (2019).
22. Yang, K. et al. In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*. 280-281 (IEEE).
23. Pamula, V. R. et al. In *2018 IEEE Symposium on VLSI Circuits*. 1–2.
24. Satpathy, S. et al. In *2018 IEEE Symposium on VLSI Circuits*. 169–170.
25. Jayaraj, A., Gujarathi, N. N., Venkatesh, I. & Sanyal, A. 0.6–1.2 V, 0.22 pJ/bit True Random Number Generator Based on SAR ADC. *IEEE Trans. Circuits Syst. II: Express Briefs.* **67**, 1765–1769 (2020).

26. Bae, S., Kim, Y., Park, Y. & Kim, C. 3-Gb/s High-Speed True Random Number Generator Using Common-Mode Operating Comparator and Sampling Uncertainty of D Flip-Flop. *IEEE J. Solid-State Circuits.* **52**, 605–610 (2017).
27. Cao, Y., Zhao, X., Zheng, W., Zheng, Y. & Chang, C. H. A New Energy-Efficient and High Throughput Two-Phase Multi-Bit per Cycle Ring Oscillator-Based True Random Number Generator. *IEEE Trans. Circuits Syst. I: Regul. Pap.* **69**, 272–283 (2022).
28. Jaiswal, A., Fong, X. & Roy, K. Comprehensive scaling analysis of current induced switching in magnetic memories based on in-plane and perpendicular anisotropies. *IEEE J. Emerg. Sel. Top. Circuits Syst.* **6**, 120–133 (2016).
29. Sengupta, A., Panda, P., Wijesinghe, P., Kim, Y. & Roy, K. Magnetic tunnel junction mimics stochastic cortical spiking neurons. *Sci. Rep.* **6**, 1–8 (2016).
30. Daniels, M. W., Madhavan, A., Talatchian, P., Mizrahi, A. & Stiles, M. D. Energy-Efficient Stochastic Computing with Superparamagnetic Tunnel Junctions. *Phys. Rev. Appl.* **13**, 034016 (2020).
31. Vodenicarevic, D. et al. Low-Energy Truly Random Number Generation with Superparamagnetic Tunnel Junctions for Unconventional Computing. *Phys. Rev. Appl.* **8**, 054045 (2017).
32. Li, M.-Y., Su, S.-K., Wong, H.-S. P. & Li, L.-J. (Nature Publishing Group, 2019).
33. Jacob, A. P. et al. Scaling challenges for advanced CMOS devices. *Int. J. High. Speed Electron. Syst.* **26**, 1740001 (2017).
34. Uchida, K. et al. In *Digest. International Electron Devices Meeting*. 47–50 (IEEE).
35. Manzeli, S., Ovchinnikov, D., Pasquier, D., Yazyev, O. V. & Kis, A. 2D transition metal dichalcogenides. *Nat. Rev. Mater.* **2**, 17033 (2017).
36. Akinwande, D. et al. Graphene and two-dimensional materials for silicon technology. *Nature* **573**, 507–518 (2019).
37. Chhowalla, M., Jena, D. & Zhang, H. Two-dimensional semiconductors for transistors. *Nat. Rev. Mater.* **1**, 1–15 (2016).
38. Schwierz, F., Pezoldt, J. & Granzner, R. Two-dimensional materials and their prospects in transistor electronics. *Nanoscale* **7**, 8261–8283 (2015).
39. Liu, C. et al. Two-dimensional materials for next-generation computing technologies. *Nat. Nanotechnol.* **15**, 545–557 (2020).
40. Iannaccone, G., Bonaccorso, F., Colombo, L. & Fiori, G. Quantum engineering of transistors based on 2D materials heterostructures. *Nat. Nanotechnol.* **13**, 183–191 (2018).
41. Liu, Y. et al. Promises and prospects of two-dimensional transistors. *Nature* **591**, 43–53 (2021).
42. Shen, P.-C. et al. Ultralow contact resistance between semimetal and monolayer semiconductors. *Nature* **593**, 211–217 (2021).
43. English, C. D., Smithe, K. K. H., Xu, R. L. & Pop, E. In *2016 IEEE International Electron Devices Meeting (IEDM)*. 5.6.1-5.6.4.
44. Price, K. M., Schauble, K. E., McGuire, F. A., Farmer, D. B. & Franklin, A. D. Uniform Growth of Sub-5-Nanometer High-κ Dielectrics on MoS2 Using Plasma-Enhanced Atomic Layer Deposition. *ACS Appl. Mater. Interfaces.* **9**, 23072–23080 (2017).
45. Sebastian, A., Pendurthi, R., Choudhury, T. H., Redwing, J. M. & Das, S. Benchmarking monolayer MoS2 and WS2 field-effect transistors. *Nat. Commun.* **12**, 693 (2021).
46. Sebastian, A., Pannone, A., Subbulakshmi Radhakrishnan, S. & Das, S. Gaussian synapses for probabilistic neural networks. *Nat. Commun.* **10**, 4199 (2019).
47. Subbulakshmi Radhakrishnan, S., Sebastian, A., Oberoi, A., Das, S. & Das, S. A biomimetic neural encoder for spiking neural network. *Nat. Commun.* **12**, 2143 (2021).
48. Jayachandran, D. et al. A low-power biomimetic collision detector based on an in-memory molybdenum disulfide photodetector. *Nat. Electron.* **3**, 646–655 (2020).
49. Schranghamer, T. F., Oberoi, A. & Das, S. Graphene memristive synapses for high precision neuromorphic computing. *Nat. Commun.* **11**, 5474 (2020).
50. Das, S., Dodda, A. & Das, S. A biomimetic 2D transistor for audiomorphic computing. *Nat. Commun.* **10**, 3450 (2019).
51. Das, S. et al. Transistors based on two-dimensional materials for future integrated circuits. *Nat. Electron.* **4**, 786–799 (2021).
52. Zhu, K. et al. The development of integrated circuits based on two-dimensional materials. *Nat. Electron.* **4**, 775–785 (2021).
53. Wachter, S., Polyushkin, D. K., Bethge, O. & Mueller, T. A microprocessor based on a two-dimensional semiconductor. *Nat. Commun.* **8**, 14948 (2017).
54. Polyushkin, D. K. et al. Analogue two-dimensional semiconductor electronics. *Nat. Electron.* **3**, 486–491 (2020).
55. Nikonov, D. E. & Young, I. A. Benchmarking of beyond-CMOS exploratory devices for logic integrated circuits. *IEEE J. Exploratory Solid-State Computational Devices Circuits* **1**, 3–11 (2015).
56. Sylvia, S. S., Alam, K. & Lake, R. K. Uniform benchmarking of low-voltage van der Waals FETs. *IEEE J. Exploratory Solid-State Computational Devices Circuits* **2**, 28–35 (2016).
57. Lee, C.-S., Cline, B., Sinha, S., Yeric, G. & Wong, H. S. P. 32-bit Processor core at 5-nm technology: Analysis of transistor and interconnect impact on VLSI system performance. 28.23.21-28.23.24, https://doi.org/10.1109/iedm.2016.7838498 (2016).
58. Agarwal, T. et al. Benchmarking of monolithic 3D integrated MX_2 FETs with Si FinFETs. 5.7.1–5.7.4, https://doi.org/10.1109/iedm.2017.8268336 (2017).
59. 2DCC. *2d-crystal-consortium*, <https://www.mri.psu.edu/2d-crystal-consortium/user-facilities/thin-films/list-thin-film-samples-available>
60. Pendurthi, R. et al. Heterogeneous Integration of Atomically Thin Semiconductors for Non-von Neumann CMOS. *Small* **18**, 2202590 (2022).
61. Radhakrishnan, S. S. et al. A Sparse and Spike-timing-based Adaptive Photo Encoder for Augmenting Machine Vision for Spiking Neural Networks. *Adv. Materials*. 2202535 (2022).
62. Dodda, A., Trainor, N., Redwing, J. & Das, S. All-in-one, bio-inspired, and low-power crypto engines for near-sensor security based on two-dimensional memtransistors. *Nat. Commun.* **13**, 1–12 (2022).
63. Oberoi, A., Dodda, A., Liu, H., Terrones, M. & Das, S. Secure Electronics Enabled by Atomically Thin and Photosensitive Two-Dimensional Memtransistors. *ACS Nano.* **15**, 19815–19827 (2021).
64. Li, H. et al. From bulk to monolayer MoS2: evolution of Raman scattering. *Adv. Funct. Mater.* **22**, 1385–1390 (2012).
65. Das, S., Chen, H.-Y., Penumatcha, A. V. & Appenzeller, J. High performance multilayer MoS2 transistors with scandium contacts. *Nano Lett.* **13**, 100–105 (2013).
66. Schulman, D. S., Arnold, A. J. & Das, S. Contact engineering for 2D materials and devices. *Chem. Soc. Rev.* **47**, 3037–3058 (2018).
67. Chuang, S. et al. MoS2 p-type transistors and diodes enabled by high work function MoO x contacts. *Nano Lett.* **14**, 1337–1342 (2014).
68. Arnold, A. J. et al. Mimicking Neurotransmitter Release in Chemical Synapses via Hysteresis Engineering in MoS2 Transistors. *ACS Nano.* **11**, 3110–3118 (2017).
69. Illarionov, Y. Y. et al. The role of charge trapping in $MoS_2/SiO_2$and MoS_2/hBN field-effect transistors. *2D Materials* **3**, https://doi.org/10.1088/2053–1583/3/3/035004 (2016).
70. Illarionov, Y. Y. et al. Energetic mapping of oxide traps in MoS2 field-effect transistors. *2D Mater.* **4**, 025108 (2017).
71. Jiang, J. et al. Defect engineering for modulating the trap states in 2D photoconductors. *Adv. Mater.* **30**, 1804332 (2018).
72. Tsai, H.-S. et al. Ultrafast exciton dynamics in scalable monolayer MoS2 synthesized by metal sulfurization. *ACS Omega.* **5**, 10725–10730 (2020).

73. Docherty, C. J. et al. Ultrafast transient terahertz conductivity of monolayer MoS2 and WSe2 grown by chemical vapor deposition. *ACS Nano*. **8**, 11147–11153 (2014).
74. Bhagdikar, S. & Mahapatra, S. In *2019 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*. 1-4.
75. Grasser, T. Stochastic charge trapping in oxides: From random telegraph noise to bias temperature instabilities. *Microelectron. Reliab.* **52**, 39–70 (2012).
76. Waltl, M. In *2019 IEEE International Integrated Reliability Workshop (IIRW)*. 1-9.
77. Kirton, M. & Uren, M. Noise in solid-state microstructures: A new perspective on individual defects, interface states and low-frequency (1/*f*) noise. *Adv. Phys.* **38**, 367–468 (1989).
78. Ibe, O. *Markov processes for stochastic modeling*. (Newnes, 2013).
79. Uren, M., Kirton, M. & Collins, S. Anomalous telegraph noise in small-area silicon metal-oxide-semiconductor field-effect transistors. *Phys. Rev. B* **37**, 8346 (1988).
80. Xuan, Y. et al. Multi-scale modeling of gas-phase reactions in metal-organic chemical vapor deposition growth of WSe2. *J. Crys.Growth* **527**, https://doi.org/10.1016/j.jcrysgro.2019.125247 (2019).
81. Dodda, A. et al. Stochastic resonance in MoS2 photodetector. *Nat. Commun.* **11**, 4406 (2020).

Acknowledgements

The work was supported by Army Research Office (ARO) through Contract Number W911NF1920338 and National Science Foundation (NSF) through a CAREER Award under grant no. ECCS-2042154. 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-2039351.

Author contributions

S.D. conceived the idea and designed the experiments. Y.Z., H.R., and T.F.S. fabricated the memtransistors. Y.Z., H.R., and S.D. performed the measurements, analyzed the data, discussed the results, and agreed on their implications. N.T. grew MOCVD MoS_2 under the supervision of J.M. R. All authors contributed to the preparation of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-33053-x.

Correspondence and requests for materials should be addressed to Saptarshi Das.

Peer review information *Nature Communications* thanks Punyashloka Debashis, Hyungjin Kim and the other, anonymous, reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022

NANO · MICRO small
www.small-journal.com

Heterogeneous Integration of Atomically Thin Semiconductors for Non-von Neumann CMOS

*Rahul Pendurthi, Darsith Jayachandran, Azimkhan Kozhakhmetov, Nicholas Trainor, Joshua A. Robinson, Joan M. Redwing, and Saptarshi Das**

Atomically thin, 2D, and semiconducting transition metal dichalcogenides (TMDs) are seen as potential candidates for complementary metal oxide semiconductor (CMOS) technology in future nodes. While high-performance field effect transistors (FETs), logic gates, and integrated circuits (ICs) made from n-type TMDs such as MoS_2 and WS_2 grown at wafer scale have been demonstrated, realizing CMOS electronics necessitates integration of large area p-type semiconductors. Furthermore, the physical separation of memory and logic is a bottleneck of the existing CMOS technology and must be overcome to reduce the energy burden for computation. In this article, the existing limitations are overcome and for the first time, a heterogeneous integration of large area grown n-type MoS_2 and p-type vanadium doped WSe_2 FETs with non-volatile and analog memory storage capabilities to achieve a non–von Neumann 2D CMOS platform is introduced. This manufacturing process flow allows for precise positioning of n-type and p-type FETs, which is critical for any IC development. Inverters and a simplified 2-input-1-output multiplexers and neuromorphic computing primitives such as Gaussian, sigmoid, and tanh activation functions using this non–von Neumann 2D CMOS platform are also demonstrated. This demonstration shows the feasibility of heterogeneous integration of wafer scale 2D materials.

1. Introduction

Complementary metal oxide semiconductor (CMOS) technology is the brain of countless modern electronics gadgets, instruments, and computing tools that are used in communications, health care, defense, transportation, manufacturing, agriculture, and many other applications transforming society for the better. It is remarkable how the CMOS technology has sustained continuous growth in performance, energy efficiency, and cost per functionality for more than six decades thriving on the relentless scaling of the silicon-based metal oxide semiconductor field effect transistors (MOSFETs) following the Moore's law.[1] Today, the CMOS technology is so powerful that it serves as the foundation for most other emerging technologies such as artificial intelligence, virtual reality, the Internet of Things, smart sensing, clean energy, autonomous vehicles, and many more.[2] While still on full throttle, the silicon CMOS has started to show early signs of trouble for future technology nodes.

There are two major technology roadblocks ahead: 1) For sub-10 nm nodes, aggressive reduction in the channel thickness is necessary to preserve the gate electrostatics and hence the desirable device performance.[3] Silicon and other bulk semiconductors are vulnerable to thickness scaling due to detrimental quantum confinement effects and increased scattering of charge carriers at the channel-to-dielectric interfaces resulting in severe mobility degradation.[4] 2) The physical separation of the arithmetic core, or the logic unit, from the data storage or the memory unit, which is mandated by the von Neuman architecture, limits the energy efficiency of CMOS processors for compute and data heavy applications.[5]

The solution to the first problem is innovation in the ultrathin channel materials. The ultimate channel thickness that one can envision for an FET would be in the sub-1 nm range. 2D semiconductors such as transition metal dichalcogenides (TMDs) with < 0.7 nm body thicknesses at the monolayer limit are, therefore, promising choices for the advanced CMOS technology nodes.[6–14] Moreover the absence of dangling bonds owing to the van der Waals nature of these 2D materials ensure better channel-to-dielectric interfaces that translates into better

R. Pendurthi, D. Jayachandran, S. Das
Engineering Science and Mechanics
Penn State University
University Park, PA 16802, USA

A. Kozhakhmetov, N. Trainor, J. A. Robinson, J. M. Redwing, S. Das
Materials Science and Engineering
Penn State University
University Park, PA 16802, USA
E-mail: sud70@psu.edu

N. Trainor, J. A. Robinson, J. M. Redwing, S. Das
2D Crystal Consortium - Materials Innovation Platform (2DCC-MIP)
Materials Research Institute
Penn State University
University Park, PA 16802, USA

S. Das
Electrical Engineering and Computer Science
Penn State University
University Park, PA 16802, USA

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/smll.202202590.

DOI: 10.1002/smll.202202590

© 2022 Wiley-VCH GmbH

Check for updates

16136829, 2022, 33, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/smll.202202590 by Boise State University Albertsons Library, Wiley Online Library on [30/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

device performance. In fact, the early reports on 2D FETs[15,16] based on mechanically exfoliated single crystal flakes were so promising that it inspired material scientists across the globe to advance the large area growth techniques for the scalable manufacturing of 2D monolayers. For example, Kang et al.[17] demonstrated wafer-scale batch fabrication of high-performance monolayer MoS_2 FETs grown via metal organic chemical vapor deposition (MOCVD) technique as the first step toward the realization of very large scale integrated (VLSI) circuits. More recent developments include the demonstration of ultra-scaled MOCVD MoS_2 FETs with 42 nm contact pitch and high ON current > 250 μA μm^{-1}.[18] Some of the early criticism of 2D FETs have also been successfully mitigated through the realization of low contact resistance < 125 Ω μm,[19] integration of ultra-thin and high-k gate dielectric,[20] and reduction in the device-to-device variability.[21–23] In fact, MOCVD MoS_2 FET based microprocessors,[24] analogue operational amplifier,[25] and RF electronics components[26] have already been reported highlighting the rapid transition of the 2D technology from single devices to ICs.

Unfortunately, however, the aforementioned advancements are based on n-type 2D materials, that is, MoS_2 and WS_2, and no CMOS capability has yet been demonstrated. Achieving 2D CMOS requires both n-type and p-type FETs, which remain a challenge due to the lack of large area grown p-type 2D materials. Hence several previous 2D CMOS demonstrations have been based on exfoliated materials,[27,28] which while promising cannot be translated to IC technology owing to non-scalability of the exfoliation process. Some of our earlier work does indicate the possibility of obtaining p-type 2D materials including carbon plasma treatment of CVD monolayer WS_2[29] and observation of ambipolar transport in WSe_2.[30,31] More recently, we have demonstrated controllable p-type doping of MOCVD WSe_2 via Vanadium substitution.[32] Although these research efforts highlight the potential of TMDs for CMOS integrated circuit (IC) technology, certain processing, fabrication, and design constraints must be mitigated. This includes large area uniform growth, clean and damage-free transfer, and heterogeneous integration and precise positioning of p-type and n-type material on the same substrate for efficient circuit design.

The solution to the second problem encountering the CMOS is non–von Neumann architectures, where the physical separation between logic and memory is eliminated. In-memory computing is considered as the most potential solution, where the memory units can perform computational tasks, helping to "jump the memory wall".[5] While the memristor technology offers powerful solution[33] in this space, hybrid design involving CMOS interface is necessary since two-terminal memristors are inadequate for peripheral analog and digital computations. Recently, three-terminal 2D memtransistors or reconfigurable 2D FETs have garnered attention for in-memory compute[34–38] owing to logic functionality achievable on-chip without CMOS peripherals. However, all demonstrations in this context that involve large area synthetic 2D material are also based on n-type MoS_2. Numerous reports of in-memory compute exist based on exfoliated p-type 2D materials such as WSe_2 and black phosphorus,[39–44] but no reports can be found that involve large area p-type 2D semiconductor for in-memory compute. Needless to say, 2D CMOS based in-memory compute also remains to be demonstrated.

In this article, we circumvent the aforementioned limitations and demonstrate 1) a bottom-up fabrication process flow for heterogeneous integration of large area MOCVD grown n-type MoS_2 and p-type Vanadium doped WSe_2 (V-WSe_2) for the acceleration of 2D CMOS technology and 2) eliminate the physical separation of logic and memory through the integration of non-volatile and analog memory storage capabilities with these 2D FETs for post-von-Neumann CMOS. Using these capabilities, we demonstrate logic elements such as inverters with reconfigurable inversion threshold, 2-input-1-output multiplexers (MUXs), and neuromorphic computing primitives such as sigmoid, hyperbolic tangent (tanh), and Gaussian neurosynaptic devices on the same hardware platform. To the best of our knowledge this is the first step toward the realization of ICs based on non–von Neumann 2D CMOS technology exploiting large area growth of atomically thin monolayer semiconductors.

2. Large Area Growth of p-Type V-WSe_2 and n-Type MoS_2

V-WSe_2 films were deposited on a 10 mm × 10 mm *c*-plane sapphire substrates in a custom-built, cold-wall MOCVD reactor. Tungsten hexacarbonyl ($W(CO)_6$), hydrogen diselenide (H_2Se), and vanadocene ($V(C_5H_5)_2$) were used as metal, chalcogen, and dopant precursors, respectively, with H_2 as the carrier gas. The MOCVD synthesis is carried out using a three-step growth method, where the dopants are simultaneously introduced with the metal precursor into the reaction chamber at all the stages of the growth. The deposition temperature and pressure were kept constant at 800 °C and 700 Torr, respectively, and the total growth time was < 1 h. **Figure 1**a shows an optical image of fully coalesced monolayer V-WSe_2 on the growth substrate. Raman spectroscopy shown in Figure 1b reveals the increased intensity of defect activated ZA(M) and LA(M) modes at 129.2 and 109.9 cm^{-1}, respectively along with the A+E mode of WSe_2 at 248.7 cm^{-1} which is the direct evidence of perturbed lattice of V-WSe_2 by the intentional impurity introduction. In addition, large crystal domain size of ≈ 800 nm is observed, which can be attributed to the reduction of carbon contamination, stemming from the precursors, through the use of H_2S. Moreover, as shown in Figure 1c, the PL signal is completely quenched as a result of non-radiative recombination of positive trions where its density is expected to increase due to the increased number of holes in the film. Figure S1, Supporting Information, shows the X-ray photoelectron spectroscopy scan of V-WSe_2. Extrinsic WSe_2 exhibits characteristic V 2p3/2 and V 2p1/2 peaks at 513.5 and 521.3 eV corresponding to V-Se bonding as well as V 2p3/2 (oxide) peak at 515.9 eV suggesting the presence of V–O bonding. XPS data are also used to estimate the Vanadium doping concentration considering W 4f, Se 3d, and V 2p spectra. We found that the doping concentration is ≈ 1.9%.

Monolayer MoS_2 was deposited on epi-ready 2″ *c*-sapphire substrate by 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.[45] Molybdenum hexacarbonyl ($Mo(CO)_6$)

© 2022 Wiley-VCH GmbH

16136829, 2022, 33, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/smll.202202590 by Boise State University Albertsons Library, Wiley Online Library on [30/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

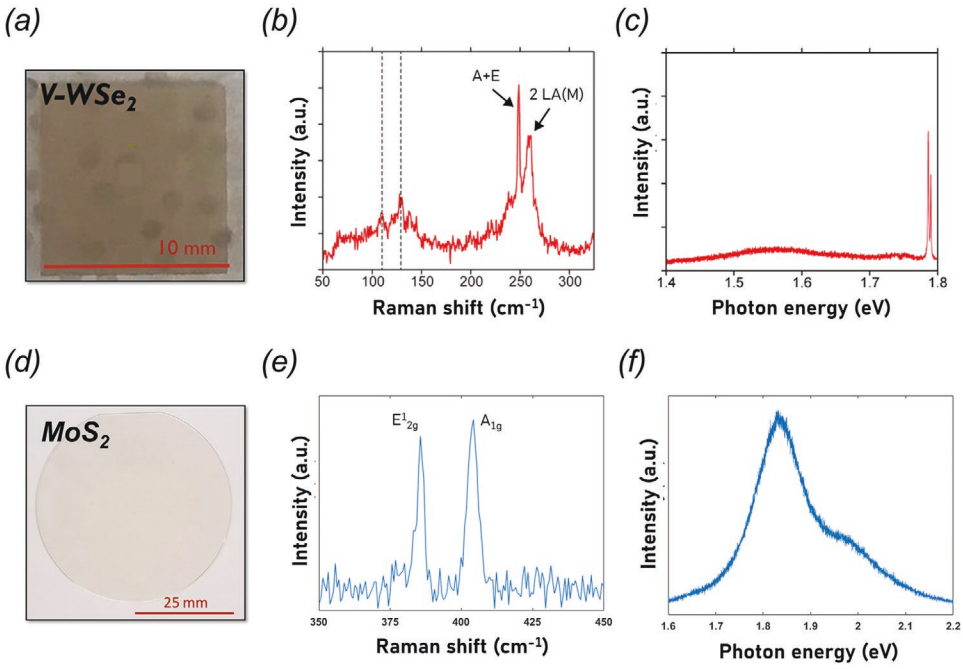


Figure 1. Large area p-type V-WSe_2 and n-type MoS_2. a) Optical image of MOCVD grown and fully coalesced monolayer V-WSe_2 on c-plane sapphire substrate. b) Raman and c) photoluminescence spectra of V-WSe_2. d) Optical image of MOCVD grown and fully coalesced monolayer MoS_2 on *c*-plane sapphire substrate. e) Raman and f) photoluminescence spectra of MoS_2.

and hydrogen sulfide (H_2S) were used as precursors. $Mo(CO)_6$ maintained at 10 °C and 650 Torr in a stainless-steel bubbler was used to deliver 1.1×10^{-3} sccm of the metal precursor for the growth, while 400 sccm of H_2S was used for the process. MoS_2 deposition was carried out at 950 °C and 50 Torr in H_2 ambient, where monolayer growth was achieved in 18 min. The substrate was first heated to 950 °C in H_2 and maintained for 10 min before the growth was initiated. After growth, the substrate was cooled in H_2S to 300 °C to inhibit decomposition of the MoS_2 films. More details can be found in our earlier work.[23,36,46] Figure 1d shows an optical image of fully coalesced monolayer MoS_2 on the growth substrate. Figure 1e shows the Raman spectra obtained for monolayer MoS_2 with two characteristics peaks at 383 and 404 cm^{-1} corresponding to the in-plane E^1_{2g} and out-of-plane A_{1g} modes, respectively, with the expected monolayer peak separation of ≈19 cm^{-1}.[47] Figure 1f shows the photoluminescence (PL) spectra with peak at 1.83 eV confirming monolayer MoS_2.

3. Integration of p-Type V-WSe_2 and n-Type MoS_2

Circuits based on complementary TMD materials were fabricated on commercially purchased 285 nm SiO_2/p^{++}-Si substrate. **Figure 2**a summarizes the entire fabrication process flow that allows precise positioning of individually gated p-type and n-type 2D FETs on the SiO_2/p^{++}-Si substrate. The first step was to fabricate local back gate islands. For this, electron-beam (e-beam) lithography was employed to define 250 µm × 250 µm squares, followed by sputtering of 20 nm TiN and 50 nm Pt. Platinum was selected as the back gate contact due to its large work function. TiN layer was deposited to promote adhesion between Pt and SiO_2. Following the back gate contact placement, 50 nm Al_2O_3 was deposited using atomic layer deposition (ALD) to serve as the back gate oxide. The oxide capacitance was found to be ≈1.8 pF with a capacitance–voltage (*C*–*V*) measurement done on a metal–insulator–metal (MIM) structure, as shown in Figure S2, Supporting Information. The dielectric constant of the ALD grown 50 nm Al_2O_3 back oxide was derived to be ≈8.22. Optical lithography was then employed to define regions where the platinum gate electrode can be accessed, followed by a reactive ion etching (RIE) step using BCl_3 gas to etch the exposed alumina. The etching was done in four cycles of 20 s exposure to avoid substrate heating effects.

After the fabrication of the Al_2O_3/Pt/TiN back-gate island stack, V-WSe_2 was transferred from the growth substrate to the island substrate using a PMMA (polymethyl-methacrylate) assisted wet transfer technique described elsewhere[48] and in the Experimental Section. Although, any other transfer techniques that avoid mechanical damage to the film and ensure clean post-transfer interface can be employed.[49] RIE using sulfur hexafluoride (SF_6) gas at 5 °C for 30 s was then used to isolate 5 µm × 20 µm V-WSe_2 strips, which were defined using e-beam lithography. These isolated V-WSe_2 films lie on top of the islands, which are required to serve as p-type FETs. Note that, in this way, p-type FETs can be fabricated anywhere in the substrate, thereby allowing the design of any complicated CMOS circuit. Finally, the p-type FET fabrication was completed with the e-beam evaporation of 40 nm Ni/30 nm Au contact pads which were defined using e-beam lithography. The channel length and width for all p-type V-WSe_2 FETs were designed to be 1 and 5 µm, respectively.

Next, another e-beam lithography step was performed followed by 30 nm alumina (Al_2O_3) evaporation to encapsulate

© 2022 Wiley-VCH GmbH

16136829, 2022, 33, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/smll.202202590 by Boise State University Albertsons Library, Wiley Online Library on [30/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

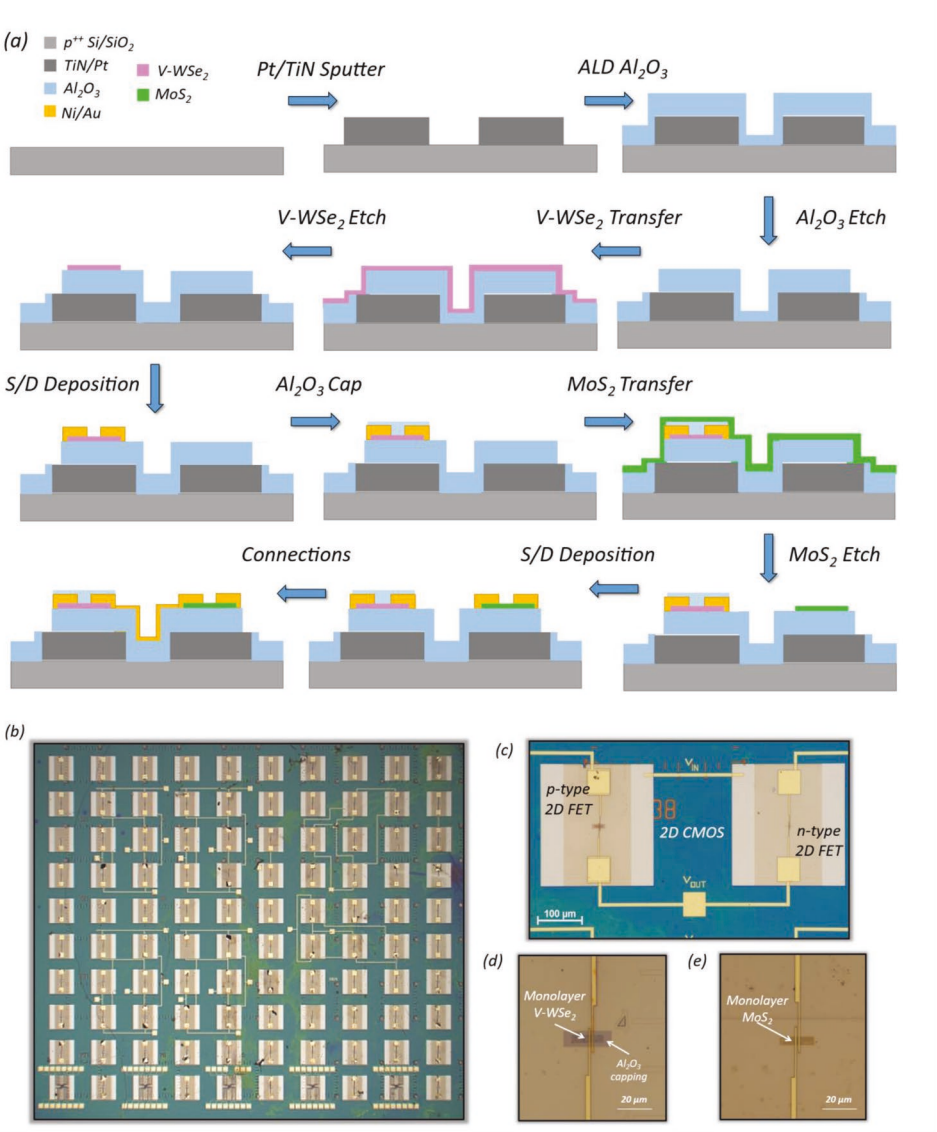


Figure 2. Integration of p-type V-WSe_2 and n-type MoS_2. a) Fabrication process flow for the realization of 2D CMOS based integrated circuits using n-type MoS_2 and p-type V-WSe_2. b) A representative optical image showing the layout of the chip. c) A 2D CMOS inverter. d) Optical image of a monolayer V-WSe_2 FET with Al_2O_3 capping layer protecting the underlying material from subsequent fabrication steps. e) Optical image of a monolayer MoS_2 FET.

all p-type FETs in order to protect them from subsequent fabrication steps. MOCVD MoS_2 film was then transferred on to the substrate using the same wet transfer technique, which was used to transfer V-WSe_2. The same fabrication steps used for making p-type FETs were then repeated to fabricate MoS_2 based n-type FETs on the required islands. The channel length and width for all n-type MoS_2 FETs were also designed to be 1 and 5 µm, respectively. Finally, necessary circuit connections between p-type FETs and n-type FETs were made by defining 7 µm wide lines using e-beam lithography followed by 60 nm Ni/ 30 nm Au evaporated using e-beam deposition. Figures 2b, 2c, and 2d, respectively, show the optical image of the fully fabricated 2D CMOS chip, a representative circuit, and individual n-type MoS_2 and p-type V-WSe2 FETs. Note that although we have used SiO_2/Si substrate in this work, the device fabrication flow outlined here is applicable to a variety of nonconducting substrates as each island is individually gated.

4. Non–von Neumann 2D CMOS

Electrical characterization on both p-type and n-type FETs were performed before depositing the connecting lines necessary for circuit demonstrations. **Figures 3**a,b show the transfer characteristics, that is, source-to-drain current (I_{DS}) plotted against the back gate voltage (V_{BG}) under various drain-to-source voltages (V_{DS}), and output characteristics, that is, I_{DS} plotted as a function of V_{DS} for different V_{BG}, respectively, for a representative p-type FET based on V-WSe_2. The same two plots for a MoS_2 based n-type FET are shown in Figure 3c,d. The p-type V-WSe_2

© 2022 Wiley-VCH GmbH

16136829, 2022, 33, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/smll.202202590 by Boise State University Albertsons Library, Wiley Online Library on [30/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

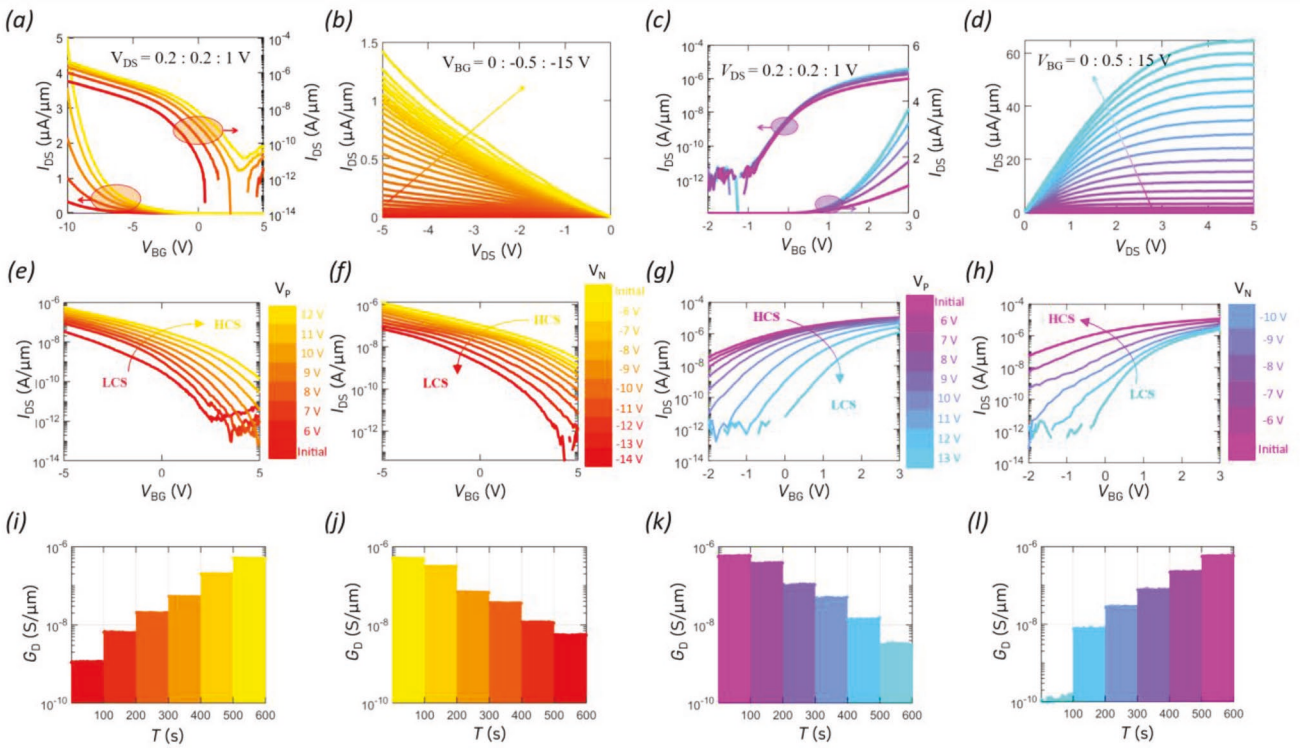


Figure 3. Electrical characterization of non–von Neumann 2D CMOS. Non–von Neumann 2D CMOS. a) Transfer characteristics, that is, source-to-drain current (I_{DS}) plotted against the back gate voltage (V_{BG}) under various drain-to-source voltages (V_{DS}) and b) output characteristics, that is, I_{DS} versus V_{DS} for different V_{BG} for a representative p-type FET based on monolayer V-WSe_2. c) Transfer characteristics and d) output characteristics for a representative n-type FET based on monolayer MoS_2. Both p- and n-type 2D FETs have 1 µm channel length and 5 µm channel width. Programming of V-WSe_2 FET using e) large positive (V_P) and f) large negative (V_N) voltage pulses of different amplitudes. Programming of MoS_2 FET using g) large positive (V_P) and h) large negative (V_N) voltage pulses of different amplitudes. Each voltage pulse was applied for a time period of 100 ms. Non-volatile retention of i) post-programmed and j) post-erased conductance states for V-WSe_2 FET. Non-volatile retention of k) post-programmed and l) post-erased conductance states for MoS_2 FET.

FET exhibits current on/off ratio ($r_{ON/OFF}$) ≈ 10^5 and subthreshold slope (SS) ≈ 470 mV decade^{-1}. The threshold voltage (V_{TH-p}) was found to be ≈ −3 V extracted at iso-current of 10 nA µm^{-1} and the hole field effect mobility (μ_p) extracted from the peak trans-conductance was found to be ≈ 2 cm^2 V^{-1} s^{-1}. The on current (I_p) reached ≈ 1.5 µA µm^{-1} for an inversion carrier density of ≈1.5 × 10^{13} cm^{-2} at V_{DS} = 5 V. The n-type MoS_2 FET exhibits $r_{ON/OFF}$ ≈ 10^7, SS ≈ 450 mV decade^{-1}, threshold voltage (V_{TH-n}) of ≈ 2.8 V and the electron field effect mobility (μ_n) of ≈ 17 cm^2 V^{-1} s^{-1}. The on current (I_{On}) reached as high as ≈ 65 µA µm^{-1} for an inversion carrier density of ≈1.5 × 10^{13} cm^{-2} at V_{DS} = 5 V. Note that the n-type MoS_2 FETs reach higher on-current compared to V-WSe_2 FETs for similar inversion charge carrier density. This can be attributed to the difference in the mobility values as well as low Schottky barrier (SB) height at Ni/MoS_2 interface compared to Ni/V-WSe_2 interface leading to significant difference in the contact resistance as evident from their respective output characteristics. Ideally, low, and high work function metals should be the preferred contact metal choices for n-type and p-type FETs, respectively. However, metal contacts are known to have their fermi level pinned differently for different 2D semiconductors.[50] For example, most metals including Sc, Ni, Ti, and Pt pin near the conduction band for MoS_2, irrespective of their work function, making it easier to observe n-type transport. However, in the case of WSe_2, most metals pin mid band gap resulting in ambipolar transport, that is, presence of both electron and hole conduction at the cost of reduced ON-state current. When it comes to fabrication, different contact metals can be chosen for V-WSe_2 and MoS_2 FETs since they are deposited in separate lithography steps. Therefore further improvements in the hole current for V-WSe_2 can be achieved by using a high work function metal that pins close to the valence band of V-WSe_2, such as palladium,[51] which will be explored in our future work.

Figures S3a and S3b, Supporting Information, show the device-to-device variation in the transfer characteristics across 50 p-type V-WSe_2 and 50 n-type MoS_2 FETs, respectively. Although V-WSe_2 demonstrates some device-to-device variability, MoS_2 FETs exhibit minimal variation. Figures S4a and S4b, Supporting Information, show that each set of 50 V-WSe_2 and MoS_2 FETs, respectively, exhibit minimal gate leakage. In addition, gate hysteresis, shown in Figures S5a and S5b, Supporting Information, is measured for 20 V-WSe_2 and MoS_2 FETs, respectively. Although there is very little hysteresis in the MoS_2 FETs, the V-WSe_2 shows more noticeable hysteresis window. This shows that more optimization is required for V-WSe_2 synthesis since every other fabrication step is same as that of MoS_2 FET fabrication. Hence V-WSe_2 synthesis will be one of the major focus in our future studies. These results indicate a relatively uniform monolayer MOCVD film growth, damage-free film transfer, and clean 2D FET fabrication processing.

© 2022 Wiley-VCH GmbH

16136829, 2022, 33, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/smll.202202590 by Boise State University Albertsons Library, Wiley Online Library on [30/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

16136829, 2022, 33, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/smll.202202590 by Boise State University Albertsons Library, Wiley Online Library on [30/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2D FETs characterized above also exhibit analog programming and non-volatile retention of the conductance states as shown in Figure 3e–h. This capability can be attributed to the phenomena of charge trapping/detrapping at/near the dielectric/2D interface. Application of high positive programming voltage pulses (V_P) to the local back gate of these devices shifts the threshold voltage of both devices to the positive side. The conductance measured at a given V_{BG} will increase for the p-type FET and decrease for the n-type FET for large positive V_P values, as shown in Figures 3e and g, respectively. Similarly, application of high negative voltage pulses (V_N) will lead to a threshold voltage shift to the negative side, causing a change from high conductance state to low conductance state for the p-type FET and vice versa for the n-type FET, as shown in Figures 3f and h, respectively. For each curve in Figure 3e–h, each voltage pulse was applied for a time period of $\tau_{P/E}$ = 100 ms. The memory retention at each of these programmed conductance states were found to be stable for both types of devices, irrespective of the magnitude and duration of the programming voltage pulses used. Figure 3i–l shows the evolution of different post-programmed and post-erased conductance states (G_{DS}) for 100 s extracted at a read voltage of V_{BG} = V_R = 0 V for p-type V-WSe_2 and n-type MoS_2 FETs, respectively. Figures S6a and S6b, Supporting Information, show the retention of the low and high conductance states measured for 1 h, for both V-WSe_2 and MoS_2 devices, respectively, under the same measurement conditions. Note that, while conventional memory devices require non-volatile retention for years, many in-memory compute applications including neuromorphic and edge computing relax the requirement for long-term retention and can be well served with short-term memory retention of several hours to days. Nevertheless, the key achievement here is the ability to program and store analog conductance states in both p-type and n-type FETs enabling non–von Neumann CMOS for the first time. Also note that programmable 2D FETs or memtransistors can be used as artificial synapses for acceleration of various neural networks and brain-inspired computations as well as hardware security applications.[52–58]

5. 2D CMOS-Based In-Memory Compute Primitives

Using the non-volatile programming capability of 2D FETs described above, here we show both digital and neuromorphic computing primitives that are reconfigurable. To do so, electrical connections were established between the p-type V-WSe_2 and n-type MoS_2 FETs as described in the fabrication process flow as the final lithography step.

Figure 4a shows the circuit schematic for a 2D CMOS inverter. Inverters are core logic elements in IC technology and require both p-type and n-type transistors to achieve high gain. Although inverters based on only n-type MoS_2 FETs with resistive-load have been demonstrated,[24,38,59] this design is known to suffer from higher power consumption and lower noise margin. Figure 4b shows the inverter characteristics, that is, the output voltage (V_{out}) plotted against the input voltage (V_{in}). The gain of the inverter was found to be ≈ 4. While inverters are useful logic elements, programmable inverters are of high interest for both digital and analog circuit designs. For example, a programmable inverter can replace a voltage comparator, which typically consists of an operational amplifier and analog to digital converter to determine if a signal is above or below a given reference level. An inverter with programmable switching threshold can set the voltage reference level based on the application and environmental requirements and thereby reduce the circuit footprint and the energy expenditure.[60] Figure 4c shows that the switching threshold (V_{SW}) of the 2D CMOS inverter, which is defined as the magnitude of V_{in} at which V_{out} reaches $V_{DD}/2$ can be adjusted to desired reference level exploiting the non-volatile and analog programming capability of our hardware platform.

Figure 4d shows the circuit schematic for a 2 × 1 MUX. A MUX is a common electronic device used to selectively switch between various inputs. In the case of a 2 × 1 MUX, the selector line, S, can be used to switch between two inputs, A and B. Note that we have used a simplified MUX design by connecting the p-type V-WSe_2 FET in series with n-type MoS_2 FET. The local back gates of the two devices are shorted and connected to the selector line, whereas the inputs A and B are applied to the drain terminal of the p-type FET and source terminal of the n-type FET, respectively. Ideally, when S = 0, input A should be reflected at the output (V_{out}), and V_{out} should be input B when S = 1. Figure 4e shows the waveforms of selector S, and inputs A and B and Figure 4f shows the V_{out}. Clearly, our 2 transistor circuit can be used as a 2 × 1 MUX reducing energy and area overhead.

Next, reconfigurable neuromorphic primitives such as sigmoid, hyperbolic tangent (tanh), and Gaussian activation functions are realized. Note that activation functions mimic neurons inside the brain and are as important as synapse that store the connection strength between the neurons. Figure 4g shows the circuit schematic comprising of series connection of p-type V-WSe_2 FET and n-type MoS_2 FET to achieve the sigmoid activation function. The input (V_{in}) is applied to the common back gate terminal of the two devices, whereas V_{DD} and GND are applied to the source and drain terminals of the n-type and p-type FETs, respectively. Figure 4h shows the output voltage (V_{out}) measured at the common node terminal of the n-type and p-type FETs as a function of V_{in}, which resembles the sigmoid activation function. Similar to the case of an inverter, the sigmoid activation curve can also be programmed to different thresholds by programming the p-type and n-type FETs, accordingly. Various sigmoid curves with different switching thresholds were measured and are shown in Figure 4i. To the same sigmoid circuit design, by applying - V_{DD} at drain terminal of the p-type FET, as shown in Figure 4j, a hyperbolic tangent (tanh) curve was obtained as shown in Figure 4k. Here, the function switches from –1 to +1. Similar to the above cases, this function can also be programmed to different states as shown in Figure 4l.

Finally, the Gaussian activation function is achieved using the circuit design shown in Figure 4m. Here, the output current (I_{out}) is measured, which follows a Gaussian curve as the input voltage (V_{in}), which is applied to the common back gate terminal of the p- and n-type FETs is swept. The emergence of Gaussian function shown in Figure 4n can be explained from the fact that, when V_{in} values are large positive or large negative one of the two FETs connected in series is off preventing any current to flow in the circuit. However, as the n-type

 © 2022 Wiley-VCH GmbH

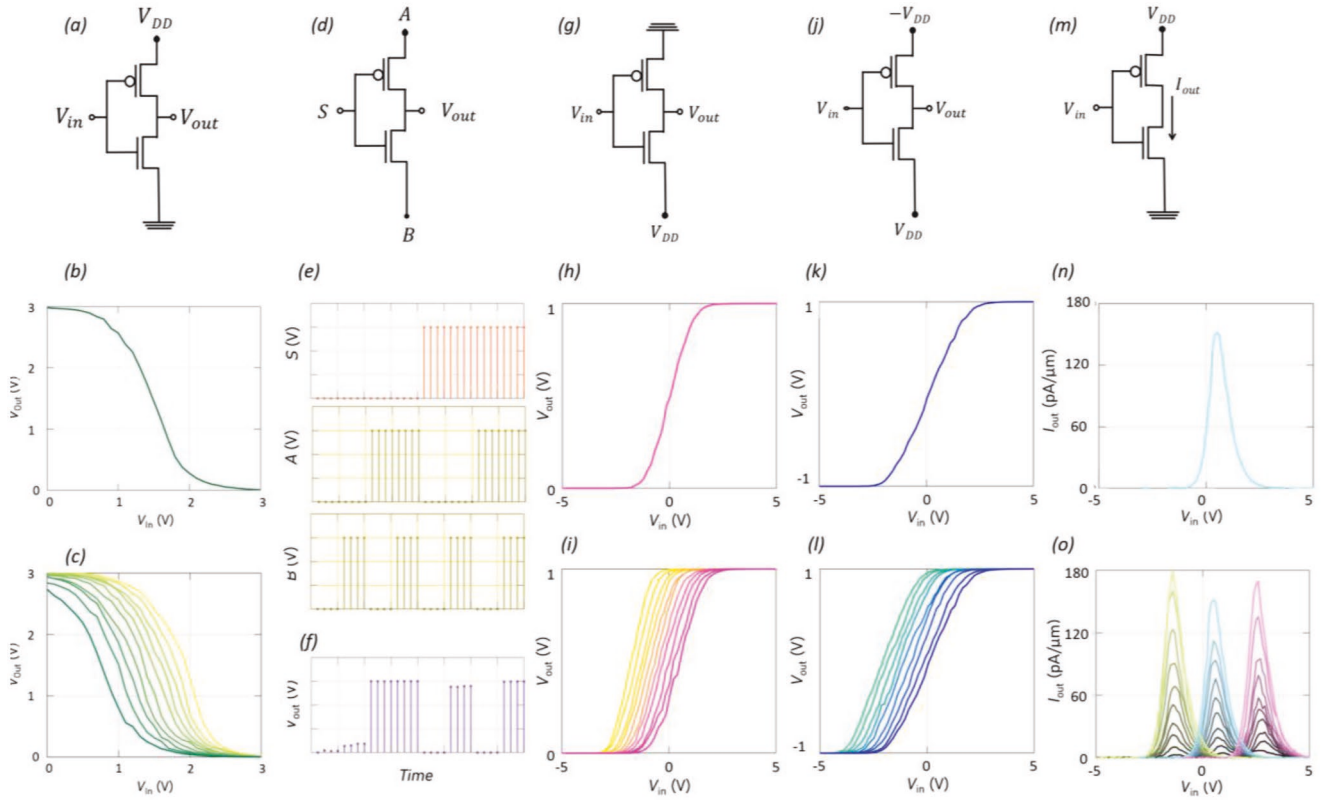


Figure 4. 2D CMOS based in-memory compute primitives. 2D CMOS based in-memory compute primitives. a) Circuit schematic, b) output (V_{out}) versus input (V_{in}) characteristics, and c) reconfigurable switching threshold (V_{SW}) for a 2D CMOS based programmable inverter. d) Circuit schematic; e) voltage waveforms applied to the selector, S, and inputs, A and B; and f) output voltage (V_{out}) waveform of a 2 × 1 MUX. g) Circuit schematic, h) output (V_{out}) versus input (V_{in}) characteristics, and i) reconfigurable sigmoid activation function realized using programmable 2D CMOS. j) Circuit schematic, k) output (V_{out}) versus input (V_{in}) characteristics, and l) reconfigurable hyperbolic tangent activation function realized using programmable 2D CMOS. m) Circuit schematic, n) output (I_{out}) versus input (V_{in}) characteristics, and o) reconfigurable Gaussian activation function realized using programmable 2D CMOS.

MoS_2 FET switches from off state to on state, current conduction begins and increases exponentially with V_{in} following the subthreshold characteristics and reaches its peak magnitude determined by the V_{DD}. Beyond this peak, the current starts to decrease exponentially following the subthreshold characteristics of the p-type V-WSe_2 FET. As a result, the series connection of the p-type V-WSe_2 and n-type MoS_2 FETs exhibits non-monotonic transfer characteristics with exponential tails that mimics a Gaussian distribution. Nevertheless, the amplitude of Gaussians can be modified by changing the applied $V_{DD,}$ while the mean can be tuned by programming the underlying p- and n-type FETs, accordingly. Gaussian curves, that is, I_{out} plotted as a function of V_{In}, with varying amplitudes and means are shown in Figure 4o. Gaussian functions are useful for Bayesian networks and probabilistic computing.[61]

6. Conclusion

In conclusion, we have demonstrated heterogeneous integration of p-type V-WSe_2 and n-type MoS_2 based programmable FETs for the realization of various digital and neuromorphic computing primitives. Our results are the first step toward the realization of 2D CMOS based integrated circuits although some challenges remain to be mitigated. For example, improvements in the synthesis of p-type V-WSe_2 can lead to better device and circuit performance, longevity, and yield. In addition, metal work function engineering can be exploited for further performance improvement for both n- and p-type 2D FETs. Our future work will also focus on the development of more complex 2D CMOS circuits by proposing solutions at the material, device, and circuit level. Optimization in the material synthesis, development and integration of a floating gate, and contact engineering are all avenues that will be explored to realize VLSI of 2D TMDs.

7. Experimental Section

Film Transfer to Local Back-Gate Islands: To fabricate the 2D FETs, MOCVD grown monolayer films were transferred from the sapphire to SiO_2/p^{++}-Si substrate with local back-gate islands using PMMA (polymethyl-methacrylate) assisted wet transfer process. First, 2D film 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 caused the NaOH to be drawn into the substrate/film

© 2022 Wiley-VCH GmbH

16136829, 2022, 33, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/smll.202202590 by Boise State University Albertsons Library, Wiley Online Library on [30/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

interface, separating the PMMA/2D film from the sapphire substrate. The separated film was rinsed multiple times inside a water bath and finally transferred onto the SiO_2/p^{++}-Si substrate with local back-gate islands and then baked at 50 °C and 70 °C for 10 min to anneal the film to the substrate. The PMMA layer is then removed by placing the sample in an acetone bath for 10 min, followed by an IPA bath for 10 min to clean the sample.

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

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

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

This work was supported by Army Research Office (ARO) through Contract Number W911NF1920338. The 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 and DMR-2039351. The work was also supported by the National Science Foundation (NSF) through a CAREER award under grant ECCS-2042145.

Conflict of Interest

The authors declare no conflict of interest.

Authors Contribution

R.P. and D.J. contributed equally to this work. S.D. conceived the idea and designed the experiments. R.P. and D.J. fabricated the CMOS chip, performed the electrical measurements, and analyzed the data. A.K. grew and characterized MOCVD V-WSe_2. N.T. grew and characterized MOCVD MoS_2. All authors contributed to the preparation of the manuscript.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords

complementary logic, field-effect transistors, heterogeneous integration, integrated circuits, two-dimensional materials

Received: May 28, 2022
Revised: June 21, 2022
Published online: July 17, 2022

[1] G. E. Moore, *Proc. IEEE* **1998**, *86*, 82.
[2] S. Salahuddin, K. Ni, S. Datta, *Nat. Electron.* **2018**, *1*, 442.
[3] A. P. Jacob, R. Xie, M. G. Sung, L. Liebmann, R. T. P. Lee, B. Taylor, *Int. J. High Speed Electro. Syst.* **2017**, *26*, 1740001.
[4] K. Uchida, H. Watanabe, A. Kinoshita, J. Koga, T. Numata, S. Takagi, in *Int. Electron Devices Meeting*, IEEE, Piscataway, NJ **2002**, pp. 47–50.
[5] A. Sebastian, M. Le Gallo, R. Khaddam-Aljameh, E. Eleftheriou, *Nat. Nanotechnol.* **2020**, *15*, 529.
[6] IEEE, *International Roadmap for Devices and Systems (IRDS)*, IEEE, Piscataway, NJ **2020**.
[7] S. Manzeli, D. Ovchinnikov, D. Pasquier, O. V. Yazyev, A. Kis, *Nat. Rev. Mater.* **2017**, *2*, 17033.
[8] D. Akinwande, C. Huyghebaert, C.-H. Wang, M. I. Serna, S. Goossens, L.-J. Li, H.-S. P Wong, F. H. L. Koppens, *Nature* **2019**, *573*, 507.
[9] M. Chhowalla, D. Jena, H. Zhang, *Nat. Rev. Mater.* **2016**, *1*, 16052.
[10] F. Schwierz, J. Pezoldt, R. Granzner, *Nanoscale* **2015**, *7*, 8261.
[11] C. Liu, H. Chen, S. Wang, Qi Liu, Yu-G Jiang, D. W. Zhang, M. Liu, P. Zhou, *Nat. Nanotechnol.* **2020**, *15*, 545.
[12] G. Iannaccone, F. Bonaccorso, L. Colombo, G. Fiori, *Nat. Nanotechnol.* **2018**, *13*, 183.
[13] Y. Liu, X. Duan, H.-J. Shin, S. Park, Yu Huang, X. Duan, *Nature* **2021**, *591*, 43.
[14] S. Das, A. Sebastian, E. Pop, C. J. McClellan, A. D. Franklin, T. Grasser, T. Knobloch, Y. Illarionov, A. V. Penumatcha, J. Appenzeller, Z. Chen, W. Zhu, I. Asselberghs, L.-J. Li, U. E. Avci, N. Bhat, T. D. Anthopoulos, R. Singh, *Nat. Electron.* **2021**, *4*, 786.
[15] Q. H. Wang, K. Kalantar-Zadeh, A. Kis, J. N. Coleman, M. S. Strano, *Nat. Nanotechnol.* **2012**, *7*, 699.
[16] B. Radisavljevic, A. Radenovic, J. Brivio, V. Giacometti, A. Kis, *Nat. Nanotechnol.* **2011**, *6*, 147.
[17] K. Kang, S. Xie, L. Huang, Y. Han, P. Y. Huang, K. F. Mak, C.-J. Kim, D. Muller, J. Park, *Nature* **2015**, *520*, 656.
[18] Q. Smets, G. Arutchelvan, J. Jussot, D. Verreck, I. Asselberghs, A. N. Mehta, A. Gaur, D. Lin, S. El Kazzi, B. Groven, M. Caymax, I. Radu, in *2019 IEEE Int. Electron Devices Meeting (IEDM)*, IEEE, Piscataway, NJ **2019**, pp. 23.2.1–23.2.4.
[19] P.-C. Shen, C. Su, Y. Lin, A.-S. Chou, C.-C. Cheng, J.-H. Park, M.-H. Chiu, A.-Y. Lu, H.-L. Tang, M. M. Tavakoli, G. Pitner, X. Ji, Z. Cai, N. Mao, J. Wang, V. Tung, J. Li, J. Bokor, A. Zettl, C.-I. Wu, T. S. Palacios, L.-J. Li, J. Kong, *Nature* **2021**, *593*, 211.
[20] K. M. Price, K. E. Schauble, F. A. Mcguire, D. B. Farmer, A. D. Franklin, *ACS Appl. Mater. Interfaces* **2017**, *9*, 23072.
[21] I. Asselberghs, Q. Smets, T. Schram, B. Groven, D. Verreck, A. Afzalian, G. Arutchelvan, A. Gaur, D. Cott, T. Maurice, S. Brems, K. Kennes, A. Phommahaxay, E. Dupuy, D. Radisic, J.-F. de Marneffe, A. Thiam, W. Li, K. Devriendt, C. Huyghebaert, D. Lin, M. Caymax, P. Morin, I. P. Radu, in *2020 IEEE International Electron Devices Meeting (IEDM)*, IEEE, Piscataway, NJ **2020**, pp. 40.2.1–40.2.4.
[22] K. K. H. Smithe, S. V. Suryavanshi, M. Muã±Oz Rojo, A. D. Tedjarati, E. Pop, *ACS Nano* **2017**, *11*, 8456.
[23] A. Sebastian, R. Pendurthi, T. H. Choudhury, J. M. Redwing, S. Das, *Nat. Commun.* **2021**, *12*, 693.
[24] S. Wachter, D. K. Polyushkin, O. Bethge, T. Mueller, *Nat. Commun.* **2017**, *8*, 14948.
[25] D. K. Polyushkin, S. Wachter, L. Mennel, M. Paur, M. Paliy, G. Iannaccone, G. Fiori, D. Neumaier, B. Canto, T. Mueller, *Nat. Electron.* **2020**, *3*, 486.
[26] Q. Gao, Z. Zhang, X. Xu, J. Song, X. Li, Y. Wu, *Nat. Commun.* **2018**, *9*, 4778,.
[27] A. B. Sachid, M. Tosun, S. B. Desai, C.-Y. Hsu, D.-H. Lien, S. R. Madhvapathy, Y.-Z. Chen, M. Hettick, J. S. Kang, Y. Zeng, J.-H. He, E. Y. Chang, Y.-L. Chueh, A. Javey, C. Hu, *Adv. Mater.* **2016**, *28*, 2547.

© 2022 Wiley-VCH GmbH

16136829, 2022, 33, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/smll.202202590 by Boise State University Albertsons Library, Wiley Online Library on [30/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

[28] H. Chen, X. Xue, C. Liu, J. Fang, Z. Wang, J. Wang, D. W. Zhang, W. Hu, P. Zhou, *Nat. Electron.* **2021**, *4*, 399.
[29] F. Zhang, Y. Lu, D. S. Schulman, T. Zhang, K. Fujisawa, Z. Lin, Y. Lei, A. L. Elias, S. Das, S. B. Sinnott, M. Terrones, *Sci. Adv.* **2019**, *5*, eaav5003.
[30] S. M. Eichfeld, L. Hossain, Y.-C. Lin, A. F. Piasecki, B. Kupp, A. G. Birdwell, R. A. Burke, N. Lu, X. Peng, J. Li, A. Azcatl, S. Mcdonnell, R. M. Wallace, M. J. Kim, T. S. Mayer, J. M. Redwing, J. A. Robinson, *ACS Nano* **2015**, *9*, 2080.
[31] X. Zhang, F. Zhang, Y. Wang, D. S. Schulman, T. Zhang, A. Bansal, N. Alem, S. Das, V. H. Crespi, M. Terrones, J. M. Redwing, *ACS Nano* **2019**, *13*, 3341.
[32] A. Kozhakhmetov, S. Stolz, A. M. Z. Tan, R. Pendurthi, S. Bachu, F. Turker, N. Alem, J. Kachian, S. Das, R. G. Hennig, O. Grãning, B. Schuler, J. A. Robinson, *Adv. Funct. Mater.* **2021**, *31*, 2105252.
[33] P. Yao, H. Wu, B. Gao, J. Tang, Q. Zhang, W. Zhang, J. J. Yang, H. Qian, *Nature* **2020**, *577*, 641.
[34] H. S. Lee, V. K. Sangwan, W. A. G. Rojas, H. Bergeron, H. Y. Jeong, J. Yuan, K. Su, M. C. Hersam, *Adv. Funct. Mater.* **2020**, *30*, 2003683.
[35] V. K. Sangwan, H.-S. Lee, H. Bergeron, I. Balla, M. E. Beck, K.-S. Chen, M. C. Hersam, *Nature* **2018**, *554*, 500.
[36] D. Jayachandran, A. Oberoi, A. Sebastian, T. H. Choudhury, B. Shankar, J. M. Redwing, S. Das, *Nat. Electron.* **2020**, *3*, 646.
[37] C. Pan, C.-Y. Wang, S.-J. Liang, Y. Wang, T. Cao, P. Wang, C. Wang, S. Wang, B. Cheng, A. Gao, E. Liu, K. Watanabe, T. Taniguchi, F. Miao, *Nat. Electron.* **2020**, *3*, 383.
[38] G. M. Marega, Y. Zhao, A. Avsar, Z. Wang, M. Tripathi, A. Radenovic, A. Kis, *Nature* **2020**, *587*, 72.
[39] D. Li, M. Chen, Q. Zong, Z. Zhang, *Nano Lett.* **2017**, *17*, 6353.
[40] C. Liu, X. Yan, X. Song, S. Ding, D. W. Zhang, P. Zhou, *Nat. Nanotechnol.* **2018**, *13*, 404.
[41] H. Tian, B. Deng, M. L. Chin, X. Yan, H. Jiang, S.-J. Han, V. Sun, Q. Xia, M. Dubey, F. Xia, H. Wang, *ACS Nano* **2016**, *10*, 10428.
[42] C.-Y. Wang, S.-J. Liang, S. Wang, P. Wang, Z. A. Li, Z. Wang, A. Gao, C. Pan, C. Liu, J. Liu, H. Yang, X. Liu, W. Song, C. Wang, B. Cheng, X. Wang, K. Chen, Z. Wang, K. Watanabe, T. Taniguchi, J. J. Yang, F. Miao, *Sci. Adv.* **2020**, *6*, eaba6173.
[43] G. Wu, B. Tian, L. Liu, W. Lv, S. Wu, X. Wang, Y. Chen, J. Li, Z. Wang, S. Wu, H. Shen, T. Lin, P. Zhou, Qi Liu, C. Duan, S. Zhang, X. Meng, S. Wu, W. Hu, X. Wang, J. Chu, J. Wang, *Nat. Electron.* **2020**, *3*, 43.
[44] D. Xiang, T. Liu, J. Xu, J. Y. Tan, Z. Hu, B. Lei, Y. Zheng, J. Wu, A. H. C. Neto, L. Liu, W. Chen, *Nat. Commun.* **2018**, *9*, 2966.
[45] Y. Xuan, A. Jain, S. Zafar, R. Lotfi, N. Nayir, Y. Wang, T. H. Choudhury, S. Wright, J. Feraca, L. Rosenbaum, J. M. Redwing, V. Crespi, A. C. T. Van Duin, *J. Cryst. Growth* **2019**, *527*, 125247.
[46] A. Dodda, A. Oberoi, A. Sebastian, T. H. Choudhury, J. M. Redwing, S. Das, *Nat. Commun.* **2020**, *11*, 4406.
[47] H. Li, Q. Zhang, C. C. R. Yap, B. K. Tay, T. H. T. Edwin, A. Olivier, D. Baillargeat, *Adv. Funct. Mater.* **2012**, *22*, 1385.
[48] F. Zhang, C. Erb, L. Runkle, X. Zhang, N. Alem, *Nanotechnology* **2018**, *29*, 025602.
[49] T. F. Schranghamer, M. Sharma, R. Singh, S. Das, *Chem. Soc. Rev.* **2021**, *50*, 11032.
[50] D. S. Schulman, A. J. Arnold, S. Das, *Chem. Soc. Rev.* **2018**, *47*, 3037.
[51] S. Das, J. Appenzeller, *Appl. Phys. Lett.* **2013**, *103*, 103501.
[52] A. Sebastian, S. Das, S. Das, *Adv. Mater.* **2021**, *34*, 2107076.
[53] S. Subbulakshmi Radhakrishnan, S. Chakrabarti, D. Sen, M. Das, T. F. Schranghamer, A. Sebastian, S. Das, *Adv. Mater.* **2022**, https://doi.org/10.1002/adma.202202535.
[54] A. Dodda, N. Trainor, Redwing, M. Joan, S. Das, *Nat. Commun.* **2022**, *13*, https://doi.org/10.1038/s41467-022-31148-z.
[55] A. Oberoi, A. Dodda, H. Liu, M. Terrones, S. Das, *ACS Nano* **2021**, *15*, 19815.
[56] A. Dodda, S. Das, *ACS Nano* **2021**, *15*, 16172.
[57] A. Dodda, S. Subbulakshmi Radhakrishnan, T. F. Schranghamer, D. Buzzell, P. Sengupta, S. Das, *Nat. Electron.* **2021**, *4*, 364.
[58] T. F. Schranghamer, A. Oberoi, S. Das, *Nat. Commun.* **2020**, *11*, https://doi.org/10.1038/s41467-020-19203-z.
[59] H. Wang, L. Yu, Y.-H. Lee, Y. Shi, A. Hsu, M. L. Chin, L.-J. Li, M. Dubey, J. Kong, T. Palacios, *Nano Lett.* **2012**, *12*, 4674.
[60] J. Segura, J. L. Rossello, J. Morra, H. Sigg, *IEEE J. Solid-State Circuits* **1998**, *33*, 1262.
[61] A. Sebastian, A. Pannone, S. S. Radhakrishnan, S. Das, *Nat. Commun.* **2019**, *10*, 4199.

© 2022 Wiley-VCH GmbH

16136829, 2022, 33, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/smll.202202590 by Boise State University Albertsons Library, Wiley Online Library on [30/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Check for updates

ADVANCED MATERIALS
www.advmat.de

A Sparse and Spike-Timing-Based Adaptive Photoencoder for Augmenting Machine Vision for Spiking Neural Networks

*Shiva Subbulakshmi Radhakrishnan, Shakya Chakrabarti, Dipanjan Sen, Mayukh Das, Thomas F. Schranghamer, Amritanand Sebastian, and Saptarshi Das**

The representation of external stimuli in the form of action potentials or spikes constitutes the basis of energy efficient neural computation that emerging spiking neural networks (SNNs) aspire to imitate. With recent evidence suggesting that information in the brain is more often represented by explicit firing times of the neurons rather than mean firing rates, it is imperative to develop novel hardware that can accelerate sparse and spike-timing-based encoding. Here a medium-scale integrated circuit composed of two cascaded three-stage inverters and one XOR logic gate fabricated using a total of 21 memtransistors based on photosensitive 2D monolayer MoS_2 for spike-timing-based encoding of visual information, is introduced. It is shown that different illumination intensities can be encoded into sparse spiking with time-to-first-spike representing the illumination information, that is, higher intensities invoke earlier spikes and vice versa. In addition, non-volatile and analog programmability in the photoencoder is exploited for adaptive photoencoding that allows expedited spiking under scotopic (low-light) and deferred spiking under photopic (bright-light) conditions, respectively. Finally, low energy expenditure of less than 1 µJ by the 2D-memtransistor-based photoencoder highlights the benefits of in-sensor and bioinspired design that can be transformative for the acceleration of SNNs.

1. Introduction

Spiking neural networks (SNNs) are the latest manifestation of artificial neural networks, promising to bridge the energy gap between natural and artificial intelligence by deriving inspiration from spike-based information encoding and processing found inside biological brains.[1–4] Spikes, also referred to as action potentials, are electrical impulses that biological neurons use to encode and communicate sensory information among each other over short and long distances, ultimately allowing the brain to learn, adapt, and make decisions with remarkable energy efficiency. However, doing so requires specialized sensory (afferent) neurons to convert visual, audio, tactile, gustatory, and olfactory stimuli, which are predominantly analog and continuous variable over time, into corresponding spike trains using the process of sensory transduction.[5–9]

Early studies based on neurophysiological recording of neurons alluded to the fact that, during sensory transduction, information about the stimulus is transcribed into the mean firing rate of the neurons and not into individual spikes.[10] Rate-based spike encoding is, therefore, the most widely used encoding algorithm adopted by SNNs. However, a minimum of two spikes are required to estimate spike rates using this method; for more reliable estimates, neural responses must be averaged over a larger number of spikes. As a result, rate-based spike encoding imposes severe constraints on rapid information processing. This lies in strict contrast to responses seen in animals surviving in the wild. For example, rhesus monkeys can identify faces within ≈80 to 160 ms[11] despite the neural pathway between the photoreceptors in the retina and visual cortex involving more than ten synaptic stages. This implies that each synaptic layer needs to process information within only 10 ms. Given that the firing rates of cortical neurons are in the range of 0–100 spikes-per-second, rate-based spike encoding fails to explain the behavior of visual animals. Instead, recent developments suggest that the precise timing of the spikes is more useful in encoding the information when a fast response is needed.[12–14] Such spike-timing-based encoding is not only fast but also extremely energy efficient since only one or few spikes are required to encode the information.[15–19] Therefore, hardware acceleration of more biorealistic and energy efficient SNNs will necessitate artificial afferent neurons which can mimic the spike-timing-based encoding found in sensory neurobiology.

S. Subbulakshmi Radhakrishnan, D. Sen, M. Das, T. F. Schranghamer, A. Sebastian, S. Das
Engineering Science and Mechanics
Penn State University
University Park, PA 16802, USA
E-mail: sud70@psu.edu
S. Chakrabarti, S. Das
Electrical Engineering and Computer Science
Penn State University
University Park, PA 16802, USA
S. Das
Materials Science and Engineering
Penn State University
University Park, PA 16802, USA
S. Das
Materials Research Institute
Penn State University
University Park, PA 16802, USA
The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/adma.202202535.

DOI: 10.1002/adma.202202535

 © 2022 Wiley-VCH GmbH

In the past, silicon (Si)-based complementary metal oxide semiconductor (CMOS) technology has been used to design circuits that can generate different biomimetic spiking profiles[20] following various models, such as the integrate-and-fire (IF) model,[21] the leaky integrate-and-fire model,[22] the Fitzhugh–Nagumo model,[23] and the Morris–Lecar model.[24] Among these, the IF and LIF models are the most popular for their relative ease of hardware implementation despite limited bioresemblance. In comparison, the Fitzhugh–Nagumo and Morris–Lecar models are more successful at emulating non-linear neural dynamics but require extensive hardware investment, which significantly increases the necessary area and energy budgets.[25] Recently, memristor-based artificial neurons and synapses have received significant attention for SNN acceleration.[26–30] However, memristive devices are yet to demonstrate spike encoding of analog input signals. Memristive devices mostly use software-defined pulses or rely on CMOS-based peripheral hardware to encode analog information, which are subsequently used to adjust the synaptic weights of the memristive crossbar arrays.[31] Nevertheless, neither the state-of-the-art CMOS technology nor the emerging memristive devices offer a solution for in-sensor encoding, that is, direct conversion of external stimuli into sparse and adaptive spike-trains exploiting biorealistic and energy efficient spike-timing-based encoding.

Here, we introduce a reconfigurable optoelectronic circuit based on medium scale integration (MSI) of photosensitiv 2D memtransistors for in-sensor and adaptive encoding of visual stimuli into corresponding electrical spikes following the spike-timing-based encoding algorithm. Our photoencoder is based on monolithic integration of two cascaded three-stage inverters (TSIs) and one XOR logic gate fabricated using a total of 21 2D-memtransistors. The optoelectronic circuit exploits the gate-tunable persistent photoconductivity and non-volatile electrical programmability of our 2D memtransistor technology for transcribing visual stimuli into sparse spiking, with the information in the illumination levels being encoded in the time-to-first-spike. The photoencoder can also adapt to scotopic (low-light) and photopic (bright-light) conditions by expediting or deferring the spiking, respectively. Finally, our design has a small footprint with an active device area of ≈100 μm^2 and consumes a miniscule energy of less than 1 μJ to accomplish its adaptive photoencoding functionality. We believe that our demonstration of energy and area efficient in-sensor and adaptive photoencoder can benefit the hardware acceleration of future SNNs.

2. Spike-Timing-Based Photoencoder Using 2D Memtransistors

Figures 1a and 1b, respectively, show the optical image and 3D schematic of a representative 2D memtransistor based on monolayer MoS_2; as with all devices used in this study, this device was locally back-gated using a stack composed of atomic layer deposition (ALD) grown 50 nm Al_2O_3 on sputter-deposited 50/20 nm Pt/TiN. All back-gate islands were placed on a commercially purchased SiO_2/p^{++}-Si substrate. As we will discuss later, the interface of monolayer MoS_2/Al_2O_3 plays a pivotal role in device performance, not only enhancing the optical response of the 2D memtransistors via the phenomenon of electrically tunable photogating effect but also enabling analog and non-volatile programming of the channel conductance via exploiting the trapping/de-trapping of charge carriers at and near the interface when subjected to gate biases of different polarities. The monolayer MoS_2 used in this study was grown using a metal–organic chemical vapor deposition (MOCVD) technique on an epitaxial sapphire substrate at relatively high temperature (950 °C);[32,33] this was done to ensure the growth of a uniform and high quality 2D film, which is important for the successful realization of 2D memtransistor-based MSI circuits for the acceleration of spike-timing-based photoencoding. The as-grown monolayer MoS_2 film was subsequently transferred from the growth substrate to the SiO_2/p^{++}-Si substrate with pre-defined islands of Al_2O_3/Pt/TiN for 2D memtransistor fabrication. Details on the monolayer MoS_2 synthesis, film transfer, and fabrication of the local back-gate islands, MoS_2 memtransistors, and MSI circuit can be found in the Experimental Section.

Figures 1c and 1d, respectively, show the circuit schematic and corresponding optical image of the bioinspired solid-state afferent neuron (spike-timing-based photoencoder) composed of two TSIs and one XOR logic gate. Note that the entire MSI circuit is constructed using a total of 21 2D memtransistors (MT_1–MT_{21}). The output voltages, V_A and V_B, from the two TSIs serve as inputs to the XOR gate and the encoded spike trains (V_C) are obtained at the output of the XOR gate. A clock signal toggling between two pre-defined levels is applied to the input terminal (V_{in}) of both TSIs to exploit the photogating effect for the transduction of visual information into electrical impulses, as discussed later. Furthermore, the two TSIs are electrically pre-programmed with slightly different inverting thresholds (V_{SW}) such that there is a finite delay in the spiking profile of V_A and V_B under illumination, which, when passed through the XOR logic gate, results in one or few spikes in V_C. Figure 1e shows an example experimental demonstration of spike profiles for V_A, V_B, and V_C under dark and two different illumination levels. As expected, no spiking is seen in V_C under dark, whereas illumination leads to finite spiking. More interestingly, the arrival time of the first spike in V_C is found to be inversely proportional to the illumination intensity, that is, higher intensity invokes earlier spikes and vice versa. This clearly shows that our 21 memtransistor-based optoelectronic circuit is capable of spike-timing-based photoencoding. In the following sections, we will explain the functioning of the circuits and the underlying physical mechanisms involved.

3. Characterization of 2D Memtransistor and Reconfigurable Cascaded Three-Stage Inverters

First, the quality of the monolayer MoS_2 film was assessed using optical measurements. **Figure 2**a shows the Raman spectra obtained from the channel of a representative 2D memtransistor using a 532 nm laser; two characteristic monolayer MoS_2 peaks were noted at 386 and 403 cm^{-1}, corresponding to the in-plane E^1_{2g} and out-of-plane A_{1g} modes, respectively, with an expected peak separation of ≈17 cm^{-1}.[34] Figure 2b shows the photoluminescence (PL) spectra for the same channel with a

© 2022 Wiley-VCH GmbH

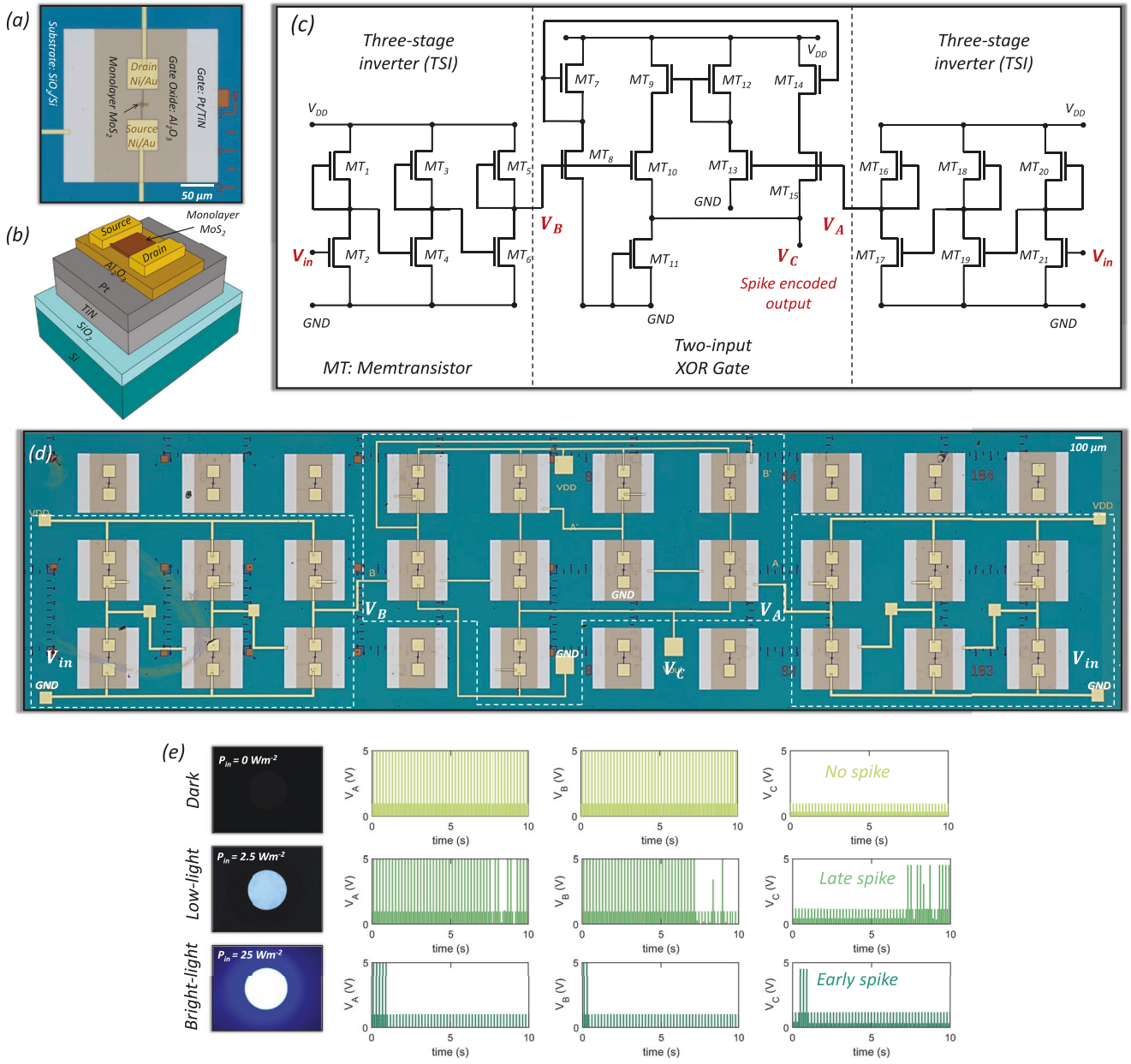


Figure 1. Bioinspired 2D afferent neuron for spike-timing-based photoencoding. a) Optical image and b) 3D schematic of a representative large-area grown monolayer MoS_2-based 2D memtransistor, which is locally back-gated using a stack comprising 50 nm Al_2O_3 on 40/30 nm Pt/TiN. All back-gate islands were placed on a commercially purchased SiO_2/Si substrate. c) Circuit schematic and d) optical image of our bioinspired solid-state afferent neuron (spike-timing-based photoencoder) comprising two cascaded three-stage inverters (TSIs) and one XOR logic gate fabricated using a total of 21 memtransistors. A clock signal toggling between two pre-defined levels is applied to the input terminal (V_{in}) of two pre-programmed cascaded inverters with slightly different switching thresholds to exploit the photogating effect and transduce the visual information into electrical impulses at the output of the TSIs, that is, V_A and V_B, with some lag. Since the outputs from the TSIs are fed as the inputs to the XOR gate, the spike-timing-based encoded information is obtained at the output of the XOR gate, that is, V_C. e) Example, experimental demonstration of spike-timing-based photoencoding under dark and two different illumination conditions and corresponding profiles for V_A, V_B, and V_C. No spikes appear under dark conditions, whereas brighter illumination leads to earlier spiking and weaker illumination invokes late spiking.

peak at 1.83 eV, another signature of monolayer MoS_2 owing to the indirect-to-direct bandgap transition at the monolayer limit.[35] Section S1, Supporting Information, shows the Raman and PL maps of the transferred monolayer MoS_2 film over a 10 μm × 10 μm area. Both the Raman peak separation and PL peak position were found to be uniform across the film area investigated. Following optical characterization, we performed electrical transport measurements to evaluate the performance of the 2D memtransistors. Figure 2c shows the transfer characteristics, that is, 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 2D memtransistor with channel length

© 2022 Wiley-VCH GmbH

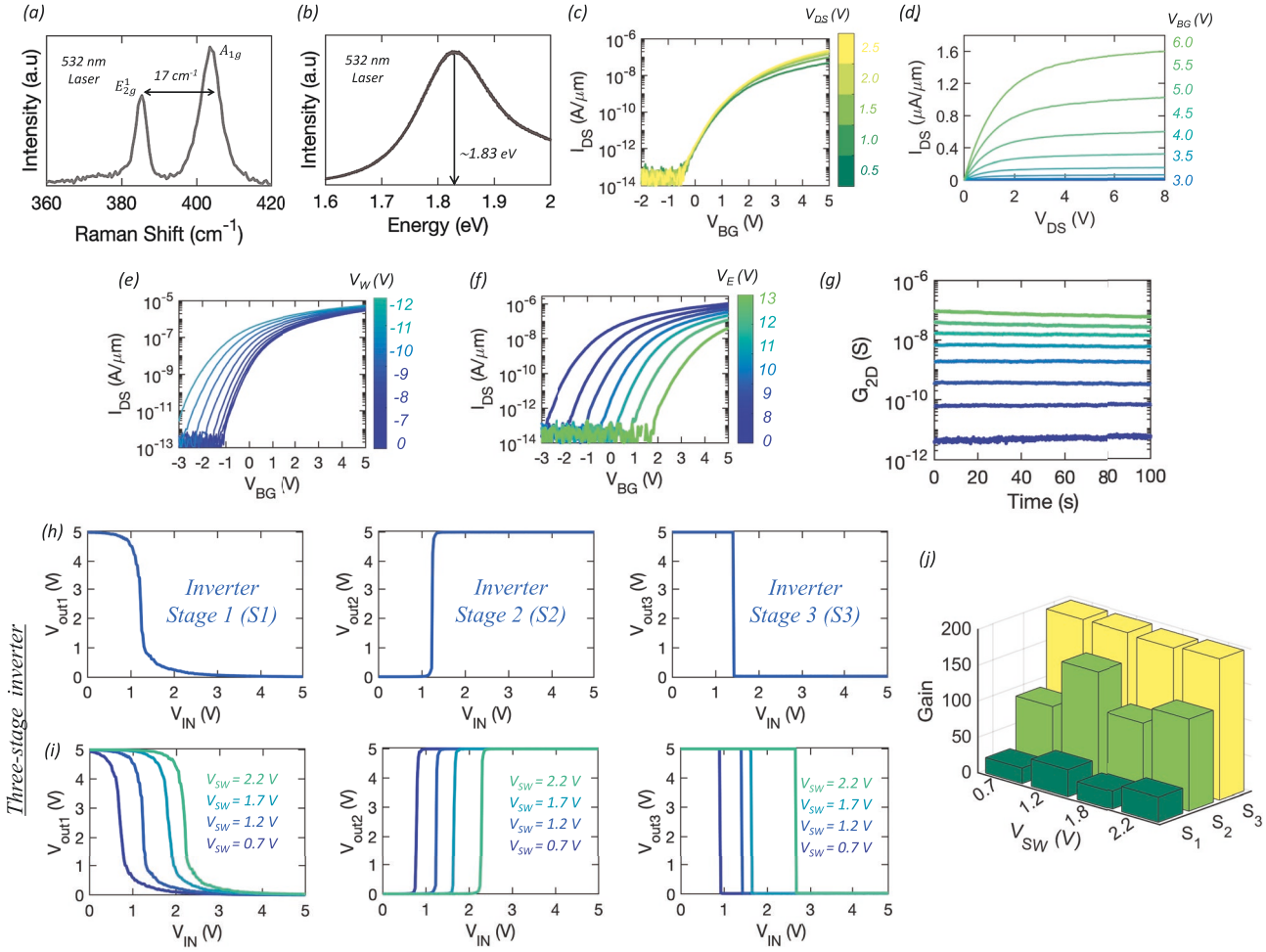


Figure 2. Characterization of monolayer MoS_2-based memtransistor and reconfigurable three-stage inverters (TSIs). a) Raman and b) photoluminescence (PL) spectra of MOCVD grown monolayer MoS_2. c) Transfer characteristics, that is, the source to drain current (I_{DS}) as a function of the local back-gate voltage (V_{BG}) at different drain biases (V_{DS}) and d) output characteristics, that is, I_{DS} versus V_{DS} for different V_{BG} for a monolayer MoS_2-based memtransistor with channel length, $L = 1$ µm, and channel width, $W = 5$ µm. e) Post-programmed and f) post-erased transfer characteristics of a monolayer MoS_2-based memtransistor when subjected to negative "Write" (V_W) and positive "Erase" (V_E) voltage pulses for a duration of 100 ms. g) Retention of post-programmed conductance states (G_{2D}) for 100 s. h) Output characteristics obtained from stage 1, 2, and 3 of a TSI when the input voltage (V_{in}) is swept from 0 to 5 V. i) Switching threshold (V_{SW}) can be adjusted for all three stages of the TSI by electrically programming one or more 2D memtransistors. j) Gain for all the stages of the TSI for different V_{SW}.

$L = 1$ µm and channel width $W = 5$ µm. Note that monolayer MoS_2 primarily shows n-type transport owing to the pinning of the metal Fermi level near the conduction band.[36–38] Also note that, unlike two-terminal memristors, three-terminal memtransistors allow for electrostatic gate control with current on/off ratio ($r_{ON/OFF}$) > 10^6, which is key toward the design of ICs based on 2D memtransistors. The subthreshold slope (SS) was found to be ≈400 mV decade^{-1} averaged over three orders of magnitude change in I_{DS}. The deviation of the SS from its ideal value of 60 mV decade^{-1} can be ascribed to the presence of interface trap states. 2D memtransistors also demonstrate minimal gate hysteresis when measured in air and low gate leakage current. The threshold voltage (V_{TH}) was found to be ≈2.3 V extracted at an isocurrent of 10 nA µm^{-1} and the electron field effect mobility (μ_{FE}) extracted from the peak transconductance was found to be ≈5.1 cm^2 V^{-1} s^{-1}. Figure 2d shows the output characteristics, that is, I_{DS} versus V_{DS} for different V_{BG}, for the same MoS_2 memtransistor. The on-current (I_{ON}) reached ≈1.5 µA µm^{-1} for an inversion carrier density of ≈10^{12} cm^{-2} at $V_{DS} = 5$ V. Section S2, Supporting Information, shows the transfer characteristics for 25 as-fabricated MoS_2 memtransistors; minimal variation was found between μ_{FE}, current on/off ratio ($r_{ON/OFF}$), V_{TH} extracted at isocurrent of 500 nA µm^{-1} and subthreshold slope (SS) over three orders of magnitude change in I_{DS}. Taken together, these results indicate that the MOCVD-grown films are of relatively high quality, that our film-transfer is relatively damage free, and that our memtransistor fabrication processes are sufficiently clean.

Next, we demonstrate the analog and non-volatile electrical programming capability of our 2D memtransistors. Figures 2e and 2f, respectively, show the post-programmed and post-erased transfer characteristics of a representative 2D memtransistor when subjected to negative "Write" (V_W) and positive "Erase" (V_E) voltage pulses of different amplitudes applied to the local

© 2022 Wiley-VCH GmbH

back-gate electrode, each for a duration of $\tau_{P/E}$ = 100 ms. The shift in the transfer characteristics can be attributed to charge trapping/de-trapping at and near the MoS_2/Al_2O_3 interface; negative and positive shifts in the transfer characteristics with increasing magnitude of V_W and V_E are indicative of electron trapping and de-trapping in the local back-gate stack, respectively. We confirmed the charge trapping in the oxide through bias temperature instability test as shown in Section S3, Supporting Information. Interestingly, the trapping and de-trapping processes were found to be non-volatile, as evident from the retention measurements displayed in Figure 2g for representative post-programmed conductance states (G_{2D}) for 100 s. Note that G_{2D} was extracted using V_{BG} = 0 V. We also examined long-term memory retention for two representative conductance states and found that the states were distinguishable even after ≈3 h. Note that the strict definition of the term memtransistors should refer to transistors with a clear physical picture of non-volatile charge storage, for example, ferroelectric transistors with non-volatile remnant polarization states or floating gate transistors with a floating gate for charge storage. This will ensure retention times in excess of 10 years. However, with the emergence of bioinspired and neuromorphic computing, the conventional definition of memory is relaxed. For example, the brain has both short-term and long-term memory. In fact, for many Internet of Things (IoT) edge devices that operate within tight energy budgets, one important trade-off arises from the fact that these devices may not require the high retention times usually guaranteed for traditional memory devices. It is in this context that we refer to our 2D transistors as memtransistors, as they allow us to harness charge trapping/de-trapping at and near the channel/dielectric non-ideal interface to achieve short-term memory retention. We also investigated the cycle-to-cycle and device-to-device programming variation in 2D memtransistors as shown in Section S4, Supporting Information.

The analog and non-volatile programmability of 2D memtransistors was exploited to design reconfigurable TSIs for enabling our photoencoder. A TSI is composed of three inverters cascaded in series, that is, the output of one inverter is fed as the input to the next inverter. In our design, we have used two TSIs as shown in Figure 1c. Note that, in the absence of a p-type 2D memtransistor, our inverter design uses an n-type depletion load, which is obtained by shorting the gate and source terminals of the 2D memtransistor. Figure 2h shows the outputs, V_{out1}, V_{out2}, and V_{out3}, obtained from stages 1, 2, and 3 of the TSI, respectively, when the input voltage (V_{in}) is swept from 0 to 5 V. Figure 2i shows that the switching threshold (V_{SW}), which is defined as the magnitude of V_{in} at which the output voltage reaches $V_{DD}/2$, can be adjusted for all three-stages of the TSI by electrically programming the corresponding 2D memtransistors as described in Figure 2e,f. This programming capability is critical for the ultimate demonstration of spike-timing-based adaptive photoencoding since it allows us to configure the V_{SW} of the two TSIs slightly differently. Also note that the transition between the two logic levels (V_{DD} = 5 and 0 V) becomes more abrupt when the output is obtained at a later stage of the TSIs, that is, the gain of the transfer function increases with the number of stages. Figure 2j shows the gain of all three stages for different V_{SW}. As we will show later, higher gain of the TSIs is critical to achieve equal magnitude spike trains at the output of the XOR gate.

4. Photogating Effect and Photoencoding using a Cascaded Three-Stage Inverter

Next, we demonstrate the adaptive photosensing capability of our 2D memtransistors. **Figure 3**a shows the post-illumination transfer characteristics for a representative 2D memtransistor after being exposed to different illumination intensities from a blue light emitting diode for 1 s while biased at V_{DS} = 1 V and different V_{BG}. Two distinct photoresponses are seen in the monolayer MoS_2-based 2D memtransistors. For V_{BG} > 0 V, that is, illumination in the on-state and in the subthreshold regime, there is no change in the post-illumination transfer characteristics, whereas for V_{BG} < 0 V, that is, illumination in the off-state, the post-illumination transfer characteristics show negative shift in V_{TH}; thus, the device is considered to exhibit persistent photoconductivity via the photogating effect. The photogating effect is a direct consequence of photocarrier trapping in the dielectric and is most dominant in the off-state. Many neuromorphic devices exploit the phenomenon of photogating effect.[39–42] In the on-state, photocarriers generated in the 2D channel are instead swept across by the applied V_{DS} and hence there is no persistent photocurrent beyond the optical exposure. Next, we exploit the persistent photoconductivity seen in a 2D memtransistor for photoencoding. Figure 3b shows the evolution of the post-illumination transfer characteristics of a 2D memtransistor as a function of illumination duration ($\tau_{illumin}$) for P_{in} = 25 W m^{-2}. During illumination the device was biased at V_{DS} = 1 V and V_{BG} = –4 V. Section S5, Supporting Information, shows the cycle-to-cycle and device-to-device variation in photoresponse for different $\tau_{illumin}$. Figure 3c shows the evolution of persistent photoconductance (G_{2D}) measured at V_{BG} = 0 V as a function of $\tau_{illumin}$ for different P_{in}. As expected, higher P_{in} values lead to faster change in G_{2D}. To transcribe the change in G_{2D} into spike trains, first, we use the TSI as described earlier. Figure 3d shows the post-illumination V_{out1}, V_{out2}, and V_{out3} as a function of V_{in} when the TSI is exposed to different $\tau_{illumin}$ for P_{in} = 25 W m^{-2}. During illumination, the TSI was biased using V_{DD} = 5 V and V_{in} = –4 V. The gradual shift in V_{SW} seen for all stages of the TSI is a consequence of the negative shift in V_{TH} observed in the memtransistor owing to the previously discussed photogating effect. Figure 3e,f shows the temporal evolution of V_{out1}, V_{out2}, and V_{out3} when V_{in} is toggled between –4 and 0.5 V and V_{DD} is toggled between 1 and 5 V in dark (P_{in} = 0 W m^{-2}) and under illumination (P_{in} = 25 W m^{-2}), respectively. Note that, during each clock cycle, the combination of V_{in} = –4 V and V_{DD} = 1 V is used to exploit the photogating effect, whereas the combination of V_{in} = 0.5 V and V_{DD} = 5 V is used to readout the logic level. While no changes are observed under dark condition, the output voltage from all three stages goes through logic transitions under illumination. This is because in dark V_{SW} is static, whereas under illumination V_{SW} decreases monotonically; once V_{SW} passes the V_{in} of 0.5 V, the logic transitions. This transition is most abrupt for V_{out3} due to higher gain, as described earlier.

© 2022 Wiley-VCH GmbH

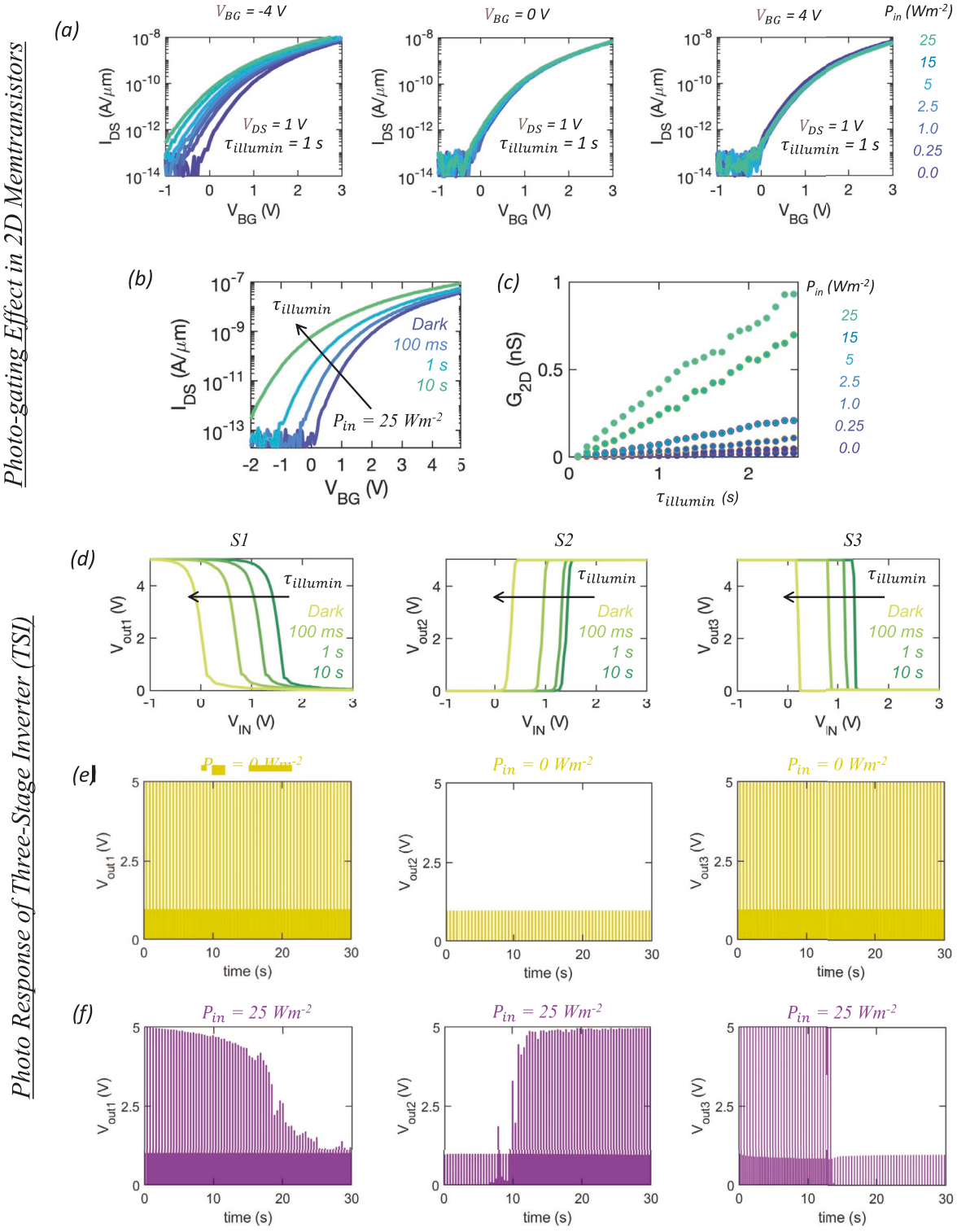


Figure 3. Photogating in 2D memtransistor and phototransduction using TSI. a) Post-illumination transfer characteristics for a monolayer MoS_2-based memtransistor after being exposed to different illumination intensities (P_{in}) for $\tau_{illumin}$ = 1 s at different V_{BG} using V_{DS} = 1 V. b) Evolution of the post-illumination transfer characteristics of a monolayer MoS_2-based memtransistor as a function of illumination duration ($\tau_{illumin}$) for P_{in} = 25 W m^{-2}. c) Evolution of persistent photoconductance (G_{2D}) measured at V_{BG} = 0 V as a function of $\tau_{illumin}$ for different P_{in}. d) Post-illumination output characteristics, that is, V_{out1}, V_{out2}, and V_{out3}, as a function of V_{in} when the TSI is exposed to different $\tau_{illumin}$ at P_{in} = 25 W m^{-2}. During illumination, the TSI was biased using V_{DD} = 5 V and V_{in} = –4 V. The gradual shift in V_{SW} seen for all stages of the TSI is a consequence of the negative shift in V_{TH} observed in the memtransistor due to the photogating effect. e,f) Temporal evolution of the output voltage, V_{out1}, V_{out2}, and V_{out3}, when V_{in} is toggled between –4 and 0.5 V and V_{DD} is toggled between 1 and 5 V in the dark (P_{in} = 0 W m^{-2}) (e) and under illumination (P_{in} = 25 W m^{-2}) (f).

© 2022 Wiley-VCH GmbH

5. Acceleration of Spike-Timing-Based Adaptive and Sparse Encoding of Visual Stimuli

To transform the logic transition into spike trains, we feed the outputs of two TSIs into a XOR gate as shown in **Figure 4**a. Our n-type memtransistor-based XOR gate design is shown in Figure 1c,d. Figure 4b shows the input and output voltage waveforms, V_A, V_B, and V_C, for the XOR gate, confirming correct logic operation, that is, V_C attains logic "1" (≈5 V) only when either V_A or V_B is at logic "1." Figure 4c shows the temporal evolution of V_A and V_B obtained at the output of the two TSIs configured with slightly different switching threshold around V_{SW} = 1 V and the corresponding V_C under different P_{in}. Since the logic transition for V_A and V_B for any given P_{in} occurs with a lag due to the difference in V_{SW}, V_C attains logic "1" or spiking only for that lag duration following the XOR logic. More interestingly, the time of occurrence of the first spike (τ_{tfs}) is inversely proportional to P_{in} as summarized in Figure 4d. The brightest illumination (P_{in} = 25 W m^{-2}) invokes spikes within ≈700 ms, whereas the weakest illumination (P_{in} = 0.25 W m^{-2}) requires more than ≈35 s to invoke the first spike. Figure 4e shows the spike-timing-based encoding energy (E_{encode}) for different P_{in} ranging from ≈100 nJ to ≈1 µJ. Since brighter illuminations are detected earlier, the corresponding energy consumptions are less. Note that the power consumption by the clocking signals (E_{clk}) were found to be ≈210 fJ per cycle, which is insignificant compared to E_{encode} (see Section S6, Supporting Information, for E_{clk}). Nevertheless, this demonstration confirms energy efficient and spike-timing-based photoencoding of input optical stimuli by our bioinspired 2D afferent neuron made using monolayer MoS_2-based photosensitive memtransistors.

Note that the number of spikes obtained for different P_{in} shows some stochasticity in their count. This can be attributed to the device-to-device and cycle-to-cycle variation in the photoresponse of the 2D memtransistors. Furthermore, some of the photocarriers trapped when V_{in} = –4 V can get released when the logic is read using V_{in} = 0.5 V, leading to finite fluctuation in V_{SW} as it approaches 0.5 V. If the monotonic change in V_{SW} occurs fast enough compared to the stochastic fluctuation in trapping/de-trapping of charge carriers, the transition is abrupt and permanent as seen in the case of brighter P_{in}. However, if the change in V_{SW} occurs slowly, the stochastic fluctuation can reverse the logic transition one or more times before reaching the steady state, as seen for weaker illuminations. Also note that the number of spike count can be adjusted by changing the difference in V_{SW} (Δ) for the two TSIs. A higher Δ will lead to more spiking and vice versa. Nevertheless, since our main objective is to achieve spike-timing-based encoding, we have not explored the impact of Δ on spike encoding. The reader may also wonder if a single inverter can be used instead of a TSI for the photoencoding application. Section S7, Supporting Information, shows V_C when V_A and V_B are obtained from the output of the first stage of the TSI, that is, V_{out1}. Clearly, in the absence of an abrupt transition under illumination the input logic levels to the XOR gate never differ significantly and hence no spiking is seen at the output; thus, use of a TSI is required for the discussed photoencoding capabilities to be realized. Section S8, Supporting Information, shows that the photoencoder can also achieve periodic output spikes in response to a constant illumination by introducing a periodic "Erase" pulse to V_{in}.

Next, we demonstrate adaptive photoencoding capability in our solid-state afferent neuron under scotopic and photopic conditions. For example, to expedite the spiking response under weaker illumination the TSI units may be reconfigured with lower V_{SW} by programming the constituent memtransistors as described earlier. Figure 4f shows the TSI transfer characteristics with different V_{SW} and the corresponding spike encoded output, V_C, for P_{in} = 1 W m^{-2}. Figure 4g shows τ_{tfs} as a function of V_{SW} for P_{in} = 1 W m^{-2}. As expected, lower V_{SW} = 0.6 V can enable spiking as early as at ≈450 ms for weaker illumination and vice versa. Similarly, the spiking response for brighter illumination can be delayed by increasing V_{SW}. Figure 4h shows the TSI transfer characteristics with different V_{SW} and corresponding spike encoded output, V_C, for P_{in} = 25 W m^{-2}. Figure 4i shows τ_{tfs} as a function of V_{SW} for P_{in} = 25 W m^{-2}. As expected, higher V_{SW} = 1.8 V can defer spiking to as late as ≈8 s under photopic conditions. These demonstrations confirm adaptive photoencoding capability in our artificial afferent neuron.

While our demonstration is based on visual stimuli, the concept can be extended to encode any type of sensory information such as audio, chemical, and tactile. One of the motivations behind the use of a 2D semiconductor such as monolayer MoS_2 as the channel material for our memtransistor technology can be accredited to the fact that there have been numerous demonstration of photodetectors,[43] chemical sensors,[44] biological sensors,[44] touch sensors,[45] and radiation sensors[46] using this material. Therefore, 2D memtransistors can naturally serve as artificial and multifunctional afferent neurons to transcribe diverse sensory stimuli into spike trains by exploiting spike-timing-based encoding using our MSI optoelectronic circuit.

6. Conclusion

We have proposed and experimentally demonstrated an MSI optoelectronic circuit composed of 21 photosensitive 2D memtransistors for spike-timing-based encoding of external visual information. As experimentally demonstrated, this photoencoder can transcribe optical stimuli into sparse spike trains with the information on illumination intensity being encoded into time-to-first spike, allowing it to function as a bioinspired solid-state afferent neuron. We also exploit the analog and non-volatile programming capabilities of our memtransistors to expedite or delay spiking under scotopic and photopic conditions, respectively, mimicking the neuroplasticity of visual neurons in the optic nerves of primates. Notably, our adaptive photoencoder consumes miniscule hardware and energy resources to accomplish these encoding functionalities, covering an area of only ≈100 µm^2 and consuming less than 1 µJ per operation. Finally, to the best of our knowledge, this is the first demonstration of hardware acceleration of spike-timing-based adaptive and sparse photoencoding, indicating the potential of this work to aid in the future development of SNNs. Our demonstration also highlights the importance of multifunctional 2D semiconductors, which are shown to offer sensing,

© 2022 Wiley-VCH GmbH

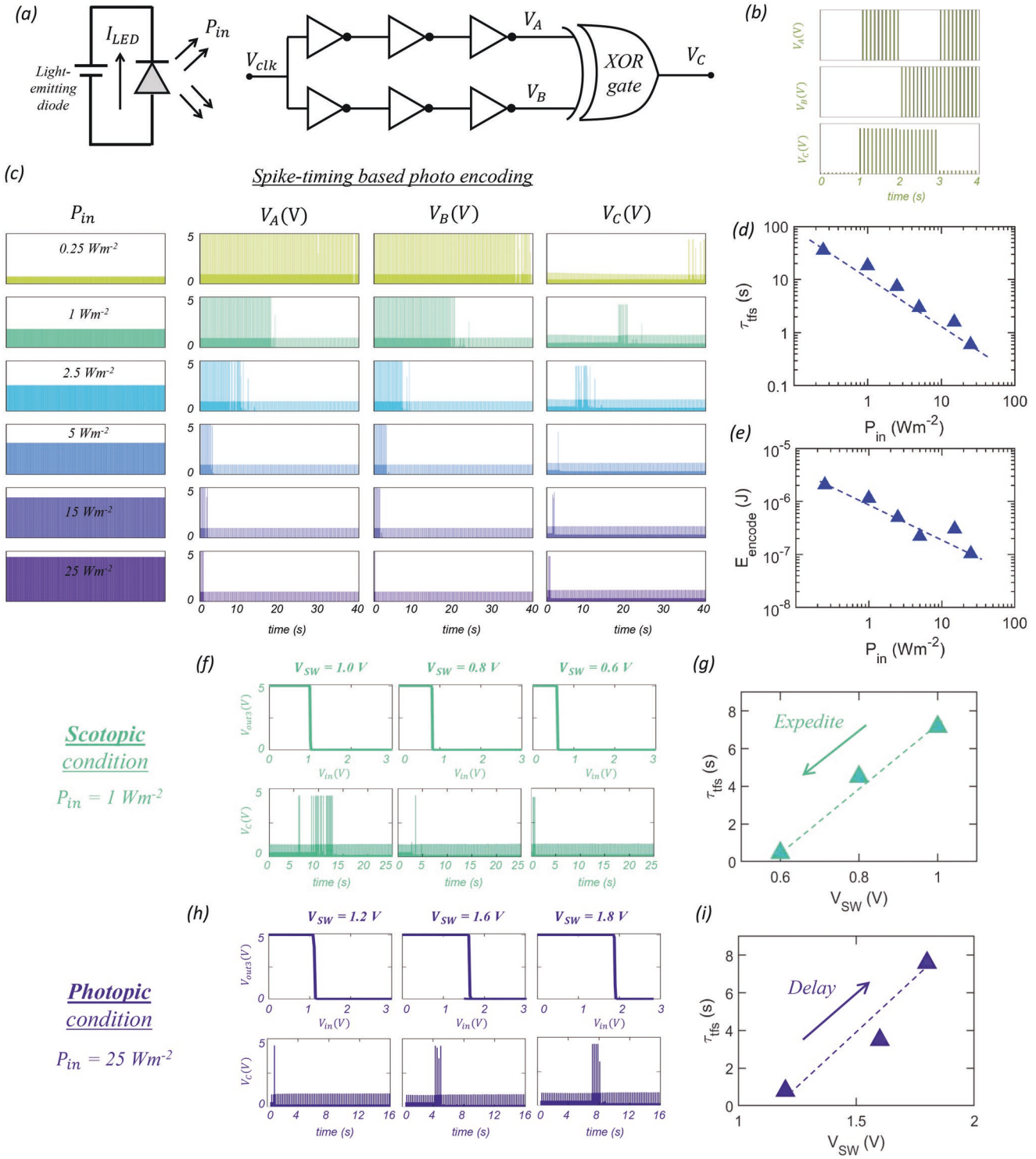


Figure 4. Demonstration of spike-timing-based sparse and adaptive photoencoding. a) Schematic representation of the artificial afferent neuron shown in Figure 1c,d constructed using two TSIs and a XOR gate. b) Input and output voltage waveforms, V_A, V_B, and V_C, for the XOR gate, confirming correct logic operation, that is, V_C attains logic "1" (≈5 V) only when either V_A or V_B is at logic "1". c) Temporal evolution of V_A and V_B obtained at the output of the TSIs configured with slightly different V_{SW}, and corresponding V_C under different P_{in}. d) Time-to-first spike (τ_{tfs}) as a function of P_{in}. The brightest illumination (P_{in} = 25 W m^{-2}) invokes spikes within ≈700 ms, whereas the weakest illumination (P_{in} = 0.25 W m^{-2}) requires more than ≈35 s to invoke the first spike. e) Encoding energy expenditure (E_{encode}) as a function of P_{in}. E_{encode} is found to be in the range of 100 nJ to 1 µJ. f) Output characteristics of the TSI with different V_{SW} and corresponding spike encoded output, V_C, for P_{in} = 1 W m^{-2}. g) τ_{tfs} as a function of V_{SW} for P_{in} = 1 W m^{-2}. As expected, a lower V_{SW} expedites spiking under scotopic condition. h) Output characteristics of the TSI with different V_{SW} and corresponding spike encoded output, V_C, for P_{in} = 25 W m^{-2}. i) τ_{tfs} as a function of V_{SW} for P_{in} = 25 W m^{-2}. As expected, spiking response for brighter illumination can be delayed by increasing V_{SW}.

© 2022 Wiley-VCH GmbH

computing, and non-volatile and analog storage capabilities; such a wide range of functionalities is difficult to accomplish through other state-of-the-art and emerging materials and devices, indicating the potential for monolithically integrated platforms based on 2D semiconductors.

7. Experimental Section

Fabrication of Local Back-Gate Islands: To define the back-gate island regions, a commercially bought substrate (285 nm SiO_2 on p^{++}-Si) was spin-coated at 4000 RPM for 45 s with bilayer photoresist consisting of Lift-Off-Resist (LOR 5A) and Series Photoresist (SPR 3012) and baked at 185 and 95 °C, respectively. The bilayer photoresist was then exposed using a Heidelberg Maskless Aligner (MLA 150) to define the island and developed using MF CD26 Microposit, followed by a deionized (DI) water rinse. The back-gate electrode of 50/20 nm Pt/TiN 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. An ALD process was then implemented to grow 50 nm Al_2O_3 uniformly across the entire substrate, including the island regions. To access the individual Pt back-gate electrodes, etch patterns were defined using the same bilayer photoresist (LOR 5A and SPR 3012) used previously. Again, the bilayer photoresist was exposed using a MLA 150 and developed using MF CD26 Microposit. The 50 nm Al_2O_3 was subsequently dry etched using a BCl_3 reactive ion etch chemistry at 5 °C for a total of 80 s. Note that this process was split into four 20 s etches to minimize heating in the substrate and thus ensure a uniform etch rate/depth. Finally, the photoresist was removed to give access to the individual Pt electrodes.

Large Area Monolayer MoS_2 Film Growth: Monolayer MoS_2 was deposited on epi-ready 2″ c-sapphire substrate by 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.[47] Molybdenum hexacarbonyl ($Mo(CO)_6$) and hydrogen sulfide (H_2S) were used as the precursors. $Mo(CO)_6$ maintained at 10 °C and 650 Torr in a stainless-steel bubbler was used to deliver 1.1×10^{-3} sccm of the metal precursor for the growth, while 400 sccm of H_2S was used for the process. MoS_2 deposition was carried out at 950 °C and 50 Torr in H_2 ambient, and monolayer growth was achieved in 18 min. The substrate was first heated to 1000 °C in H_2 and maintained for 10 min before the growth was initiated. After growth, the substrate was cooled in H_2S to 300 °C to inhibit decomposition of the MoS_2 films. Further details can be found in earlier works.[33,48,49]

MoS_2 Film Transfer to Local Back-Gate Islands: A poly(methyl methacrylate) (PMMA)-assisted wet-transfer process was used to transfer the monolayer MoS_2 film from the sapphire growth substrate to the previously discussed SiO_2/p^{++}-Si substrate with local back-gate islands. The growth substrate was first spin-coated with PMMA 495 A6 and allowed to sit for 24 h to promote good PMMA/MoS_2 adhesion. The edges of the spin-coated film were then scratched using a razor blade and the substrate was immersed inside of a 1 M NaOH solution kept at 90 °C. Capillary action served to draw the NaOH solution into the substrate/film interface, separating the PMMA/MoS_2 stack from the sapphire growth substrate. A clean glass slide was then used to retrieve the free-floating film and transfer it to a DI water bath for 15 min to rinse off any NaOH residues; this process was then repeated two more times. After the third water bath, the local back-gate island substrate was used to fish out the film before being baked at 50 °C and 70 °C for 10 min each to remove moisture and promote adhesion. Finally, the PMMA supporting layer was removed using an acetone bath (1 h) and the substrate was cleaned using an IPA bath (15 min).

Fabrication of 2D Memtransistor: The memtransistor channel region was defined using electron beam (e-beam) lithography. The transferred monolayer MoS_2 film was spin-coated with PMMA 950 A6 (4000 rpm for 45 s) and baked at 180 °C for 90 s. Then, the resist was patterned using e-beam lithography and developed with a 1:1 mixture of 4-methyl-2-pentanone (MIBK) and IPA for 60 s and pure IPA for 45 s. The e-beam exposed area of monolayer MoS_2 was then dry etched using a sulfur hexafluoride (SF_6) reactive ion etch at 5 °C for 30 s, isolating the channel regions of each device. The PMMA resist was then removed using acetone and IPA solution. To define source and drain contacts, a bilayer resist of methyl methacrylate (MMA) and PMMA A3 was spun (4000 RPM for 45 s) onto the substrate and baked at 150 °C for 90 s and 180 °C for 90 s, respectively. The source and drain contacts were then patterned using e-beam lithography and developed using the same process as previously. 40 nm nickel (Ni) and 30 nm gold (Au) were deposited using e-beam evaporation; all excess metal was removed via lift-off performed by keeping the sample in acetone for 30 min, with a subsequent 30 min IPA bath for cleaning.

Monolithic Integration: To define the connections between the respective memtransistors for the fabrication of the photoencoder, the substrate was again spin-coated with MMA and PMMA, patterned via e-beam lithography, and developed using the processes detailed previously. E-beam evaporation was used of to deposit 60 nm Ni/30 nm Au. The increased thickness of the Ni layer as compared to the original source/drain contacts was to ensure conformality and adhesion to the substrate. Finally, the e-beam resist was rinsed away via the same lift-off process using acetone and IPA.

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

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

The work was supported by Army Research Office (ARO) through Contract Number W911NF1920338 and the National Science Foundation (NSF) through CAREER Award under Grant Number ECCS-2042154. The authors also acknowledge Mr. Rahul Pendurthi, Mr. Darsith Jayachandran, Mr. Najam U. Sakib, and Mr. Muhtasim Ul Karim Sadaf for their help with device fabrication and characterization and the materials support from The Pennsylvania State University 2D Crystal Consortium–Materials Innovation Platform (2DCCMIP) under NSF cooperative agreement DMR-1539916.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

S.D conceived the idea and designed the experiments. S.S.R., and T.F.S. fabricated the memtransistor-based circuits. S.C. made the circuit layouts. S.D., S.S.R., S.C., D.S., and M.D. performed the measurements analyzed the data, discussed the results, and agreed on their implications. All the authors contributed to the preparation of the manuscript.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

© 2022 Wiley-VCH GmbH

Keywords

2D materials, memtransistors, photoencoders, sparse spike-based encoding

Received: March 18, 2022
Revised: May 31, 2022
Published online:

[1] S. Ghosh-Dastidar, H. Adeli, *Int. J. Neural Syst.* **2009**, *19*, 295.
[2] F. Ponulak, A. Kasinski, *Acta Neurobiol. Exp.* **2011**, *71*, 409.
[3] B. Han, A. Sengupta, K. Roy, in *2016 Int. Joint Conf. on Neural Networks (IJCNN)*, IEEE, Piscataway, NJ, USA **2016**, pp. 971–976.
[4] M. Davies, N. Srinivasa, T.-H. Lin, G. Chinya, Y. Cao, S. H. Choday, G. Dimou, P. Joshi, N. Imam, S. Jain, Y. Liao, C. K. Lin, A. Lines, R. Liu, D. Mathaikutty, S. Mccoy, A. Paul, J. Tse, G. Venkataramanan, Y.-H. Weng, A. Wild, Y. Yang, H. Wang, *IEEE Micro* **2018**, *38*, 82.
[5] M. Rudnicki, O. Schoppe, M. Isik, F. Völk, W. Hemmert, *Cell Tissue Res.* **2015**, *361*, 159.
[6] S. Reiter, C. C. Rodriguez, K. Sun, M. Stopfer, *J. Neurosci.* **2015**, *35*, 12309.
[7] R. M. Hallock, P. M. Di Lorenzo, *Neurosci. Biobehav. Rev.* **2006**, *30*, 1145.
[8] M. Meister, L. Lagnado, D. A. Baylor, *Science* **1995**, *270*, 1207.
[9] S.-Y. Choi, B. Borghuis, R. Rea, E. S. Levitan, P. Sterling, R. H. Kramer, *Neuron* **2005**, *48*, 555.
[10] E. D. Adrian, Y. Zotterman, *J. Physiol.* **1926**, *61*, 151.
[11] D. I. Perrett, E. T. Rolls, W. Caan, *Exp. Brain Res.* **1982**, *47*, 329.
[12] R. S. Johansson, I. Birznieks, *Nat. Neurosci.* **2004**, *7*, 170.
[13] P. Reinagel, R. C. Reid, *J. Neurosci.* **2000**, *20*, 5392.
[14] J. Huxter, N. Burgess, J. O'keefe, *Nature* **2003**, *425*, 828.
[15] S. J. Thorpe, in *Parallel Processing in Neural Systems*, Elsevier, Amsterdam, The Netherlands **1990**, pp. 91–94.
[16] D. A. Butts, C. Weng, J. Jin, C. I. Yeh, N. A. Lesica, J. M. Alonso, G. B. Stanley, *Nature* **2007**, *449*, 92.
[17] S. M. Chase, *J. Neurosci.* **2006**, *26*, 3889.
[18] D. S. Reich, F. Mechler, J. D. Victor, *J. Neurophysiol.* **2001**, *85*, 1039.
[19] R. Gütig, *Curr. Opin. Neurobiol.* **2014**, *25*, 134.
[20] G. Indiveri, B. Linares-Barranco, T. J. Hamilton, A. Van Schaik, R. Etienne-Cummings, T. Delbruck, S.-C. Liu, P. Dudek, P. Häfliger, S. Renaud, J. Schemmel, G. Cauwenberghs, J. Arthur, K. Hynna, F. Folowosele, S. Saighi, T. Serrano-Gotarredona, J. Wijekoon, Y. Wang, K. Boahen, *Front. Neurosci.* **2011**, *5*, 73.
[21] G. Indiveri, in *Proc. 2003 Int. Symp. on Circuits and Systems, 2003. ISCAS'03*, IEEE, Piscataway, NJ, USA **2003**; https://doi.org/10.1109/ISCAS.2003.1206342.
[22] S. A. Aamir, P. Müller, A. Hartel, J. Schemmel, K. Meier, in *ESSCIRC Conf. 2016: 42nd European Solid-State Circuits Conf.*, IEEE, Piscataway, NJ, USA **2016**, pp. 71–74.
[23] B. Linares-Barranco, E. Sanchez-Sinencio, A. Rodriguez-Vazquez, J. L. Huertas, *IEEE J. Solid-State Circuits* **1991**, *26*, 956.
[24] I. Sourikopoulos, S. Hedayat, C. Loyez, F. Danneville, V. Hoel, E. Mercier, A. Cappy, *Front. Neurosci.* **2017**, *11*, 123.
[25] C. Zhao, B. T. Wysocki, Y. Liu, C. D. Thiem, N. R. Mcdonald, Y. Yi, *ACM J. Emerging Technol. Comput. Syst.* **2015**, *12*, 23.
[26] Z. Wang, S. Joshi, S. Savel'ev, W. Song, R. Midya, Y. Li, M. Rao, P. Yan, S. Asapu, Y. Zhuo, H. Jiang, P. Lin, C. Li, J. H. Yoon, N. K. Upadhyay, J. Zhang, M. Hu, J. P. Strachan, M. Barnell, Q. Wu, H. Wu, R. S. Williams, Q. Xia, J. J. Yang, *Nat. Electron.* **2018**, *1*, 137.
[27] I. Gupta, A. Serb, A. Khiat, R. Zeitler, S. Vassanelli, T. Prodromakis, *Nat. Commun.* **2016**, *7*, 12805.
[28] S. Kim, C. Du, P. Sheridan, W. Ma, S. Choi, W. D. Lu, *Nano Lett.* **2015**, *15*, 2203.
[29] M. Al-Shedivat, R. Naous, G. Cauwenberghs, K. N. Salama, *IEEE J. Emerging Sel. Top. Circuits Syst.* **2015**, *5*, 242.
[30] Z. Zhao, L. Qu, L. Wang, Q. Deng, N. Li, Z. Kang, S. Guo, W. Xu, *IEEE Trans. Circuits Syst. II: Express Beliefs* **2020**, *67*, 931.
[31] M. Chu, B. Kim, S. Park, H. Hwang, M. Jeon, B. H. Lee, B. G. Lee, *IEEE Trans. Ind. Electron.* **2014**, *62*, 2410.
[32] List Of Thin Film Samples Available, https://www.mri.psu.edu/2d-crystal-consortium/user-facilities/thin-films/list-thin-film-samples-available (accessed: June 2022).
[33] A. Sebastian, R. Pendurthi, T. H. Choudhury, J. M. Redwing, S. Das, *Nat. Commun.* **2021**, *12*, 693.
[34] H. Li, Q. Zhang, C. C. R. Yap, B. K. Tay, T. H. T. Edwin, A. Olivier, D. Baillargeat, *Adv. Funct. Mater.* **2012**, *22*, 1385.
[35] A. Splendiani, L. Sun, Y. Zhang, T. Li, J. Kim, C. Y. Chim, G. Galli, F. Wang, *Nano Lett.* **2010**, *10*, 1271.
[36] S. Das, H. Y. Chen, A. V. Penumatcha, J. Appenzeller, *Nano Lett.* **2013**, *13*, 100.
[37] D. S. Schulman, A. J. Arnold, S. Das, *Chem. Soc. Rev.* **2018**, *47*, 3037.
[38] S. Chuang, C. Battaglia, A. Azcatl, S. McDonnell, J. S. Kang, X. Yin, M. Tosun, R. Kapadia, H. Fang, R. M. Wallace, A. Javey, *Nano Lett.* **2014**, *14*, 1337.
[39] D. Xie, L. Wei, M. Xie, L. Jiang, J. Yang, J. He, J. Jiang, J. He, *Adv. Funct. Mater.* **2021**, *31*, 2010655.
[40] G. Feng, J. Jiang, Y. Li, D. Xie, B. Tian, Q. Wan, *Adv. Funct. Mater.* **2021**, *31*, 2104327.
[41] Y. Li, K. Yin, Y. Diao, M. Fang, J. Yang, J. Zhang, H. Cao, X. Liu, J. Jiang, *Nanoscale* **2022**, *14*, 2316.
[42] J. Jiang, W. Hu, D. Xie, J. Yang, J. He, Y. Gao, Q. Wan, *Nanoscale* **2019**, *11*, 1360.
[43] Z. Yin, H. Li, H. Li, L. Jiang, Y. Shi, Y. Sun, G. Lu, Q. Zhang, X. Chen, H. Zhang, *ACS Nano* **2012**, *6*, 74.
[44] L. Wang, Y. Wang, J. I. Wong, T. Palacios, J. Kong, H. Y. Yang, *Small* **2014**, *10*, 1101.
[45] M. Park, Y. J. Park, X. Chen, Y. K. Park, M. S. Kim, J. H. Ahn, *Adv. Mater.* **2016**, *28*, 2556.
[46] A. J. Arnold, T. Shi, I. Jovanovic, S. Das, *ACS Appl. Mater. Interfaces* **2019**, *11*, 8391.
[47] Y. Xuan, A. Jain, S. Zafar, R. Lotfi, N. Nayir, Y. Wang, T. H. Choudhury, S. Wright, J. Feraca, L. Rosenbaum, J. M. Redwing, V. Crespi, A. C. T. Van Duin, *J. Cryst. Growth* **2019**, *527*, 125247.
[48] D. Jayachandran, A. Oberoi, A. Sebastian, T. H. Choudhury, B. Shankar, J. M. Redwing, S. Das, *Nat. Electron.* **2020**, *3*, 646.
[49] A. Dodda, A. Oberoi, A. Sebastian, T. H. Choudhury, J. M. Redwing, S. Das, *Nat. Commun.* **2020**, *11*, 4406.

© 2022 Wiley-VCH GmbH

Check for updates

An Annealing Accelerator for Ising Spin Systems Based on In-Memory Complementary 2D FETs

*Amritanand Sebastian, Sarbashis Das, and Saptarshi Das**

Metaheuristic algorithms such as simulated annealing (SA) are often implemented for optimization in combinatorial problems, especially for discreet problems. SA employs a stochastic search, where high-energy transitions ("hill-climbing") are allowed with a temperature-dependent probability to escape local optima. Ising spin glass systems have properties such as spin disorder and "frustration" and provide a discreet combinatorial problem with a high number of metastable states and ground-state degeneracy. In this work, subthreshold Boltzmann transport is exploited in complementary 2D field-effect transistors (p-type WSe_2 and n-type MoS_2) integrated with an analog, nonvolatile, and programmable floating-gate memory stack to develop in-memory computing primitives necessary for energy- and area-efficient hardware acceleration of SA for Ising spin systems. Search acceleration of >800× is demonstrated for 4 × 4 ferromagnetic, antiferromagnetic, and spin glass systems using SA compared to an exhaustive search using a brute force trial at miniscule total energy expenditure of ≈120 nJ. The hardware-realistic numerical simulations further highlight the astounding benefits of SA in accelerating the search for larger spin lattices.

1. Introduction

Combinatorial optimization problems represent a set of problems where finding the best solution using an exhaustive search is often unfeasible. Interestingly, such problems appear in applications ranging from supply-chain management, airline scheduling, and industrial resource allocation to artificial intelligence, applied mathematics, and theoretical computer science. The travelling salesman problem (TSP) is a representative optimization problem where the shortest route connecting all of a given number of cities must be found.[1] In TSP, the time complexity is on the order of $n!$, where n is the number of cities. The optimal solution can be found for a small n by an exhaustive search using the brute force trial (BFT) method, wherein every combination is evaluated. However, BFT becomes impractical for large n. For example, 40 cities would require 10^{49} trials! While BFT fails, various metaheuristic algorithms such as simulated annealing (SA),[2] genetic algorithm,[3] and ant colony optimization[4] have been utilized to obtain an approximate solution which is very close to the actual solution for TSP and other optimization problems. Among these, SA is an excellent optimization technique in discrete problems where multiple local minima exist. SA provides a simple framework which can be implemented on systems with arbitrary energy landscapes, and it statistically guarantees an optimal solution. SA has hence been employed to solve optimization problems in a wide variety of domains such as circuit design,[5] data analysis,[6] imaging,[7] neural networks,[8] geophysics,[9] finance,[10] and the Ising model of magnetism.[11]

SA draws inspiration from physical annealing, in which a material is heated above its recrystallization temperature to allow atoms to rearrange and is then slowly cooled down to improve its crystallinity and reach a low energy state. SA is an optimization algorithm where the free energy (H) of a system is minimized by employing a stochastic search. It is similar to other optimization methods, such as gradient descent,[12] where transitions lowering H are accepted. However, unlike the gradient descent method, it also allows transitions increasing H ("hill-climbing"), determined by the annealing temperature.[13–16] This "hill-climbing" feature of SA makes it highly attractive for systems with multiple local minima in their energy landscape (**Figure 1**a). The cost associated with a state transition leading to ΔH change in the free energy of the system is evaluated using the Boltzmann factor and accepted if it falls below a predefined threshold, P, following Equation (1)

$$\exp\left(\frac{\Delta H}{k_B T}\right) < P \tag{1}$$

A. Sebastian, S. Das
Deparment of Engineering Science and Mechanics
Penn State University
University Park, PA 16802, USA
E-mail: sud70@psu.edu

S. Das
Department of Electrical Engineering
Penn State University
University Park, PA 16802, USA

S. Das
Department of Materials Science and Engineering
Penn State University
University Park, PA 16802, USA

S. Das
Materials Research Institute
Pennsylvania State University
University Park, PA 16802, USA

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/adma.202107076.

DOI: 10.1002/adma.202107076

© 2021 Wiley-VCH GmbH

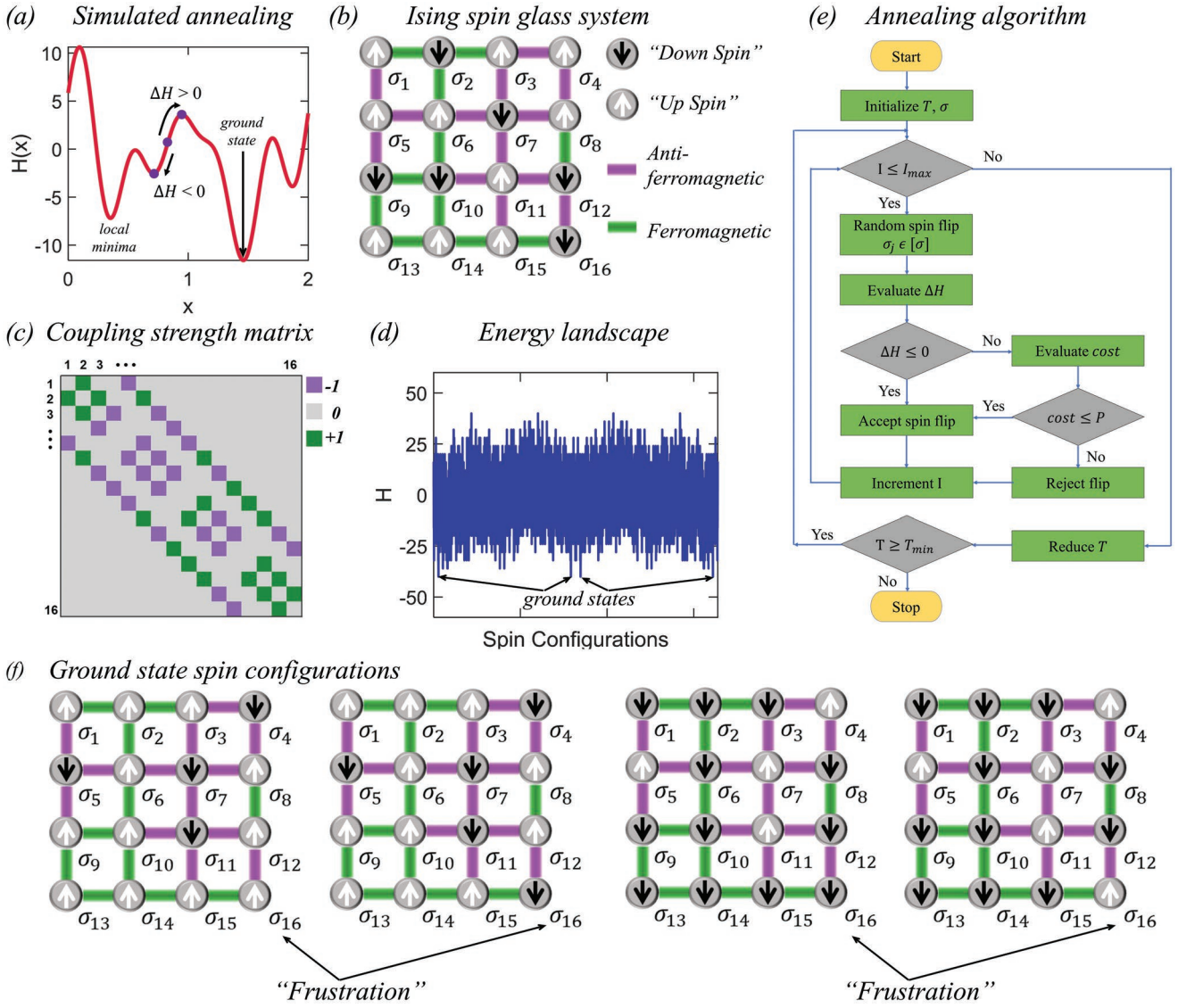


Figure 1. Simulated annealing (SA) and Ising spin glass system. a) Illustration of the basic principle of SA used for finding the ground state(s) or lowest energy state(s) of a system in a large search space with multiple local minima. Unlike many other optimization methods, SA accepts transitions increasing H ("hill-climbing") based on the annealing temperature (T). b) A 4×4 Ising spin glass system with 16 randomly oriented spins in up (white) or down (black) direction. The neighboring spin interactions can be either ferromagnetic (green) or antiferromagnetic (purple). c) The corresponding coupling strength matrix [CS]. d) Free energy (H) corresponding to the 2^{16} possible states for the spin glass system in (b). The energy landscape shows multiple local minima with ground state degeneracy of 4. The number of brute force trials increases exponentially as the size of the spin lattice increases, qualifying the spin glass system as a challenging combinatorial optimization problem where SA can accelerate the search. e) Flowchart of the SA algorithm for a spin glass system. A predetermined cooling schedule is used for T. At each T, the algorithm runs for a predetermined number of iterations (I_{max}). During each iteration, a random spin is flipped and the ΔH value associated with the spin flip is evaluated. For $\Delta H \leq 0$, the spin flip is always accepted. For $\Delta H > 0$, the spin flip is accepted if the cost is less than a predetermined value, P. f) For a sufficiently large number of iterations, the system converges to one of the 4 ground states for the spin glass systems in (b).

Here, k_B is the Boltzmann constant and T is the temperature. This particular SA approach is also referred to as the threshold accepting method and is widely used for its simple structure.[17] The basic principle of SA is illustrated in Figure 1a. To minimize the free energy function $H(x)$ corresponding to the argument set $x = \{x_1, x_2, x_3, \ldots x_N\}$, where N is a large number, a random x_i is initialized. At each iteration, a new random point (x_k) is chosen and the ΔH associated with the transition is evaluated. If $\Delta H < 0$, the transition is automatically accepted, whereas if $\Delta H > 0$, the transition is accepted conditionally, i.e., "hill-climbing" is allowed following Equation (1). The SA algorithm starts at a high T, where significant "hill-climbing" is allowed, and T is progressively reduced following an annealing schedule. At a sufficient lower T the system finds the global minima or arrives close to the same.

Most of the work on optimization algorithms such as SA is done in software,[11,13–16] with the few hardware demonstrations[18,19] based on von-Neumann architecture rendering

© 2021 Wiley-VCH GmbH

them area and energy-inefficient owing to the physical separation of memory and compute. Non-von-Neumann hardware accelerators such as graphics processing units (GPUs)[20,21] and field-programmable gate arrays (FPGAs)[22] have become increasingly popular in recent years, offering speed-up, higher energy efficiency, and a smaller physical footprint.[23] Quantum annealing, a variant of SA where the "hill-climbing" is achieved by quantum mechanical tunneling, has been implemented in a D-wave processor, a quantum computer.[24] More recently, memristive crossbar arrays have been used to implement standalone SA[25] and SA in Hopfield neural networks for various optimization problems.[26–28] While memristors offer in-memory computing, they have limitations owing to sneak paths and narrow dynamic ranges, which limit the number of achievable conductance states and cause high power consumption and metal migration in the ON state due to high currents. To improve their performance, memristors are often integrated with an access transistor in a 1T1R structure and also require peripheral circuits based on silicon complementary–metal–oxide semiconductor (CMOS) technology.

In this work, we explore SA using in-memory complementary field-effect transistors (FETs) based on ultrathin-body 2D semiconductors, i.e., p-type WSe_2 and n-type MoS_2 FETs. First, unlike quantum computers operating at cryogenic temperatures, our demonstration is based on room temperature, and second, we exploit analog subthreshold conduction and analog programmability to design unique computing primitives and annealing schedule which achieve better energy and area efficiency compared to large memristive cross-bar arrays. Note that earlier works have already demonstrated the scalability and technological viability of 2D materials.[29–31] Additionally, in-memory and near-memory technologies based on 2D materials have been used recently for various computing tasks overcoming the von-Neumann bottleneck.[32–34] This work further advances the development of 2D FET-based in-memory computing platforms. To the best of our knowledge this is the first demonstration of hardware acceleration of SA for the Ising spin glass systems using emerging materials and devices.

2. Spin Glass System

Since the early years of SA, the Ising spin glass problem has been extensively studied since it offers a combinatorially huge number of outcomes with multiple statistically equivalent ground states. Similar to the disordered nature of atomic positions in glass, in a spin glass system, the magnetic spins are disordered.[35,36] The spatial disorder in a glass is set by quenching, where its atoms are frozen in a disordered state. Similarly, spin glass is a system of quenched magnets with disorder stemming from frozen spin states and interactions. Spin glasses demonstrate interesting properties such as frustration. At low temperatures, a spin glass system has a roughly equal number of ferromagnetic and antiferromagnetic interactions, or bonds, which are frozen. However, spins can be flipped to obtain a low energy state which satisfies the frozen spin interactions. This leads to "frustration" in spins trying to settle between competing ferromagnetic and antiferromagnetic interactions.[37] A large number of metastable states with high degeneracy corresponding to different spin orientations are observed in spin glass systems. This leads to a non-monotonic energy landscape with multiple local minima. The transition from a metastable state to the lowest energy state requires specific spin flips and represents an optimization problem where high energy transitions should be allowed to escape from local minima; hence, spin glass is an ideal system to implement SA.

A $K \times K$ square spin lattice with $K = 4$ is shown in Figure 1b. The natures of the various spin interactions, i.e., ferromagnetic (green) versus antiferromagnetic (purple), are also shown. The corresponding spin vector $[\sigma]$ consisting of $K^2 = 16$ spins is given by $[\sigma]=[\sigma_1,\sigma_2,\ldots\sigma_{K^2}]$. In the Ising spin model, +1 represents "up" spin and −1 represents "down" spin, respectively. The Hamiltonian representing the free energy of the spin glass system is given by the zero-field Edwards–Anderson (EA) model described in Equation (2)[38]

$$H = -\sum_{i,j}\sigma_i J_{ij}\sigma_j = -\frac{1}{2}\sum_{i,j}\sigma_i J_{ij} N_{ij}\sigma_j = -\frac{1}{2}\sum_{i,j}\sigma_i CS_{ij}\sigma_j = -\frac{1}{2}[\sigma][CS][\sigma]^T \quad (2)$$

Here, σ_i and σ_j are the ith and jth elements of σ. J_{ij} denotes the nature of the interaction between σ_i and σ_j, i.e., $J_{ij} = +1$ for ferromagnetic interactions and $J_{ij} = -1$ for antiferromagnetic interactions. The operator <> denotes that only the nearest neighbor interactions are taken into account, i.e., $N_{ij} = +1$ if σ_i and σ_j are immediate neighbors and $N_{ij} = 0$ otherwise.[25] Note that both the $[J]$ and $[N]$ matrices are sparse matrices of size $K^2 \times K^2$. These two matrices are combined into a single $K^2 \times K^2$ matrix, referred to as the coupling strength matrix, [CS] (Figure 1c). The green squares represent +1, the purple squares represent −1, and the gray squares represent 0. Note that the $K \times K$ spin glass system with K^2 spins, each with two possible orientations ("up"/"down"), has 2^{K^2} possible states. Also note that, by changing $[N]$, long-range spin interactions can be incorporated into the model. The free energy (H) values corresponding to each of these 2^{K^2} states are shown in Figure 1d for the spin glass system shown in Figure 1b. Clearly, the energy landscape is non-monotonic with multiple local minima and degenerate global minima, or ground states. Furthermore, as the size of the spin lattice (K) increases, the search for spin configurations that satisfy all interactions becomes unfeasible using BFTs, thus qualifying the spin glass system as a challenging combinatorial optimization problem where SA can speed-up the search process.

3. Simulated Annealing for Spin Glass System

The SA algorithm to find the ground state of a spin glass system is shown in Figure 1e. To find the orientations of $K^2 = 16$ spins that result in minimum free energy, a random spin in the $K \times K$ spin lattice is flipped and the corresponding ΔH value is evaluated using Equation (2). If $\Delta H < 0$, the free energy of the system is lowered and hence the spin flip is accepted. However, if $\Delta H > 0$, the spin flip increases the free energy of the system. In this case, the spin flip can still be accepted ("hill-climbing") if the associated cost obtained from Equation (1) is lower than

© 2021 Wiley-VCH GmbH

the predetermined value, P. If both conditions are not satisfied, then the spin flip is rejected. At a given T, a predetermined number of iterations (I_{max}) are performed to traverse the energy landscape of the spin system. This process is then repeated by progressively cooling the system, i.e., by reducing T following a predefined cooling schedule. It is found that within a reasonable number of iterations, and at a sufficiently lower temperature, the system converges to its ground state (Figure 1f). Note that the spin glass systems typically exhibit ground-state degeneracy. See Figure S1a–d in the Supporting Information for the ground states and the free energy landscapes of 4 × 4 ferromagnetic and antiferromagnetic spin systems, respectively. In a ferromagnetic system, the ground state degeneracy is 2 since all interactions are satisfied if all spins are pointing "up" or "down" (Figure S1c, Supporting Information). An antiferromagnetic system also possesses ground state degeneracy of 2 since all interactions are satisfied if adjacent spins are oriented in opposite directions (Figure S1d, Supporting Information). However, in spin glass systems, the ground-state degeneracy is increased as the ground state is unable to satisfy all interactions due to the phenomenon of "frustration". For example, "up" and "down" configurations are equally valid for σ_{16} for the spin glass system to be in its ground state (Figure 1f). Video S1 in the Supporting Information shows SA for 4 × 4 ferromagnetic, antiferromagnetic, and "frustrated" spin glass systems.

4. Hardware Realization of Simulated Annealing

Hardware implementation of SA requires: 1) a random number generator for random spin flip, 2) a computational unit to calculate the change in free energy (ΔH) of the spin system associated with the random spin flip following Equations (2 and 3), 3) a computational unit to determine the cost of "hill-climbing" if $\Delta H > 0$ to accept or reject the spin flip following Equation (1), and, finally, 4) a hardware mechanism equivalent to the annealing/cooling schedule in metallurgy. While we use software code to generate the random numbers, all other computational units including the mechanism for annealing are realized in hardware. For example, a multiplier module is designed using complementary 2D FETs along with a resistor and a capacitor module to evaluate ΔH. Similarly, co-location of analog memory and analog computing enabled by nonvolatile and programmable MoS_2 FETs is used to determine the cost of "hill-climbing" and achieve an annealing schedule equivalent to changing the T.

The schematic and optical image of a back-gated p-type WSe_2 FET with 50 nm Al_2O_3 as the gate dielectric are shown in **Figure 2**a,b, respectively. Undoped WSe_2 demonstrates ambipolar transport with both electron (n-type) and hole (p-type) conduction owing to the pinning of the metal Fermi-level near the middle of the WSe_2 bandgap.[39–41] However, for the design of the multiplier module, unipolar p-type WSe_2 is preferred. Hence, WSe_2 is doped p-type using surface charge transfer doping with sub-stochiometric WO_{3-x}.[40] Doping is achieved on a multilayered WSe_2 flake by converting its top layers to WO_{3-x} through mild O_2 plasma exposure, as discussed in our earlier reports.[40,42] As shown in Figure 2a, before O_2 plasma exposure the flakes are patterned to obtain a p–i–p structure, i.e., p-doped source and drain extension regions with the middle region left intrinsic[40] (see the Experimental Section for details on p-type WSe_2 FET fabrication). The length of the intrinsic region is designed to be 500 nm. The corresponding transfer characteristics, i.e., drain current (I_{DS}) vs back-gate voltage (V_{BG}) for different drain-to-source voltage (V_{DS}), and output characteristics, i.e., I_{DS} vs V_{DS} for different V_{BG}, of the WSe_2 FET are shown in in Figure 2c,d, respectively. Clearly, the WSe_2 FET demonstrates unipolar p-type conduction with current on-off ratio of $\approx 10^8$, subthreshold slope, SS = 200 mV dec^{-1}, hole mobility, μ_P = 13 cm^2 V^{-1}s^{-1}, and on-current, I_{ON} = 11 μA μm^{-1} for an inversion carrier density of 4.4×10^{12} cm^{-2}.

The schematic and optical image of a back-gated MoS_2 FET with the same 50 nm Al_2O_3 gate dielectric and with a channel length of 500 nm are shown in Figure 2e,f, respectively. We have used monolayer MoS_2 grown using the same metal–organic chemical vapor deposition (MOCVD) technique as described in our earlier reports[32,43,44] (see the Experimental Section for details on n-type MoS_2 FET fabrication). For MoS_2, the metal Fermi-level pins closer to the conduction band at the source/drain contact interfaces, resulting in unipolar n-type conduction.[41,45] The corresponding transfer characteristics and output characteristics are shown in Figure 2g,h, respectively. For monolayer MoS_2 FET, we extracted current on-off ratio of $\approx 10^7$, subthreshold slope, SS = 310 mV dec^{-1}, electron mobility, μ_N = 15 cm^2 V^{-1}s^{-1}, and I_{ON} = 6.7 μA μm^{-1} for an inversion carrier density of 3.7×10^{12} cm^{-2}.

The unique stack of p^{++} Si/TiN/Pt/Al_2O_3 offers analog and nonvolatile memory, i.e., the threshold voltage of the FETs can be adjusted by applying a programming voltage (V_P) to the back-gate electrode for a programming pulse time (t_P). The programmability is shown in Figure 2i using the transfer characteristics of a MoS_2 FET at V_{DS} = 1 V after the application of negative programming voltages of different amplitudes for t_P = 100 ms. Figure 2j shows the retention characteristics, i.e., post-programmed I_{DS} vs time, measured at V_{BG} = 2 V, confirming nonvolatile and analog programmability of the MoS_2 FET. Similarly, by applying positive erase voltages (V_E) for an erase pulse time, t_E = 100 ms, the programmed states can be erased and the transfer characteristics can be shifted in the opposite direction as shown in Figure 2k. The corresponding nonvolatile retention characteristics are shown in Figure 2l. Note that programming (erase) operations can also be achieved by varying t_P (t_E) for a fixed V_P (V_E), as shown in Figure 2a,b in the Supporting Information. The working principle of the analog and nonvolatile back-gate memory stack has been described in detail in our earlier report.[32]

The above demonstration of p-type WSe_2 FETs, n-type MoS_2 FETs, and the analog compute and analog nonvolatile storage capability allow us to design the computational primitives necessary for the hardware realization of SA. Note that ΔH due to random spin flip events can be computed from the difference in the free energy of the spin glass system before (H_i) and after (H_f) the spin flip. According to Equation (2), evaluation of H and H' requires multiplication of the spin vector $[\sigma]$ of size 1 × K^2 with the [CS] matrix of size K^2 × K^2 followed by multiplication with the transpose of the spin vector, i.e., $[\sigma]^T$ of size K^2 × 1. However, there are several challenges: first, while vector matrix multiplication can be realized using cross-bar architectures, the [CS] matrix contains negative elements which

© 2021 Wiley-VCH GmbH

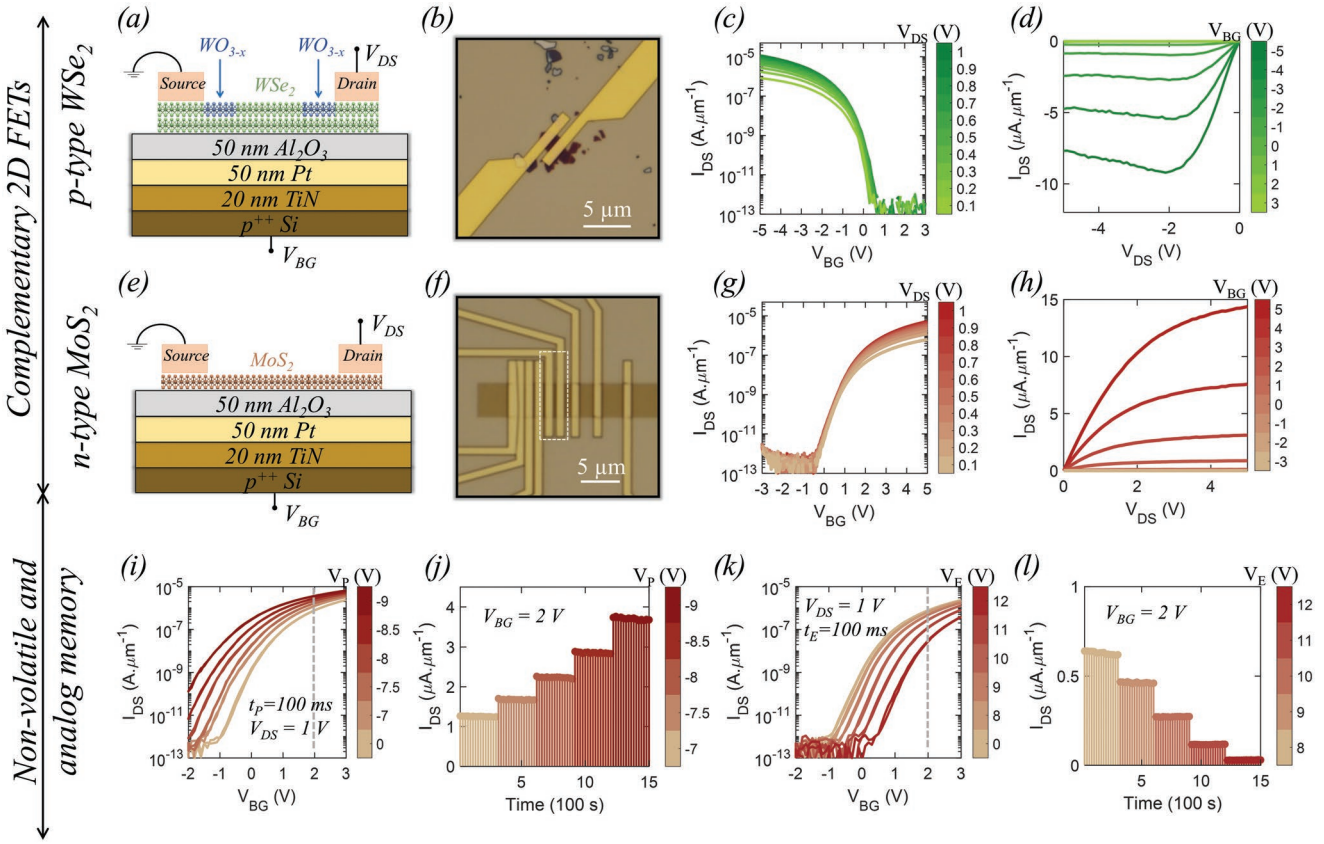


Figure 2. Analog in-memory complementary 2D field-effect transistors (FETs). a) Schematic and b) optical image of a back-gated p-type WSe_2 FET. The channel is selectively exposed to mild O_2 plasma to form a p–i–p structure with the length of the intrinsic region as 500 nm. c,d) Corresponding transfer (c) and output (d) characteristics for the WSe_2 FET. e) Schematic and f) optical image of a back-gated n-type MoS_2 FET with a channel length of 500 nm. g) Transfer and h) output characteristics for the MoS_2 FET. The p^{++}-Si/TiN/Pt/Al_2O_3 stack offers analog and nonvolatile memory where the threshold voltage of the FETs can be adjusted by applying a programming pulse to the back gate. i) Transfer characteristics of the post-programmed MoS_2 FET after applying negative programming voltages of different amplitudes for t_P = 100 ms. j) Retention characteristics, i.e., post-programmed I_{DS} vs time, measured at V_{BG} = 2 V. i) Transfer and k) corresponding retention characteristics of the post-erased MoS_2 FET after applying positive erase voltages of different amplitudes for t_E = 100 ms.

are difficult to realize using conductance states and require additional circuitry (note that earlier demonstration of SA[25] using resistive random access memory only implemented ferroelectric interactions); and second, the size of the [CS] matrix can become substantial even for relatively low values of K, imposing heavy area and energy overhead for the computation. Instead, the computational load can be significantly reduced by acknowledging the fact that only one spin is allowed to randomly flip during each iteration of SA, simplifying the computation of ΔH following Equation (3)

$$\Delta H = H_f - H_i = \left(-\frac{1}{2}\sum_{i,j} CS_{ij}\sigma_i\sigma'_j\right) - \left(-\frac{1}{2}\sum_{i,j} CS_{ij}\sigma_i\sigma_j\right) = \left[-\frac{1}{2}\Delta\sigma_j\right]\left[\sum_{i=1}^{K^2} CS_{ij}\sigma_i\right] \quad (3)$$

Here, we assume that the jth spin is flipped. Note that the first term inside the square bracket, i.e., $[-\frac{1}{2}\Delta\sigma_j]$ is computationally equivalent to the initial spin state of σ_j. For example, if σ_j flips from +1(–1) to –1(+1), $\Delta\sigma_j = -2(+2)$ and hence $[-\frac{1}{2}\Delta\sigma_j] = +1(-1)$. Therefore, following Equation (3), ΔH can be obtained just by finding the dot product of the spin vector $[\sigma]$ with the jth row vector of the [CS] matrix, i.e., $[CS]_j$, and by multiplying the result with σ_j. **Figure 3** shows the circuit modules used for computing ΔH. The multiplication module (M1) in Figure 3a is composed of a WSe_2 FET (T1) and MoS_2 FET (T2), connected in series with a common gate and a common source terminal. It multiplies the sign of two input voltages, V_{in-1} and V_{in-2}. V_{in-1} is applied to the common-gate terminal, V_{in-2} is applied to the drain terminal of T1, and $-V_{in-2}$ is applied to the drain terminal of T2. Note that T1 and T2 demonstrate dominant p-type and n-type conduction, respectively, as shown in Figure 3b. The transfer characteristics of M1, i.e., output voltage, V_{out}, vs V_{in-1} for V_{in-2} = ±0.1 V, are shown in Figure 3c. For V_{in-1} = 1 V, $V_{out} = V_{in-2}$, whereas for V_{in-1} = –1 V, $V_{out} = -V_{in-2}$. This is expected since for V_{in-1} = 1 V, T2 is more conductive than T1, thus allowing for V_{out} to follow V_{in-2} and vice versa for V_{in-1} = –1 V (Figure 3b). The nonvolatile memory of the MoS_2 FET is used here to ensure that the cross-point between the transfer characteristics of T1 and T2 occurs in the subthreshold

© 2021 Wiley-VCH GmbH

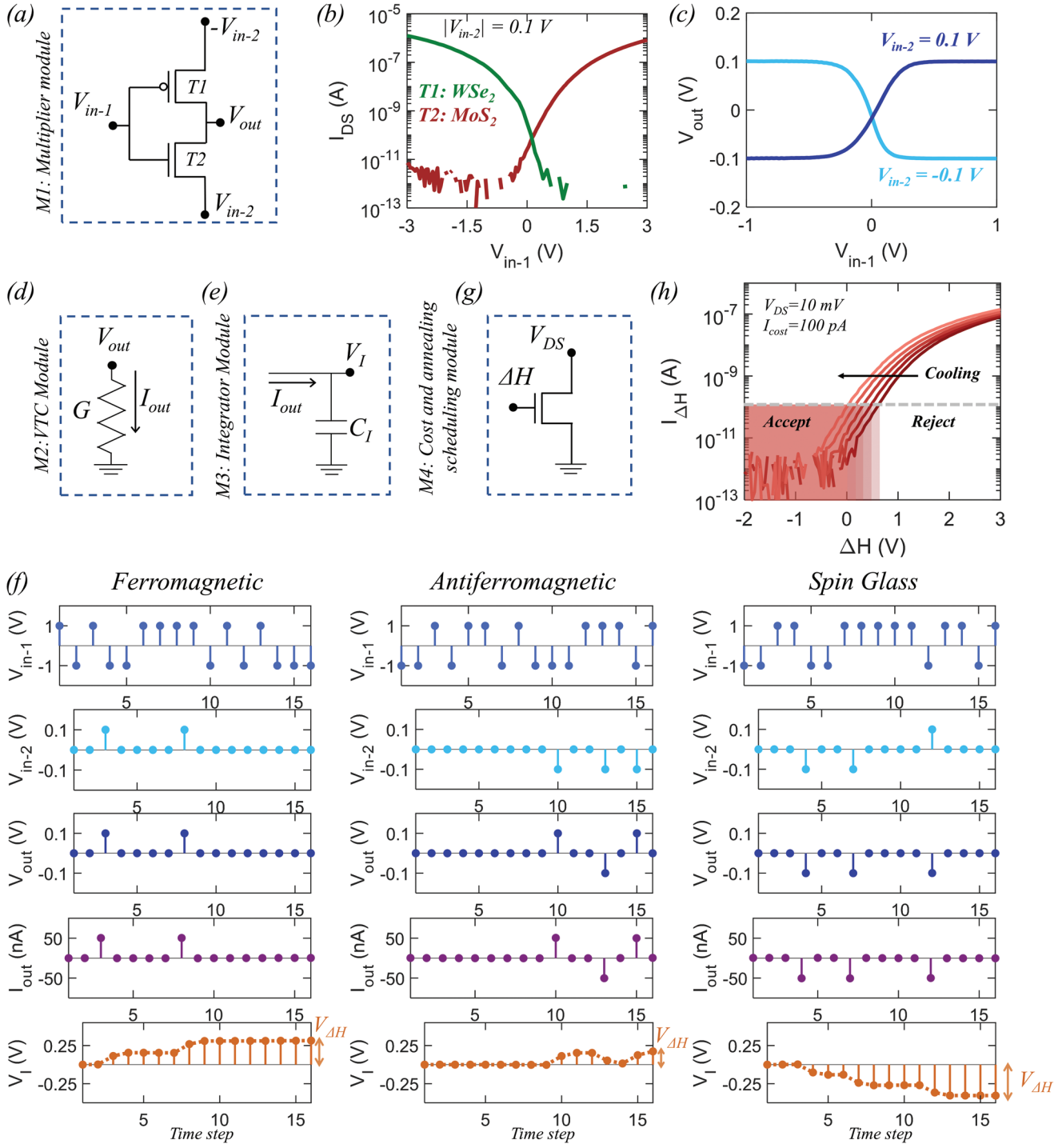


Figure 3. Circuit modules for hardware acceleration of SA. a) The multiplier module (M1) has a p-type WSe_2 FET (T1) and an n-type MoS_2 FET (T2) connected in series with a common gate and a common source terminal. It multiplies the sign of two input voltages, V_{in-1} and V_{in-2}. V_{in-1} is applied to the common-gate terminal, V_{in-2} is applied to the drain terminal of T1, and $-V_{in-2}$ is applied to the drain terminal of T2. b) Transfer characteristics of T1 and T2. c) Transfer characteristics of M1, i.e., output voltage, V_{out}, vs V_{in-1}, for $V_{in-2} = \pm 0.1$ V. Using M1, the product between the ith elements of $[\sigma]$ and $[CS]_j$ is obtained at the ith time step ($i\tau_p$) as $V_{out}(i\tau_p)$ by applying $V_{in-1}(i\tau_p) = V_1\sigma_i$ and $V_{in-2}(i\tau_p) = V_2 CS_{ij}$. Note that $i = 1{:}1{:}K^2$ and $K = 4$. We have used $\tau_p = 60$ ms, $V_1 = 1$ V, and $V_2 = 0.1$ V, resulting in $V_{out}(i\tau_p) = 0.1 \times CS_{ij}\sigma_i$. d) The voltage-to-current converter module (M2) transforms V_{out} from M1 into current, I_{out}, following $I_{out} = GV_{out}$ with $G \approx 0.5$ µS. e) The integrator module (M3), a capacitor ($C_I = 20$ nF), sums I_{out} from M2 over K^2 time steps into voltage, $V_{\Delta H}$. f) V_{in-1}, V_{in-2}, $-V_{in-2}$, V_{out}, I_{out}, and the output from M3, i.e., V_I, for representative ferromagnetic, antiferromagnetic, and spin glass systems during a given iteration of SA. $V_{\Delta H}$ and σ_j are multiplied to obtain ΔH. g) Schematic and h) transfer characteristics of a programmable MoS_2 FET used for evaluating the cost associated with the state transition as well as for realizing cooling schedule in hardware. The subthreshold conduction governed by Boltzmann statistics is exploited to evaluate the cost of "hill-climbing" by applying $V_{BG} = \Delta H$ and the spin flip is accepted if $I_{\Delta H} < I_{cost} = 100$ pA. The cooling schedule is implemented by shifting the threshold voltage of the FET through back-gate programming.

© 2021 Wiley-VCH GmbH

region, as shown in Figure 3b, and not in the OFF region. Having the cross-point in the OFF-state can lead to erroneous V_{out} values. Additionally, it is used to ensure that at V_{in-1} of ±1, V_{out} reaches $\pm V_{in-2}$ (Figure 3c). This necessitates at least 3–4 orders of magnitude difference in resistance between T1 and T2 at $V_{in-1} = \pm 1$, as the multiplier module is basically a voltage divider circuit and, thus, having comparable resistances for T1 and T2 can lead to the V_{out} being lower than $\pm V_{in-2}$. Using M1, the product between the ith elements of $[\sigma]$ and $[CS]_j$ is obtained at the ith time step ($i\tau_p$) as $V_{out}(i\tau_p)$ by applying $V_{in-1}(i\tau_p) = V_1\sigma_i$ and $V_{in-2}(i\tau_p) = V_2 CS_{ij}$. We have used $\tau_p = 60$ ms, $V_1 = 1$ V, and $V_2 = 0.1$ V, resulting in $V_{out}(i\tau_p) = 0.1 \times CS_{ij}\sigma_i$. Next, V_{out} is converted to I_{out} using a voltage-to-current converter module, M2 (Figure 3d), composed of a resistor, following $I_{out} = GV_{out}$ with $G \approx 0.5$ µS, where G is the conductance. Finally, the contribution due to all spins are summed over K^2 time steps using an integrator module, M3 (Figure 3e), composed of a capacitor ($C_I \approx 20$ nF), resulting in $V_{\Delta H}$ given by Equation (4)

$$V_{\Delta H} = \frac{\tau_p}{C_I}\sum_{i=1}^{K^2} I_{out}(i) = \frac{\tau_p}{C_I} G V_2 \sum_{i=1}^{K^2} CS_{ij}\sigma_i \tag{4}$$

Figure 3f shows V_{in-1}, V_{in-2}, $-V_{in-2}$, V_{out}, I_{out}, and the output of the integrator, V_I, for representative ferromagnetic, antiferromagnetic, and spin glass systems during a given iteration. Finally, $V_{\Delta H}$ and σ_j are multiplied in MATLAB to obtain ΔH.

Following the evaluation of ΔH, it must be determined if the spin flip is accepted. Traditionally, this is done using two separate steps: one to determine the sign of ΔH and another one to determine the cost of "hill-climbing" if ΔH is positive. In hardware, both steps are combined in M4 (Figure 3g) by exploiting the subthreshold conduction in a FET where the carrier injection from the source contact into the semiconducting channel is given by Boltzmann statistics. The cost of "hill-climbing" is, therefore, obtained by applying $V_{BG} = \Delta H$ and the spin flip is accepted if $I_{\Delta H} < I_{cost} = 100$ pA (Figure 3h) following Equation (5)

$$I_{\Delta H} = I_0 \exp\left(-\frac{q\Delta H}{mk_BT}\right) \tag{5}$$

Here, q is the electronic charge, I_0 is the static leakage current, and m is the body factor, which determines the SS following Equation (6)

$$SS = \frac{mk_BT}{q}\ln(10); \qquad m = \left(1 + \frac{C_S}{C_{OX}} + \frac{C_{IT}}{C_{OX}}\right) \tag{6}$$

Here, C_S is the semiconductor capacitance, C_{IT} is the interface trap capacitance, and C_{OX} is the oxide capacitance. For fully depleted ultrathin-body FETs, $C_S = 0$. We extracted a SS of 430 mV dec^{-1} and hence m of 7.1. Note that in Figure 3h the transfer characteristics are represented as a plot of $I_{\Delta H}$ versus ΔH. Also note that all negative ΔH (low-energy transition) naturally leads to $I_{\Delta H} < I_{cost}$, whereas positive ΔH up to ΔH_{max} satisfies the "hill-climbing" criterion of $I_{\Delta H} < I_{cost}$.

Next, we implement cooling schedule in hardware. While temperature can be used to the change the current $I_{\Delta H}$ for the same ΔH following Equation (5), physically changing the temperature of a system requires excessive energy. Alternatively, ΔH_{max} can be modulated by programming the MoS_2 FET in different states as shown in Figure 3h. As the annealing temperature is lowered, the transfer characteristics are shifted towards the left. This ensures acceptance of more positive ΔH at higher temperatures and no "hill climbing" at $T = 0$ K, i.e., $\Delta H_{max} = 0$ V. As shown in Figure 3h, our annealing schedule involved five different ΔH_{max} ,i.e., 0.6, 0.45, 0.3, 0.15, and 0 V, analogous to temperature in metallurgical annealing with maximum number of iterations, $I_{max} = 5$, 15, 15, 30, and 30, at the respective "temperatures".

The flow chart for the experimental demonstration of SA using our hardware-software approach is summarized in Figure S3 in the Supporting Information. The operations performed in hardware are highlighted in red. First, the spin system, i.e., $[\sigma]$ and [CS], and parameters of the simulated annealing algorithm such as I_{max} and ΔH_{max} are initiated in MATLAB. Using the random number generator in MATLAB, a random spin is flipped. ΔH due to this random spin flip is computed in hardware, using M1, M2 and M3. By applying ΔH to M4, $I_{\Delta H}$ is evaluated to check if the spin flip is accepted. After I_{max} iterations, ΔH_{max} is reduced according to the cooling schedule using M4 and the process is repeated for ΔH_{max} up to 0.

Figure 4a–c, respectively, show the representative experimental demonstration of SA leading to the convergence of randomly initiated ferromagnetic, antiferromagnetic, and a spin glass system to their respective ground states. See Figures S4–S6 and Videos S2-S4 in the Supporting Information for SA experiments performed on ferromagnetic, antiferromagnetic and a spin glass system initiated with 20 randomly oriented spin configurations. 11 ferromagnetic (55%), 9 antiferromagnetic (45%) and 8 spin glass systems (40%) converged and reached their respective ground states at the end of the total 95 iterations. Additionally, multiple systems are either 1 or 2 spin flips away from their ground state, and hence they are expected to converge for an increased number of iterations. We also observed "frustration" in the spin glass system. Remarkably, compared to an exhaustive search using BFT that requires a maximum of 2^{K^2} (= 65 536 for $K = 4$) spin flips, SA accelerates the search requiring orders of magnitude lower spin flip events. We define acceleration as the ratio of maximum number of spin flips using exhaustive search to the maximum number of SA spin flips to reach the ground state. The highest acceleration for ferromagnetic, antiferromagnetic and the spin glass system were found to be ≈1365×, ≈1260×, and ≈1310×, respectively, whereas, the average acceleration for the systems that converged to their ground states were ≈850×, ≈900×, and ≈810×, respectively. Evolution of H for representative ferromagnetic, antiferromagnetic, and spin glass systems are shown in Figure 4d–f, respectively (see Figure S7 in the Supporting Information for the evolution H for all 60 spin systems used in our experiments). The signature of SA can be seen in the energy landscapes, i.e., significant "hill-climbing" occurs during the initial iterations when the system is at higher "temperature", whereas at the lowest "temperature", the free energy decreases monotonically.

To obtain further insight, hardware-realistic simulation of SA is performed using the virtual source (VS) model developed in our earlier work to capture 2D FET characteristics.[46,47] **Figure 5**a–c, respectively, show the convergence accuracy of

© 2021 Wiley-VCH GmbH

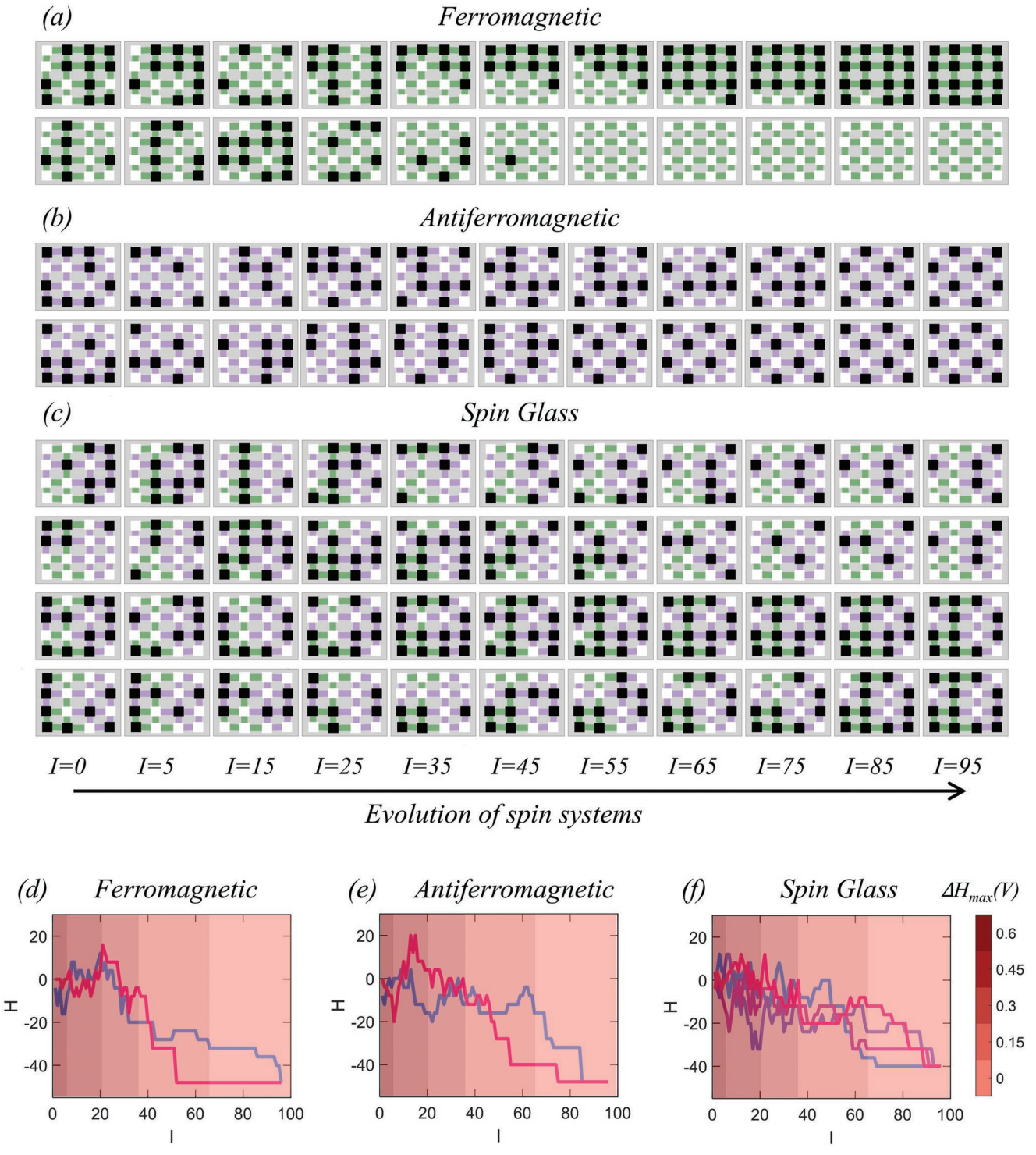


Figure 4. Experimental demonstration of SA. Randomly initiated a) ferromagnetic, b) antiferromagnetic, and c) spin glass systems converge to their ground states as successive iterations (I) are performed using SA in hardware. Compared to an exhaustive search using brute force trial (BFT) that requires a maximum of 2^{K^2} (= 65 536 for $K = 4$) spin flips, SA accelerates the search by >800×. d–f) Evolution of H with I for ferromagnetic (d), antiferromagnetic (e), and spin glass (f) systems for different ground states. The signature of SA can be seen in the energy landscapes. At higher "temperature", more "hill-climbing" is allowed, whereas at the lowest "temperature", the free energy decreases without any "hill-climbing".

1000 randomly initiated 4 × 4 ferromagnetic, antiferromagnetic, and spin glass systems subjected to SA for different total numbers of iterations (I). All ferromagnetic and antiferromagnetic systems converged after ≈600 and ≈750 iterations, resulting in average accelerations of ≈110× and ≈90×, respectively, compared to exhaustive BFTs. Note that these average acceleration

© 2021 Wiley-VCH GmbH

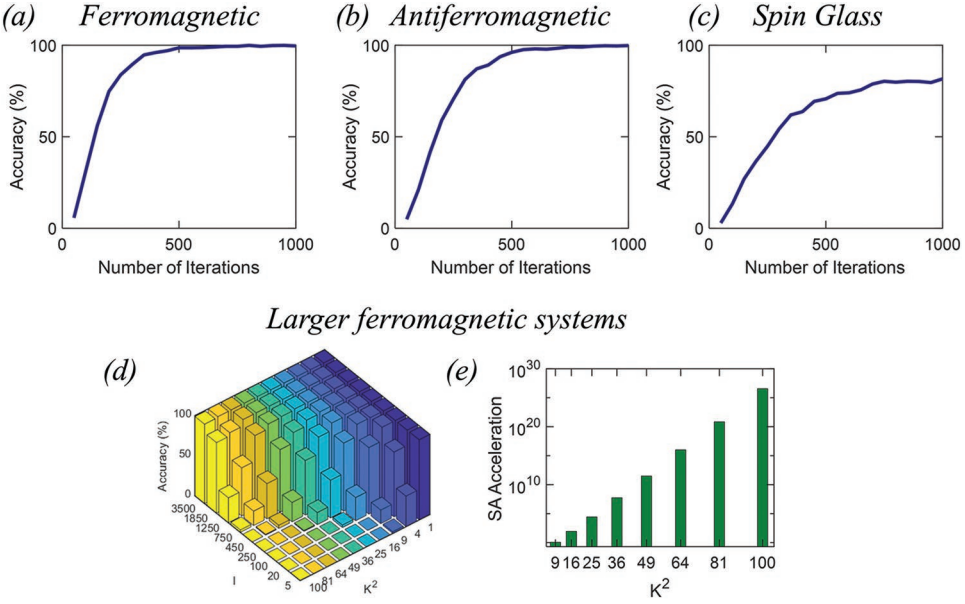


Figure 5. Hardware-realistic simulation of SA using the virtual source model. a–c) Convergence accuracy of 1000 randomly initiated 4 × 4 ferromagnetic (a), antiferromagnetic (b), and spin glass (c) systems subjected to SA for different total *I*) d) Convergence accuracy for 100 randomly oriented $K \times K$ ferromagnetic spin systems as a function of total number of spins (K^2) and total number of SA iterations. e) Acceleration factor, i.e., the ratio of maximum number of spin flips using BFT to the maximum number of SA spin flips for 100% convergence accuracy, as a function of K^2.

numbers are lower than the experimental findings since more iterations are performed (200 at each temperature). The spin glass system, however, shows ≈80% convergence accuracy after ≈700 iterations. This is owing to the fact that any spin glass system is more prone to getting stuck in a metastable state. Figure 5d shows the convergence accuracy for 100 randomly oriented $K \times K$ ferromagnetic spin systems as a function of total number of spins (K^2) and total number of SA iterations. Even for $K = 10$, SA requires only ≈3500 iterations, or maximum number of spin flips, in comparison to $\approx 10^{30}$ maximum number of spin flips required for an exhaustive search, demonstrating the tremendous improvement in acceleration ($\approx 10^{26}\times$). Figure 5e shows the acceleration for ≈100% convergence accuracy as a function of K^2. Clearly, as K increases the benefits of SA become even more astounding. Additionally, note that hardware modules don't need to scale with K. The same modules can be implemented for larger spin systems. The only difference is that, in the case of a $K \times K$ system, the length of the vectors such as V_{in-1}, V_{in-2}, $-V_{in-2}$, V_{out}, and I_{out} that are passing through various modules would be $1 \times K^2$.

Finally, we evaluate the energy expenditure (E_{SA}) by our annealing accelerator during each iteration following Equation (7)

$$E_{SA} = E_{M1} + E_{M2} + E_{M3} + E_{M4} = \sum_{i=1}^{K^2} \left\{ \left[(I_{T1} + I_{T2}) V_{in-2} \tau_p \right]_{M1} + \left[I_{out} V_{out} \tau_p \right]_{M2} + \left[\frac{1}{2} C (\Delta V_I)^2 \right]_{M3} + \left[I_{\Delta H} V_{DS} \tau_p \right]_{M4} \right\} \quad (7)$$

Here, E_{M1}, E_{M2}, E_{M3}, and E_{M4} are the energy consumption by M1, M2, M3, and M4, respectively, and I_{T1} and I_{T2} are the current in T1 and T2, respectively. Note that this calculation does not take parasitic capacitances in our FETs into account. Figure S8a in the Supporting Information shows E_{M1}, E_{M2}, E_{M3}, E_{M4}, and E_{SA} averaged over all 60 spin systems as a function of I for a 4 × 4 spin lattice. The average energy expenditure for the hardware module was found to be a miniscule ≈1.3 nJ per iteration, which corresponds to a maximum total energy expenditure of ≈120 nJ for finding the ground state of any 4 × 4 spin system. Note that due to the limitations imposed by our measurement instruments, we have used $\tau_p = 60$ ms. However, it is possible to scale τ_p and thereby reduce the energy expenditure even further. Also note that our energy calculations exclude the software operations performed using MATLAB such as the generation of random numbers.

We also attempt to estimate the energy consumption for larger spin systems. Since only the neighbor interactions are considered, out of these K^2 timesteps for a $K \times K$ system, ideally, there would be a maximum of 4 spikes and the rest would be zero. However, for the hardware implementation a nonideality is seen. When $V_{in-2} = 0$, V_{out} must be zero for the multiplier module. However, in practice, when $V_{in-2} = 0$, $V_{out} \approx 10^{-4}$. This non ideality can carry into the other modules and lead to higher power consumption for larger systems. From Figure S8a in the Supporting Information, M2 and M3, i.e., the VTC module and the integrator module, dominate energy consumption. Hence, we estimate the worst-case energy consumption for spin systems with K ranging from 1 to 1000, for both modules M3 and M4, as shown in Figure S8b in the Supporting Information. Figure S8b in the Supporting Information shows that the increase in energy consumption due to this non ideality is minimal even for very large spin systems, ensuring low energy consumption for larger spin systems. Figure S8b in the Supporting Information represents the energy consumption per iteration. Hence, note that there would be an obvious increase in energy associated with the larger number of iterations needed for larger spin systems.

© 2021 Wiley-VCH GmbH

5. Conclusion

This work successfully demonstrates hardware acceleration of SA for the Ising spin system by exploiting subthreshold conduction and analog programmability of complementary 2D FETs integrated with nonvolatile floating-gate memory stack. By designing in-memory computing primitives and annealing schedule equivalent of cooling, we were able to achieve >800× acceleration for 4 × 4 ferromagnetic, antiferromagnetic, and spin glass systems experimentally and at a frugal average energy expenditure of ≈120 nJ. Our numerical simulations show more striking benefits of SA for search acceleration of larger spin lattices.

6. Experimental Section

Device Fabrication: Back-gated MoS_2 and WSe_2 FETs were fabricated using e-beam lithography. MOCVD grown MoS_2 was transferred onto 50 nm Al_2O_3 substrate using a PMMA (poly(methyl methacrylate)) assisted wet-transfer process. The substrate was spin-coated with PMMA and baked at 180 °C for 90 s to define the channel region. The PMMA photoresist was then exposed to an e-beam and developed using a 1:1 4-methyl-2-pentanone (MIBK) and 2-propanol (IPA) mixture. Using sulfur hexafluoride (SF_6) at 5 °C for 30 s, the monolayer MoS_2 film was subsequently etched. Next, the sample was rinsed in acetone and IPA to remove the photoresist. In order to fabricate the source/drain contacts, the substrate was spin-coated with MMA and PMMA and again patterned using e-beam lithography and developed using MIBK and IPA. E-beam evaporation was then used to deposit a 40 nm Ni/30 nm Au stack to serve as the contacts. Finally, the photoresist was rinsed away in a lift-off process using acetone and IPA.

For WSe_2, micromechanical exfoliation was performed to obtain optimally thin WSe_2 flakes on the 50 nm Al_2O_3 substrate. The source/drain contacts (10 nm Pt/30 nm Au) were defined using e-beam lithography, as discussed above, with a channel length of 1 µm. Following that, in order to fabricate the p–i–p structure, the channel was spin-coated with PMMA and a subsequent e-beam exposure was used to expose 250 nm of the channel near the source and drain contact, leaving the middle 500 nm covered with PMMA. The WSe_2 FET was further doped with O_2 plasma using a Tepla M4L plasma etch tool at a power of 100 W for 300 s. O_2 and He gas flow rates of 150 sccm and 50 sccm with a chamber pressure of 500 mT were used for the O_2 plasma doping. Finally, the photoresist was rinsed away in a lift-off process using acetone and IPA.

Electrical Characterization: A Lake Shore CRX-VF probe station and a Keysight B1500A parameter analyzer were used to perform the electrical characterization at room temperature in high vacuum (≈10^{-6} Torr). The measurements with the resistor and capacitor modules were performed outside the probe station on a bread board.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

The work was supported by Army Research Office (ARO) through Contract Number W911NF1920338. The authors also acknowledge Andrew J Arnold for help with WSe_2 device fabrication. The 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.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

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

Data Availability Statement

Research data are not shared.

Keywords

2D materials, field-effect transistors, Ising spin lattices, simulated annealing

Received: September 7, 2021
Revised: October 20, 2021
Published online:

[1] E. L. Lawler, *The Travelling Salesman Problem: A Guided Tour of Combinatorial Optimization*, John Wiley & Sons, Hoboken, NJ, USA **1985**.
[2] J. R. A. Allwright, D. B. Carpenter, *Parallel Comput.* **1989**, *10*, 335.
[3] A. Singh, A. S. Baghel, *Soft Comput.* **2008**, *13*, 95.
[4] S. Nallaperuma, M. Wagner, F. Neumann, *Front. Rob. AI* **2015**, *2*, https://doi.org/10.3389/frobt.2015.00018.
[5] S. Gavrilov, D. Zheleznikov, V. Khvatov, R. Chochaev, in *2018 IEEE Conf. of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, IEEE, Piscataway, NJ, USA **2018**, pp. 1492–1495, https://doi.org/10.1109/EIConRus.2018.8317380.
[6] Z. Wang, Y. Zhao, Y. Liu, C. Lv, *Concurrency Comput.: Pract. Exp.* **2018**, *30*, 4429.
[7] X. Xiao, Y. Liu, H. Song, T. Kikkawa, *Int. J. RF Microwave Comput.-Aided Eng.* **2020**, *30*, e22364.
[8] L. M. R. Rere, M. I. Fanany, A. M. Arymurthy, *Proc. Comput. Sci.* **2015**, *72*, 137.
[9] A. Biswas, T. Acharya, *Model. Earth Syst. Environ.* **2016**, *2*, https://doi.org/10.1007/s40808-016-0256-x.
[10] F. Neri, in *Trends and Advances in Information Systems and Technologies. WorldCIST'18 2018. Advances in Intelligent Systems and Computing*, Vol. *746*, (Eds: Á. Rocha, H. Adeli, L. Reis, S. Costanzo), Springer Nature, Cham, Switzerland **2018**, pp. 755–763.
[11] S. V. Isakov, I. N. Zintchenko, T. F. Rønnow, M. Troyer, *Comput. Phys. Commun.* **2015**, *192*, 265.
[12] T. Leleu, Y. Yamamoto, S. Utsunomiya, K. Aihara, *Phys. Rev. E* **2017**, *95*, 022118.
[13] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller, *J. Chem. Phys.* **1953**, *21*, 1087.
[14] S. Kirkpatrick, C. D. Gelatt Jr., M. P. Vecchi, *Science* **1983**, *220*, 671.
[15] A. K. Peprah, S. K. Appiah, S. K. Amponsah, *Appl. Math.* **2017**, *08*, 1195.

© 2021 Wiley-VCH GmbH

[16] S. Geman, D. Geman, *IEEE Trans. Pattern Anal. Mach. Intell.* **1984**, *6*, 721.
[17] G. Dueck, T. Scheuer, *J. Comput. Phys.* **1990**, *90*, 161.
[18] J. Niittylahti, H. Raittinen, K. Kaski, in *Proc. IEEE Winter Workshop on Nonlinear Digital Signal Processing*, IEEE, Piscataway, NJ, USA **1993**, https://doi.org/10.1109/NDSP.1993.767727.
[19] D. Abramson, *Computer* **1992**, *25*, 27.
[20] A. M. Ferreiro, J. A. García, J. G. López-Salas, C. Vázquez, *J. Glob. Optim.* **2012**, *57*, 863.
[21] C. Cook, H. Zhao, T. Sato, M. Hiromoto, S. X. D. Tan, *Integration* **2019**, *69*, 335.
[22] C. Yoshimura, M. Hayashi, T. Okuyama, M. Yamaoka, in *2016 Fourth Int. Symp. on Computing and Networking (CANDAR)*, IEEE, Piscataway, NJ, USA **2016**, pp. 436–442, https://doi.org/10.1109/CANDAR.2016.0081.
[23] A. HajiRassouliha, A. J. Taberner, M. P. Nash, P. M. F. Nielsen, *Signal Proc.: Image Commun.* **2018**, *68*, 101.
[24] H. Ushijima-Mwesigwa, C. F. A. Negre, S. M. Mniszewski, in *Int. Conf. for High Performance Computing, Networking, Storage and Analysis (PMES)*, ACM, New York, **2017**, pp. 22–29, https://doi.org/10.1145/3149526.3149531.
[25] J. H. Shin, Y. J. Jeong, M. A. Zidan, Q. Wang, W. D. Lu, in *2018 IEEE Int. Electron Devices Meeting (IEDM)*, IEEE, Piscataway, NJ, USA **2018**, https://doi.org/10.1109/IEDM.2018.8614698.
[26] K. Yang, Q. Duan, Y. Wang, T. Zhang, Y. Yang, R. Huang, *Sci. Adv.* **2020**, *6*, eaba9901.
[27] F. Cai, S. Kumar, T. Van Vaerenbergh, X. Sheng, R. Liu, C. Li, Z. Liu, M. Foltin, S. Yu, Q. Xia, J. J. Yang, R. Beausoleil, W. D. Lu, J. P. Strachan, *Nat. Electron.* **2020**, *3*, 409.
[28] M. R. Mahmoodi, H. Kim, Z. Fahimi, H. Nili, L. Sedov, V. Polishchuk, D. B. Strukov, in *2019 IEEE Int. Electron Devices Meeting (IEDM)*, IEEE, Piscataway, NJ, USA **2019**, https://doi.org/10.1109/IEDM19573.2019.8993442.
[29] Q. Smets, G. Arutchelvan, J. Jussot, D. Verreck, I. Asselberghs, A. N. Mehta, in *2019 IEEE Int. Electron Devices Meeting (IEDM)*, IEEE, Piscataway, NJ, USA **2019**, https://doi.org/10.1109/IEDM19573.2019.8993650.
[30] A. Sebastian, R. Pendurthi, T. H. Choudhury, J. M. Redwing, S. Das, *Nat. Commun.* **2021**, *12*, 693.
[31] S. Wachter, D. K. Polyushkin, O. Bethge, T. Mueller, *Nat. Commun.* **2017**, *8*, 14948.
[32] D. Jayachandran, A. Oberoi, A. Sebastian, T. H. Choudhury, B. Shankar, J. M. Redwing, S. Das, *Nat. Electron.* **2020**, *3*, 646.
[33] H. Jang, C. Liu, H. Hinton, M. H. Lee, H. Kim, M. Seol, H.-J. Shin, S. Park, D. Ham, *Adv. Mater.* **2020**, *32*, 2002431.
[34] G. M. Marega, Y. Zhao, A. Avsar, Z. Wang, M. Tripathi, A. Radenovic, A. Kis, *Nature* **2020**, *587*, 72.
[35] H. Maletta, W. Felsch, *Phys. Rev. B* **1979**, *20*, 1245.
[36] M. K. Singh, W. Prellier, M. P. Singh, R. S. Katiyar, J. F. Scott, *Phys. Rev. B* **2008**, *77*, 144403.
[37] J. Vannimenus, G. Toulouse, *J. Phys. C: Solid State Phys.* **1977**, *10*, L537.
[38] S. F. Edwards, P. W. Anderson, *J. Phys. F: Met. Phys.* **1975**, *5*, 965.
[39] S. Das, J. Appenzeller, *Appl. Phys. Lett.* **2013**, *103*, 103501.
[40] A. J. Arnold, D. S. Schulman, S. Das, *ACS Nano* **2010**, *14*, 13557.
[41] D. S. Schulman, A. J. Arnold, S. Das, *Chem. Soc. Rev.* **2018**, *47*, 3037.
[42] A. Wali, S. Kundu, A. J. Arnold, G. Zhao, K. Basu, S. Das, *ACS Nano* **2021**, *15*, 3453.
[43] A. Sebastian, R. Pendurthi, T. H. Choudhury, J. M. Redwing, S. Das, *Nat. Commun.* **2021**, *12*, 693.
[44] A. Dodda, A. Oberoi, A. Sebastian, T. H. Choudhury, J. M. Redwing, S. Das, *Nat. Commun.* **2020**, *11*, 4406.
[45] S. Das, H. Y. Chen, A. V. Penumatcha, J. Appenzeller, *Nano Lett.* **2013**, *13*, 100.
[46] A. Sebastian, A. Pannone, S. S. Radhakrishnan, S. Das, *Nat. Commun.* **2019**, *10*, 4199.
[47] S. Das, A. Dodda, S. Das, *Nat. Commun.* **2019**, *10*, 3450.

© 2021 Wiley-VCH GmbH

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-22332-8 OPEN

A biomimetic neural encoder for spiking neural network

Shiva Subbulakshmi Radhakrishnan [1], Amritanand Sebastian [1], Aaryan Oberoi [1], Sarbashis Das [2] & Saptarshi Das [1,3,4]✉

Spiking neural networks (SNNs) promise to bridge the gap between artificial neural networks (ANNs) and biological neural networks (BNNs) by exploiting biologically plausible neurons that offer faster inference, lower energy expenditure, and event-driven information processing capabilities. However, implementation of SNNs in future neuromorphic hardware requires hardware encoders analogous to the sensory neurons, which convert external/internal stimulus into spike trains based on specific neural algorithm along with inherent stochasticity. Unfortunately, conventional solid-state transducers are inadequate for this purpose necessitating the development of neural encoders to serve the growing need of neuromorphic computing. Here, we demonstrate a biomimetic device based on a dual gated MoS_2 field effect transistor (FET) capable of encoding analog signals into stochastic spike trains following various neural encoding algorithms such as rate-based encoding, spike timing-based encoding, and spike count-based encoding. Two important aspects of neural encoding, namely, dynamic range and encoding precision are also captured in our demonstration. Furthermore, the encoding energy was found to be as frugal as ≈1–5 pJ/spike. Finally, we show fast (≈200 timesteps) encoding of the MNIST data set using our biomimetic device followed by more than 91% accurate inference using a trained SNN.

[1] Department of Engineering Science and Mechanics, Pennsylvania State University, University Park, PA, USA. [2] Department of Electrical Engineering, Pennsylvania State University, University Park, PA, USA. [3] Department of Materials Science and Engineering, Pennsylvania State University, University Park, PA, USA. [4] Materials Research Institute, Pennsylvania State University, University Park, PA, USA. ✉email: sud70@psu.edu

Billions of neurons coupled through trillions of synapses form the complex computational unit of the brain, that can simultaneously process a massive amount of information received from external/internal stimuli through various sensory organs. These neurons use action potential or spikes as the common language to speak with each other and to compute for learning and decision-making. Spikes are stereotypical electrical impulses or all-or-none (digital) point events in time, that allow long-distance neural communication and energy-efficient neural computation. However, external stimuli such as light, sound, smell, temperature, etc., and internal stimuli such as blood pressure, oxygen levels, feeling of pain and hunger, etc. that are primarily analog continuous variables in time. It requires specialized sensory neurons, also known as afferent neurons, to transform the specific type of analog stimulus into corresponding spike trains following one or more neural encoding algorithms, and subsequently relay the spike-encoded information to the central nervous system for processing. This process is referred to as sensory transduction. For example, in the auditory neural pathways, mechanoelectrical transduction is mediated by the hair cells within the ear[1]. Similarly, gustatory receptors in taste buds interact with chemicals in food to produce action potentials[2,3]. Phototransduction in the retina is mediated by rods and cones and eventually converted to spikes by the ganglion cells[4,5]. Sensory transduction also exhibits inherent stochasticity, which allows neurons to process information with better noise tolerance and energy efficiency[6,7].

The diversity of neurobiological architectures and neural computational algorithms found inside even the simplest of animal brains continue to fascinate computer scientists and electronic device engineers. Neuromorphic computing pioneered by Carver Mead and colleagues is a branch of research that aims to mimic the computational power of the brain on a chip[8,9]. Unfortunately, the initial growth in neuromorphic computing was rather slow owing to the contemporary dominance of von Neuman architecture, and the success of the complementary metal-oxide-semiconductor (CMOS) technology. However, the recent demise in scaling and fundamental limitations of von Neuman computing is fueling the resurgence of bio-inspired neuromorphic hardware[10–12]. Artificial neural networks (ANNs) are the most prevalent form of neuromorphic computing that have already demonstrated breakthrough progress in many fields[13]. ANNs consist of multiple layers with each layer comprising of collection of computational units, called artificial neurons, which are connected through artificial synapses. While the models of cortical hierarchies from biological neural networks (BNNs) have been mimicked through deep learning[14] in ANNs with a massive number of computational layers, only marginal similarities with brain-like computing can be recognized at the implementation level. The most obvious difference is that artificial neurons receive, process, and transmit analog information in continuous time, whereas biological neurons use action potential or spikes. Also, stochasticity is an inherent neural phenomenon, which is typically ignored by most ANNs. Spiking neural network (SNN) promises to bridge this gap by adopting a new computing paradigm based on biologically plausible neurons[15,16]. In fact, the past few years have seen tremendous progress in the development of SNNs offering unprecedented energy efficiency and faster inference owing to event-driven computation[17]. However, hardware realization of SNNs necessitates the development of neural encoders since conventional sensors are incapable of converting sensory input into spike trains.

Here, we report a biomimetic device based on a dual gated MoS_2 field effect transistor (FET) with a stochastic sampling terminal capable of encoding analog signals, for example illuminance levels of a light emitting diode (LED), into corresponding spike trains. We are also able to implement various neural encoding algorithms, such as rate-based encoding, spike timing-based encoding, and spike count-based encoding. Two key features of neural encoding, namely, dynamic range and encoding precision are also captured in our demonstration. Finally, we show a fast and accurate inference of spike encoded MNIST data set using a trained spiking neural network (SNN) with inference accuracy of more than 91%. Remarkably, energy consumption by our biomimetic neural encoder was found to be as frugal as ≈1–5 pJ/spike.

Results

The overall philosophy of biomimetic neuromorphic computing is shown in Fig. 1. Figure 1a shows the schematic of a biological neural network that involves sensory transduction of analog stimulus to corresponding spike trains by specialized neurons and subsequent processing by the central nervous system. For example, external optical stimuli are converted into corresponding graded potentials by the photoreceptor cells (rods and cones) in the human eyes followed by neural encoding into spike trains using the ganglion cells, and eventual processing of the encoded visual stimulus by the visual cortex. Figure 1b shows the corresponding neuromorphic hardware comprising of neuromorphic sensors, neuromorphic encoders, and neuromorphic processors. Figure 1c shows the schematic of our experimental demonstration with a white light-emitting diode (LED) as the visual stimulus, a silicon (Si) photodiode (PD) as the sensor, a dual gated MoS_2-based FET as the neuromorphic encoder, and a trained spiking neural network (SNN) as the neuromorphic processor.

Biomimetic neural encoder and neuromorphic transducer. We have used multilayer exfoliated MoS_2 that belongs to the family of two-dimensional (2D) layered materials[18–20] as the semiconducting channel, 285 nm SiO_2 on p^{++}-Si as the back-gate stack, and 120 nm of hydrogen silsesquioxane (HSQ) as the top-gate dielectric for the fabrication of the biomimetic neural encoder as shown schematically in Fig. 2a. The thickness of the MoS_2 flake is ≈5 nm. The optical image of the device is shown in Fig. 2b. Source, drain, and gate metal stacks were patterned using electron-beam lithography followed by the deposition of 40 nm of Nickel (Ni) and 30 nm of gold (Au) using electron-beam evaporation. More details on device fabrication can be found in the "Methods" section and in our prior work[21,22]. The channel length and width were ≈1 μm and ≈2.8 μm, respectively. Note that the use of ultrathin body MoS_2 as the semiconducting channel material is motivated by the growing interest in 2D materials, as a successor to Si as well as their promising use in neuromorphic and biomimetic devices[12,22–25]. Furthermore, various types of sensors such as photodetectors[26], chemical sensors[27], biological sensors[27], touch sensors[28], and radiation sensors[29] have been demonstrated using MoS_2 based devices, which allow direct integration of sensors and encoders in future neuromorphic hardware. The presynaptic signal obtained from the neuromorphic sensors such as the Si PD is applied to the back-gate terminal as analog voltage (V_{PSV}), whereas, encoded information in form of postsynaptic current spikes (I_{PSC}) is obtained at the drain terminal. The top-gate voltage (V_{TG}) is applied as a sequence of sampling pulses, with a pulse duration (t_p) of 10 ms and amplitude determined based on the desired encoding algorithm.

Figure 2c shows the transfer function of the neural encoder i.e., I_{PSC} vs. V_{PSV} measured at a drain bias, $V_{DS} = 1$ V, for different V_{TG}. The n-type unipolar characteristics is common for MoS_2 FETs[30,31]. The monotonic positive shift in the transfer function with decreasing V_{TG} can be explained form the principle of charge balance, i.e., the inversion charge induced by positive V_{PSV}

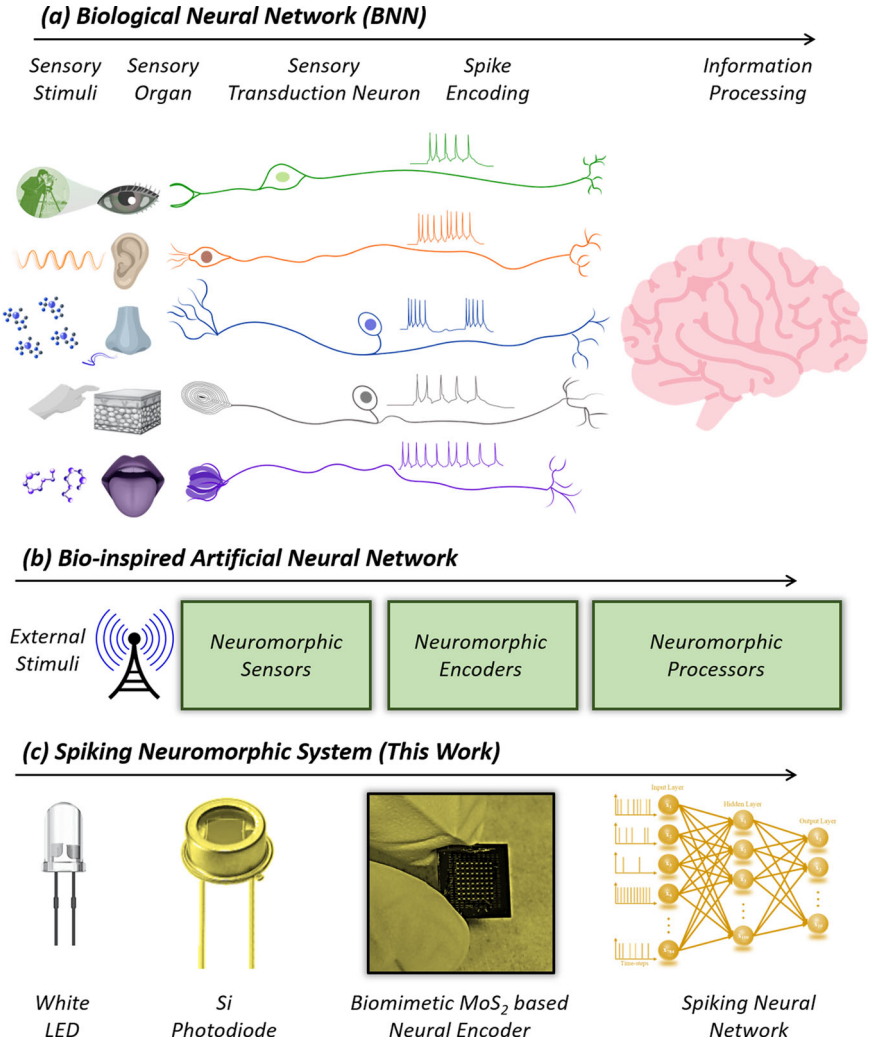


Fig. 1 Biomimetic neuromorphic computing. a Schematic of biological neural network (BNN). Specialized neurons convert analog external/internal stimuli into corresponding spike trains by a process called sensory transduction and subsequent relay the information to the central nervous system for further processing. Various encoding algorithms are found in sensory neurobiology such as spike rate-based, spike count-based, and spike timing-based encoding. **b** Schematic of bio-inspired artificial neural network comprising of neuromorphic sensors, neuromorphic encoders, and neuromorphic processors. **c** Schematic of our spiking neuromorphic system consisting of a white light emitting diode (LED) as the visual stimulus, a silicon (Si) photodiode (PD) as the sensor, a dual gated MoS_2 based field effect transistors (FETs) with a stochastic terminal as the neuromorphic encoder, and a trained spiking neural network (SNN) as the neuromorphic processor.

is compensated by the negative V_{TG} and vice versa. Figure 2d shows the spiking threshold (V_{ST}), which we define as V_{PSV} required to invoke a current spike, i.e., $I_{PSC} > I_{ST}$ as a function of V_{TG}, and for different thresholding current, I_{ST}. As expected, the spiking threshold is higher for more negative V_{TG} and higher I_{ST}. Note that the slope of V_{ST} vs. V_{TG} is constant, irrespective of I_{ST} and is proportional to the ratio of back-gate capacitance to top-gate capacitance, i.e., C_{TG}/C_{BG} which was found to be ≈2.2, consistent with the thicknesses of ≈120 nm and ≈285 nm and dielectric constants of ≈3.2 and ≈3.9 of HSQ and SiO_2, respectively. Also, note that the presynaptic terminal (back-gate) and encoding terminal (top-gate) can be interchanged (see Supplementary Fig. 1 and Supplementary Note 1). Figure 2e shows the circuit schematic for phototransduction comprising of a Si PD and a load resistor (R_L). Figure 2f shows the current (I_{PD}) vs. voltage (V_{PD}) characteristics of the neuromorphic sensor, i.e. Si PD for different intensities of the visual stimulus, i.e., LED illuminance (P_{LED}). Finally, Fig. 2g shows the phototransduction characteristics of the Si PD and corresponding V_{PSV} applied to the neuromorphic encoder as a function of P_{LED} obtained using R_L.

Hardware acceleration of various neural encoding algorithms. Next, we implement various neural encoding algorithms, found in sensory neurobiology, using our biomimetic encoder for translating analog V_{PSV} values obtained for different LED illuminations shown in Fig. 3a into corresponding spike trains. The most popular encoding principle is rate-based encoding, originally demonstrated by Adrian and Zotterman[32] using an electrophysiological experiment in sensory nerve fibers of frog muscles. In rate encoding, it is postulated that the information about the stimulus is contained in the firing rate of the neuron, and not in individual spikes. This is more so because the sequence of spikes generated by the neurons in response to a given stimulus varies from trial to trial and over time owing to the inherent stochasticity in sensory transduction, whereas the mean firing rate, i.e., inverse of interspike interval remains practically constant. Numerous studies in sensory and motor systems of various species have validated the spike rate encoding hypothesis. Based on these observations, rate encoding is widely used for SNNs. For rate-based encoding using our biomimetic encoder, the magnitude of V_{TG} pulses are randomly sampled from a Gaussian

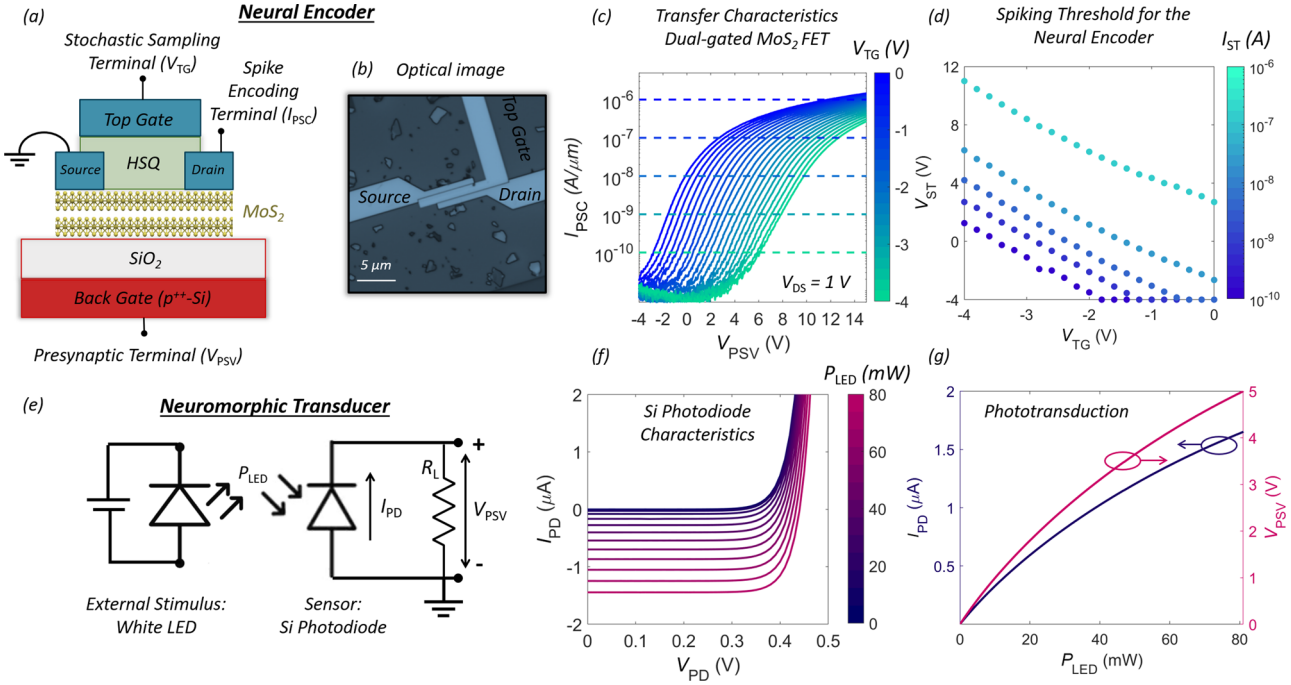


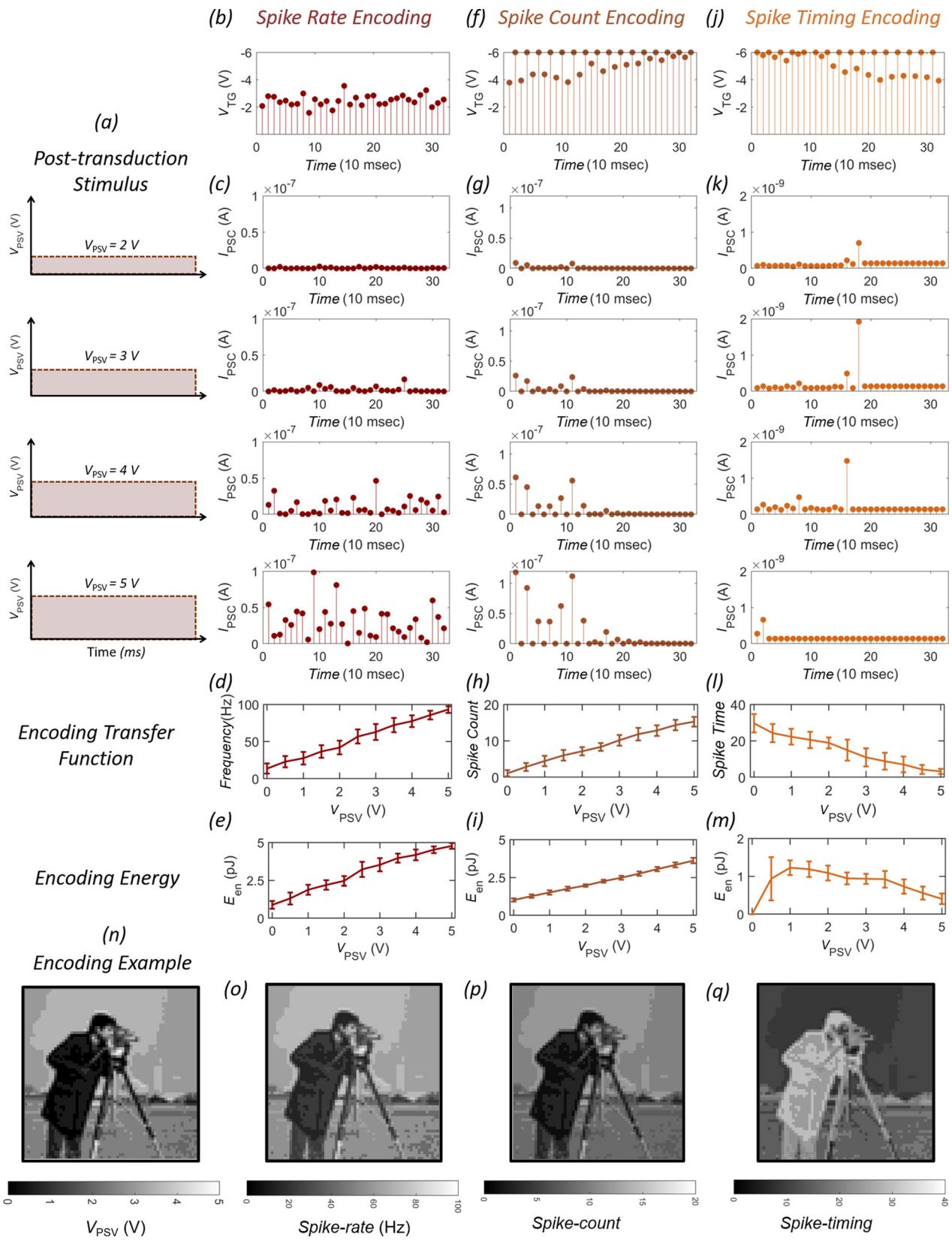
Fig. 2 Biomimetic neural encoder and neuromorphic transducer. a Schematic and **b** optical image of our biomimetic neural encoder. We have used multilayer exfoliated MoS_2 as the semiconducting channel, 285 nm SiO_2 on p^{++}-Si as the back-gate stack, and 120 nm of hydrogen silsesquioxane (HSQ) as the top-gate dielectric for the fabrication of the encoder. The channel length and width are ≈1 μm and ≈2.8 μm, respectively. Presynaptic signal obtained from the neuromorphic sensors is applied to the back-gate terminal as analog voltage (V_{PSV}), whereas encoded information in form of postsynaptic current spikes (I_{PSC}) is obtained at the drain terminal. The top-gate voltage (V_{TG}) is applied as a sequence of sampling pulses, with a pulse duration of $t_p = 10$ ms and amplitude determined based on the desired encoding algorithm. **c** Transfer function of the neural encoder i.e., I_{PSC} vs. V_{PSV} measured at a drain bias, $V_{DS} = 1$ V, for different V_{TG}. **d** Spiking threshold (V_{ST}), which we define as V_{PSV} required to invoke a current spike, i.e., $I_{PSC} > I_{ST}$ as a function of V_{TG}, for different thresholding current, I_{ST}. As expected, the spiking threshold is higher for more negative V_{TG} and higher I_{ST}. **e** An example of neuromorphic transducer comprising of Si PD and a parallel load resistor (R_L). The Si PD transduces light from the LED into photocurrent, which is converted to V_{PSV} by R_L. **f** Current (I_{PD}) vs. voltage (V_{PD}) characteristics of the Si PD for different LED illuminance (P_{LED}). **g** Phototransduction characteristics of the Si PD and corresponding presynaptic voltage (V_{PSV}) applied to the neuromorphic encoder as a function of P_{LED}.

distribution with mean, $\mu_{TG} = -2.5$ V, and standard deviation $\sigma_{TG} = 0.8$ V as shown in Fig. 3b and the responses corresponding to each V_{PSV} value with $I_{ST} = 500$ pA is displayed in Fig. 3c (see "Methods" section for discussion on current-sampling method). During each trial, V_{TG} pulses were sampled for N number of times (=32) and a total of 16 trials were conducted resulting in 512 sampling points for each V_{PSV}. See Supplementary Fig. 2 for the complete circuit used to obtain neural encoding for different illumination levels. Also, see Supplementary Movie 1 for real-time encoding of different LED intensities into stochastic spike trains. Figure 3d shows the encoding transfer function, i.e., mean firing rate (inverse of the mean interspike interval) as a function of V_{PSV} (see Supplementary Fig. 3 for distribution of interspike interval). Clearly, the firing rate increases monotonically with increasing stimulus intensity, indicating that our biomimetic encoder is capable of rate-based encoding. Finally, Fig. 3e shows the encoding energy per spike (E_{en}) for rate-based encoding, computed based on Eq. (1). The monotonic increase in E_{en} with increasing V_{PSV} is consistent with increasing firing rate, i.e., more spiking in the postsynaptic neuron.

$$E_{en} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{2}C_{TG}{V_{TG,i}}^2 + V_{PSC,i}V_{DS}t_p\right), \quad (1)$$

Typical energy consumption is around 1–5 pJ/spike. Note that the second term in Eq. (1) dominates in our demonstration since the first term contributes ≈100 fJ. Therefore, one obvious way to reduce the power consumption is through V_{DS} scaling. Note that the neural encoder exploits subthreshold device characteristics and does not impose any requirement on the current device. Hence it is possible to operate the neural encoder with ultra-low V_{DS}. Another alternative to reduce the power dissipation is to increase the sampling rate i.e., reduce t_p. However, beyond a certain point, the first term will start to dominate, which can be scaled by scaling the oxide thickness to achieve encoding at scaled V_{TG} values. Note that oxide thickness scaling increases C_{TG}, but the square term involving V_{TG} will determine the energy scaling.

Another encoding principle found in sensory neurobiology is spike count-based encoding. For example, rats show remarkable texture discriminations using their facial whiskers. It is found that the trigeminal ganglion cells that innerte the sensory receptor from each whisker use spike count to distinguish the stimuli[33]. Similar spike count encoding is observed for frequency discrimination of vibrotactile stimuli in the primary somatosensory cortex of trained monkeys[34]. Figure 3f shows the V_{TG} pulse profile used to achieve spike count-based encoding using our biomimetic encoder. In this case, the magnitude of V_{TG} pulses increases with added zero mean Gaussian noise of standard deviation $\sigma_{TG} = 0.2$ V. Each trial consists of 32 pulses, and 16 trials were recorded for each V_{PSV}. The corresponding responses of the neural encoder are displayed in Fig. 3g. Figure 3h shows the encoding transfer function i.e., the mean spike count as a function of V_{PSV} (see Supplementary Fig. 4 for total spike counts for all 16 trials for different σ_{TG}). Clearly, the mean spike count increases monotonically with increasing stimulus intensity, indicating that



our biomimetic encoder is capable of spike count-based encoding. Note that, the implementation of spike count-based encoding does not necessarily require stochasticity, i.e., similar results could be obtained using $\sigma_{TG} = 0$ V. However, in the context of SNN, stochasticity can aid as hardware realization of integrate and fire (IF) neuron can be challenging. A more realistic neuron is leaky integrate and fire (LIF) neuron, where random spiking can compensate for the loss in information due to capacitive discharging between spikes. Figure 3i shows E_{en} for spike count-based encoding, which shows a monotonic increase with V_{PSV} since the spike count increases accordingly. The energy expenditure was found to be ≈1–3.5 pJ/spike.

Fig. 3 Hardware realization of neural encoding algorithms. a Analog V_{PSV} values obtained for different LED illuminations. **b** Spike rate-based encoding using V_{TG} pulses that are randomly sampled from a Gaussian distribution with mean, $\mu_{TG} = -2.5$ V, and standard deviation $\sigma_{TG} = 0.8$ V. **c** Corresponding I_{PSC} for each V_{PSV} in **a**. During each trial, V_{TG} pulses were sampled for $N = 32$ number of times and a total of 16 trials were conducted resulting in 512 sampling points for each V_{PSV}. See Supplementary Movie 1 for real-time encoding of different LED intensities into stochastic spike trains. **d** Mean firing rate (frequency) as a function of V_{PSV}. **e** Encoding energy (E_{en}) for rate-based encoding. **f** Spike count-based encoding. Here, the magnitude of V_{TG} pulses increase with added zero mean Gaussian noise of standard deviation $\sigma_{TG} = 0.2$ V. Each trial consists of 32 pulses and 16 trials are recorded for each V_{PSV}. **g** Corresponding responses of the neural encoder. **h** Mean spike count as a function of V_{PSV}. **i** E_{en} for spike count-based encoding. **j** Spike timing-based encoding. Here, the magnitude of V_{TG} pulses decrease over time with added zero mean Gaussian noise of standard deviation $\sigma_{TG} = 0.2$ V. Each trial consists of 32 pulses for each V_{PSV}. **k** Corresponding responses of the neural encoder. A sense amplifier is used to sense the arrival of the first spike triggering the deactivation of the V_{TG} pulse sampling for the rest of the trial. **l** Mean spike-timing as a function of V_{PSV}. **m** E_{en} for spike timing-based encoding. Error bars in **d, e, h, i, l, m** represent the variation (standard deviation) across trials. **n** Original, **o** spike-rate encoded, **p** spike-count encoded, and **q** spike-timing encoded Cameraman image. The pixel values ranging from 0 to 255 were mapped linearly to the V_{PSV} range of 0 to 5 V. Supplementary movie files 2 show the time evolution of encoding of the Cameraman images for rate-based, count-based, and timing-based encoding.

Whereas, rate-based encoding and spike count-based encoding are the most broadly accepted view of neural computation, these approaches ignore the information possibly contained in the exact timing of the spikes. In fact, recent studies suggest that a straightforward firing rate or spike count-based encoding may be too simplistic to describe brain activity in its entirety. For example, neurophysiological experiments show that visual neurons in rhesus monkeys can recognize faces within ≈80–160 ms[35]. Anatomically, it involves more than ten synaptic stages between the photoreceptors of the retina and visually responsive neurons in the temporal cortex implying that each layer has, on average, only 10 ms of processing time. Since the firing rates of cortical neurons are in the range 0–100 spikes per second, a neuron in any given layer can only generate one spike before neurons in the next layer have to respond. This puts severe constraints on the way information is encoded in visual pathways. Firing-rate or spike count-based encoding seems inappropriate and evidence suggests that analog information is encoded by the relative arrival times of spikes[36–39]. Such an encoding scheme also referred to as the spike timing-based encoding, not only allows very rapid information processing but also offers tremendous energy benefits for future SNNs. Figure 3j shows the V_{TG} pulse profile used for achieving spike timing-based encoding using our biomimetic encoder. In this case, the magnitude of V_{TG} pulses decreases over time with added zero-mean Gaussian noise of standard deviation $\sigma_{TG} = 0.2$ V. Each trial consists of 32 pulses for each V_{PSV}. The corresponding responses of the neural encoder are displayed in Fig. 3k. A sense amplifier is used to sense the arrival of the first spike triggering the deactivation of the V_{TG} pulse sampling for the rest of the trial. Figure 3l shows the encoding transfer function i.e., mean spike-timing as a function of V_{PSV} (see Supplementary Fig. 5 for distribution of spike timing over multiple trials). Clearly, high-intensity stimuli invoke early spiking and vice versa indicating that our biomimetic encoder is capable of spike timing-based encoding as well. Note that, the implementation of spike timing-based encoding does not require stochasticity, i.e., similar results could be obtained by using $\sigma_{TG} = 0$ V. However, the flexibility of noise adjustment makes our neural encoder more bio-realistic. Figure 3m shows E_{en} for spike timing-based encoding. Unlike rate-based and count-based encoding, timing-based encoding shows a monotonic decrease with increasing V_{PSV}. This is owing to the fact the spiking occurs earlier for higher V_{PSV} deactivating the encoder and minimizing the energy consumption per spike. The fact that the encoding energy can be significantly lower for timing-based encoding compared to rate-based or count-based encoding is appealing for ultra-low-power neuromorphic computing using SNN (See Supplementary Note 2 showing the comparison of our neural encoder with other types of spike encoders).

Finally, Fig. 3n–q, respectively, shows the original Cameraman image and the corresponding spike rate, spike count, and spike timing-based encoding. The pixel values ranging from 0 to 255 were mapped linearly to the V_{PSV} range of 0–5 V (see the "Methods" section for details on image encoding). Clearly, the Cameraman image is accurately encoded, irrespective of the encoding algorithm. Note that the contrast of the image in Fig. 3q is reversed compared to the original image, which is expected for spike-time based encoding since the higher pixel values should spike earlier than the lower pixel values. Supplementary movie files 2 show the time evolution of encoded images over time for rate-based, count-based, and timing-based encoding. Supplementary Fig. 6 shows the time evolution of the correlation coefficient (CC) between the original image and the encoded image. The CC reaches ≈1 at the end of encoding for all three encoding algorithms.

Dynamic range and encoding precision for rate-based encoding. Now, we focus on two key aspects of neural encoding, namely, dynamic range and encoding precision. A high dynamic range (HDR) allows neurons to respond to more extreme stimuli. For example, photoreceptors in human eyes can identify objects in starlight as well as in bright sunlight despite of illumination levels differing by ≈9 orders of magnitude, i.e., over a dynamic range of 90 dB[40]. Similarly, the dynamic range of human hearing is roughly 140 dB[41]. However, HDR does not necessarily guarantee high precision (HP). For example, a whisper cannot be heard in loud surroundings. Similarly, eyes take time to adapt to different illumination levels. In fact, most sensory neurons adjust their spike encoding based on the environment[42–44]. Figure 4a–f shows how our biomimetic encoder achieves similar functionality by adjusting σ_{TG} and μ_{TG} of the Gaussian distribution used for sampling V_{TG} as well as and the thresholding current, I_{ST} for the spike rate-based encoding algorithm presented earlier. For numerical simulations, we have used the virtual source (VS) model described elsewhere[22,23]. Clearly, HDR can be achieved by using higher values of σ_{TG}, whereas smaller values of σ_{TG} allow HP (Fig. 4a, b). This is because the encoding transfer function follows the cumulative probability distribution of a random Gaussian variable since for a given V_{PSV} stimulus, there will always be a postsynaptic spike if the magnitude of the V_{TG} pulse is more positive than the one corresponding to the spiking threshold, V_{ST}, as shown in Fig. 2c. For a higher value of σ_{TG}, the cumulative probability distribution follows a linear trend allowing a larger V_{PSV} range to be encoded, whereas a lower value of σ_{TG} results in a non-linear cumulative probability distribution that restricts the encoding range but improves the encoding precision. However, both feats cannot be achieved at the same time by adjusting σ_{TG}. However, by adjusting μ_{TG} it is possible to achieve HP for different ranges of stimulus intensity similar to the

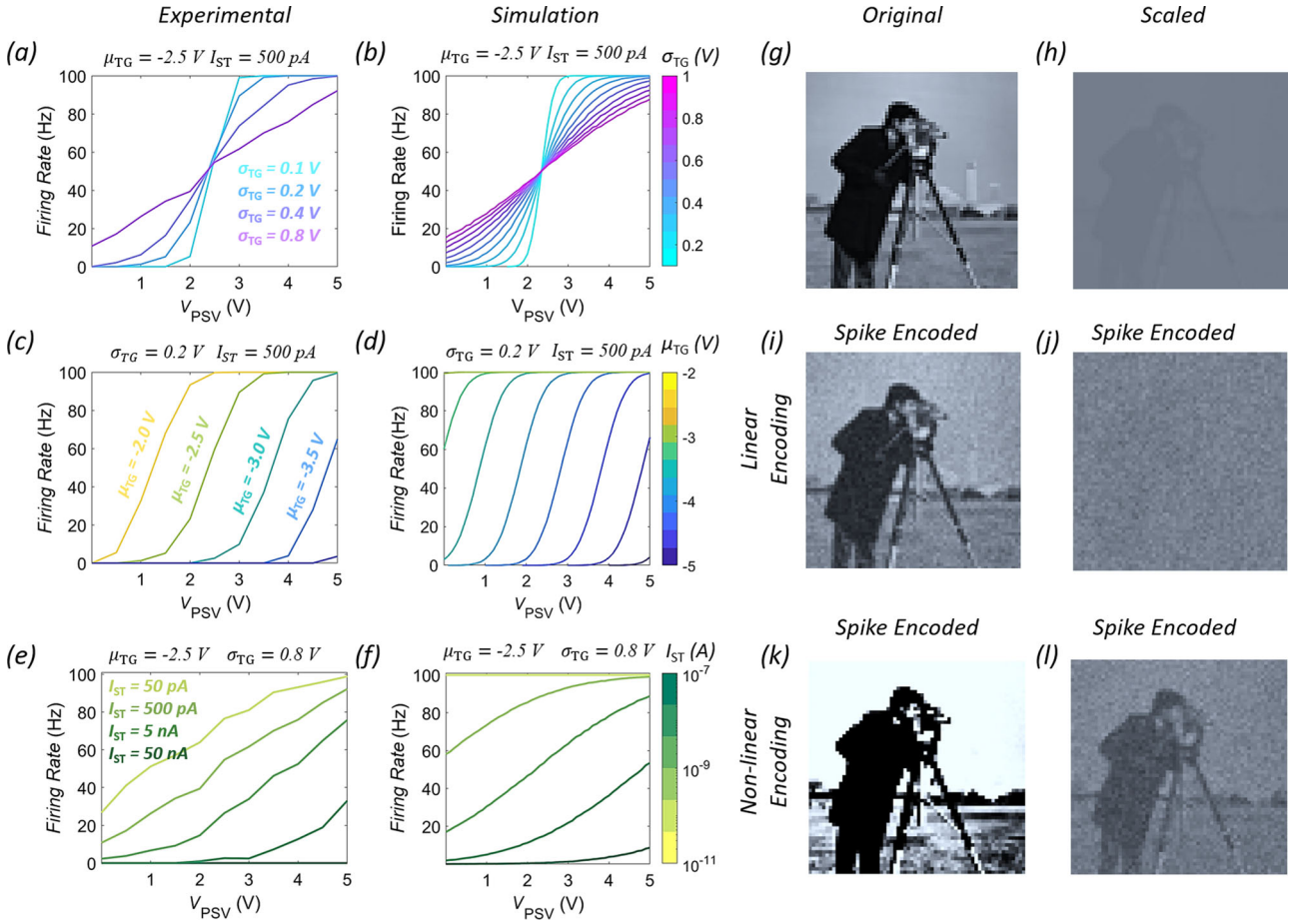


Fig. 4 Dynamic range and encoding precision for rate-based encoding. a Experimental and **b** virtual source (VS) model-based simulation of firing rate as a function of V_{PSV}, for different σ_{TG}. of the Gaussian distribution used for sampling V_{TG}. A high dynamic range (HDR) can be achieved by using higher values of σ_{TG}, whereas smalle. r values of σ_{TG}. allow high precision (HP). However, both feats cannot be achieved at the same time by adjusting σ_{TG}. **c** Experimental and **d** VS model-based simulation of firing rate as a function of, V_{PSV}, for different mean (μ_{TG}) of the Gaussian distribution used for sampling V_{TG}. By adjusting μ_{TG} it possible to achieve HP for different ranges of stimulus intensity like the sensory neurons. **e** Experimental and **f** VS model-based simulation also show that the firing rate can be tuned by adjusting I_{ST}. Lower values of I_{ST} allow more spiking events for any given V_{PSV}, whereas, higher values of I_{ST} restrict. s spiking even for higher V_{PSV}. **g** Original and **h** scaled Cameraman images. Corresponding spike rate-based **i, j** linear and **k**, **l** non-linear encoding using our biomimetic encoder. The original Cameraman image necessitates linear encoding since the pixel values have large dynamic range, whereas, the scaled Cameraman image is better encoded using high precision non-linear encoding.

sensory neurons (Fig. 4c, d). The spiking rate can also be tuned by adjusting the I_{ST} (Fig. 4e, f). Lower values of I_{ST} allow more spiking events for any given V_{PSV}, whereas, higher values of I_{ST} restrict spiking even for higher V_{PSV}. Figure 4g and h show the original and scaled Cameraman images, and Fig. 4i–l, respectively, show the corresponding spike rate-based linear and non-linear encoding using our biomimetic encoder. The original Cameraman image necessitates linear encoding since the pixel values have a large dynamic range, whereas, the scaled Cameraman image is better encoded using high precision non-linear encoding.

MNIST digit classification using our neural encoder device. Finally, we exploit our biomimetic encoder for encoding MNIST data set on digit-classification into spike trains and infer using a trained SNN. For training the SNN we have used an approach described by Sengupta et al.[45]. This approach overcomes the lower accuracy of unsupervised learning rules such as the spike-time dependent plasticity (STDP) used for training SNNs[46–50]. The lower accuracy is due to the lack of efficient algorithms to make use of the spiking neurons. To bridge this gap, ANN-SNN conversion schemes are used, where an ANN is trained using the traditional back-propagation algorithm, followed by the conversion of the ANN to SNN[45,51,52]. This approach yields higher inference accuracy owing to near-lossless ANN-SNN conversion[45]. Here, we train a fully connected two-layered artificial neural network with 100 neurons in the hidden layer and 10 neurons in the output layer for digit-classification using the MNIST dataset as shown in architecture in Fig. 5a. MNIST dataset with a size of 28×28 pixels is flattened to obtain 784 pixels, which is fed to the input layer. The ten output neurons correspond to digits from 0 to 9. During training, for every input image, the network is trained through gradient descent to ensure that the output matches the expected label. Here, the ANN is trained with a learning rate of 0.0001 to ensure high convergence accuracy. Further, the following restrictions are incorporated while training the ANN to allow smooth ANN-SNN transition: rectified linear unit (ReLU) is used as the activation function due to its functional equivalence to IF spiking neuron used in SNN, bias terms are eliminated to ensure a smaller parameter space which enables easier ANN-SNN conversion, and no regularization is used. Sixty thousand images from the MNIST data set were used to train the ANN to achieve a training accuracy of 91.5%

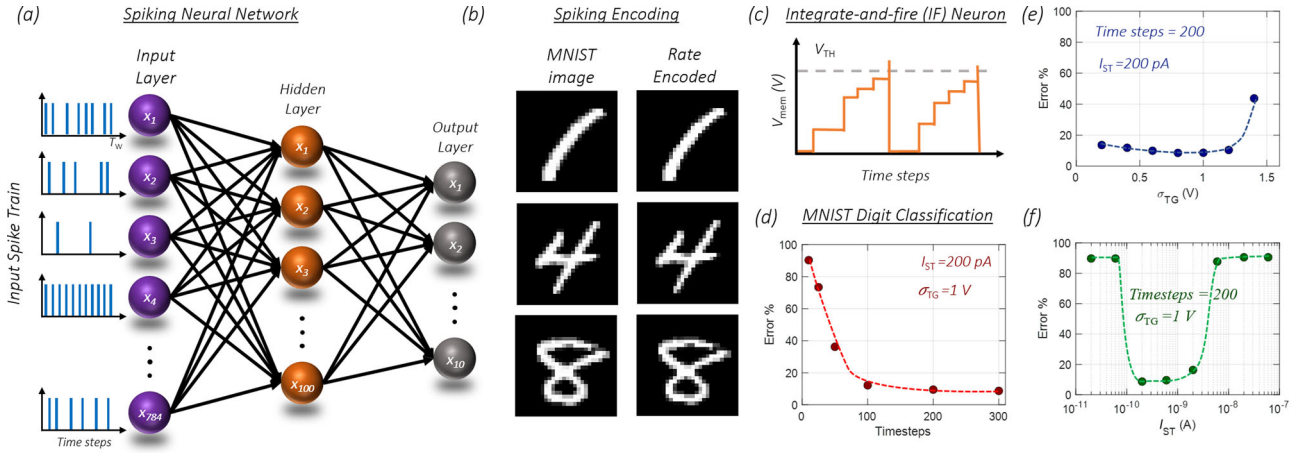


Fig. 5 Encoding of MNIST data for digit classification using SNN. a A fully connected two-layered trained ANN with 100 neurons in the hidden layer and 10 neurons in the output layer for MNIST digit-classification is transformed into SNN. **b** Examples of spike encoded digits. First, the pixel intensity values ranging from 0 to 255 are mapped on to V_{PSV} values in the range of 0–5 V. Next, we recorded I_{PSC} for each V_{PSV} by applying stochastic V_{TG} pulses with I_{ST} = 200 pA to generate binary spike trains (X) in time. We adopt rate-based encoding by applying V_{TG} pulses with the pulse magnitude determined using a random Gaussian distribution with $\mu_{TG} = -5.5$ V, and $\sigma_{TG} = 1$ V. **c** Characteristics of IF neuron which is substituted for ReLU for ANN-SNN conversion. In the IF neuron when the membrane potential crosses a certain threshold (V_{TH}), the neuron spikes, propagating spike to the next layer, and it resets back to its resting potential which is set as zero. **d** Inference error versus the number of timesteps for classifying a set of 10,000 test images. Increasing the number of timesteps is important to allow sufficient firing, to effectively encode the pixel intensities. But even with 200 timesteps, we achieve a low error of 9.5%. Dependence of inference accuracy on **e** σ_{TG}, and **f** I_{ST}. The minimum error of ≈8.6% is achieved at $\sigma_{TG} = 0.8$ V. For lower σ_{TG}, the dynamic range is insufficient to capture the variation in pixel intensities, whereas, for very high σ_{TG}, there is insufficient difference between the firing rates corresponding to different pixel intensities resulting in larger errors. Similarly, the minimum error of 8.6% is obtained for $I_{ST} = 200$ pA. For higher I_{ST}, the spiking is minimal resulting in an inadequate representation of image pixels, whereas, for lower I_{ST}, any pixel intensity results in excessive firing.

over 100 epochs. Following this, a testing accuracy of 92.7% was achieved using the remaining 10,000 images.

As discussed, SNNs use binary spikes in time which are representative of the action potential in BNNs. This requires the conversion of the analog image pixel intensities to digital spike trains. To accomplish the analog to spike conversion using our biomimetic encoder, first, the pixel intensity values ranging from 0 to 255 are mapped onto V_{PSV} range of 0–5 V. Next, we record I_{PSC}, for V_{PSV} values corresponding to each pixel over a time window, T_W, by applying stochastic V_{TG} pulses. Each I_{PSC} value subsequently undergoes a thresholding function with I_{ST} to generate binary X in time. We adopt rate-based encoding by applying V_{TG} pulses with the pulse magnitude determined using a random Gaussian distribution, as described earlier. Figure 5b shows examples of spike encoded digit using $\mu_{TG} = -5.5$ V, and $\sigma_{TG} = 1$ for the Gaussian distribution for the V_{TG} pulses, and $I_{ST} = 200$ pA. The resultant X is fed into the SNN, as shown in the input layer of Fig. 5a. For ANN-SNN conversion, ReLU activation functions are replaced by IF neuron as shown in Fig. 5c following Eq. (2).

$$V_{mean}(t+1) = V_{mean}(t) + \sum w \cdot X(t). \quad (2)$$

Here, the IF neuron is represented as the function of timesteps (t). $V_{mean}(t)$ is the membrane potential, and w denotes the weights obtained from the trained ANN. In the IF neuron when the membrane potential crosses a certain threshold (V_{th}), the neuron spikes, propagating spike to the next layer, and it resets back to its resting potential which is set as zero. To optimize the IF neuron threshold, threshold-balancing is used to set the threshold as the maximum neuron activation for the corresponding layer obtained by the dot product of the weights and spike-train at an instance t[45]. The SNN is used to classify the set of 10,000 test images. Figure 5d shows the inference error versus the number of timesteps. Increasing the number of timesteps is important to allow sufficient firing, to effectively encode the pixel intensities. But remarkably, even with 200 timesteps, we achieve a low error of 9.5%. This is further improved as the timesteps are increased with the minimum error of 8.6% at 300 timesteps. Hence a maximum accuracy of 91.4% is achieved when the SNN is simulated with our biomimetic neural encoder. Additionally, similar test accuracies are obtained from both ANN and SNN, indicating a successful ANN-SNN transformation with a minimal loss of 1.3%.

Finally, we explore the dependence of inference accuracy on the dynamic range and the firing rate, parameters that can be adjusted in our biomimetic spike encoder by adjusting σ_{TG} and I_{ST}, respectively. As shown in Fig. 5e, minimum error of ≈8.6% is achieved at σ_{TG} of 0.8 and 1, with higher errors for lower and higher σ_{TG}. As described earlier in Fig. 4a, b, for lower σ_{TG}, the dynamic range is low to capture the variation in pixel intensities, whereas for very high σ_{TG}, there is insufficient difference between t firing rates corresponding to different pixel intensities resulting in a large error. A similar non-monotonic trend is seen in inference error with respect to I_{ST} in Fig. 5f, with the minimum error of 8.6% obtained at 200 pA. For higher I_{ST}, the spiking is minimal resulting in an inadequate representation of image pixels, while for lower I_{ST}, any pixel intensity results in excessive firing as seen in Fig. 4e, f. Nevertheless, by optimizing these parameters it is possible to ensure efficient encoding of the MNIST images, and thereby achieve a maximum accuracy of 91.4%.

Discussion

In conclusion, we have developed a neural encoder based on a dual gated MoS_2 FET with a stochastic sampling terminal capable of encoding analog signals into spike trains. We also implemented three encoding algorithms, namely, spike rate-based encoding, spike count-based encoding, and spike timing-based encoding found in sensory neurobiology. As a prototype demonstration, we

show the direct conversion of analog light intensities to corresponding rate-based spike trains analogous to phototransduction mechanisms in visual pathways. We also show frugal encoding energy expenditure in the range of few pico Joules per spike. Our biomimetic encoder also allows flexibility in terms of adjusting the encoding range and encoding precision, two key features found in biological sensory transduction to enable seamless adaption to different environmental conditions. Finally, we encoded the MNIST data set for digit classification using our spike encoder and achieved an inference accuracy of 91.4% by using a trained SNN. In brief, our demonstration of the biomimetic neural encoder is a leap forward towards achieving energy-efficient and bio-realistic neuromorphic hardware.

Methods

Device fabrication. The dual-gated devices were fabricated using micro-mechanically exfoliated MoS_2 flakes on 285 nm thermally grown SiO_2 substrates with highly doped-Si as the back-gate electrode. The source/drain contacts were defined using electron-beam lithography (Vistec EBPG5200). Ni (40 nm) followed by Au (30 nm) were deposited using electron-beam evaporation for the contacts. For fabricating the top-gate, hydrogen silsesquioxane (HSQ) was used as the dielectric. It was deposited by spin coating 6% HSQ in methyl isobutyl ketone (MIBK) (Dow Corning XR-1541-006) at 4000 rpm for 45 s and baked at 80 °C for 4 min. The HSQ was patterned using an e-beam dose of 2000 μC/cm^2 and was developed at room temperature using 25% tetramethylammonium hydroxide (TMAH) for 30 s following a 90 s rinse in deionized water (DI). Next, it was cured in the air at 180 °C and then 250 °C for 2 min and 3 min, respectively. The top-gate electrode was patterned using electron-beam lithography followed by the deposition of Ni/Au using electron-beam evaporation as the contact.

Device measurements. Electrical characterization was performed at room temperature in high vacuum (≈10^{-6} Torr) on a Lake Shore CRX-VF probe station and using a Keysight B1500A parameter analyzer. We observed none to minimal hysteresis in the device characteristics for both top-gate and back-gate sweeps indicating high quality of MoS_2/SiO_2 and MoS_2/HSQ interfaces (See Supplementary Fig. 7). For current sampling, when a sampling delay is set, (for example, $T =$ 10 ms) the tool determines a time width (W) for integration based on the current value i.e., lower current values need larger W and vice versa. If $W < T$ the tool measures the current at any point within the delay period. Else if $W > T$, irrespective of the time delay, the tool measures the current every time width (W) it sets for the current value range. The 10 ms delay we chose is large enough for the current value ranges we are operating to integrate. So, the tool will measure at any point within the 10 ms delay.

Image encoding. Note that the gray scale pixel values in an 8-bit cameraman image range from 0–255, which are mapped to V_{PSV} range of 0–5 V. This would require a V_{PSV} precision of $5/255 = 0.0196 \approx 0.02$ V. However, our experimental V_{PSV} step size was 0.5 V. Therefore, we quantized the cameraman image into 11 distinct levels and correlate those levels to $V_{PSV} = 0$, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, and 5.0 V. Note that for spike-rate based encoding, our maximum frequency is 100 Hz and the average standard deviation for the encoded frequencies i.e., the encoding error bar is ≈10 Hz, allowing distinct encoding of 11 levels. One way to improve the mapping precision for spike-rate based encoding is to increase the maximum encoding frequency to 1000 Hz through faster sampling if the encoding error bar remains unaltered. Similarly, for spike-count based encoding, the maximum count is 16 and the average standard deviation for the encoded count or the encoding error bar is ≈1.5, again allowing the distinct encoding of 11 levels. Finally, for spike-timing based encoding, the maximum number of time-steps is 30 and the average standard deviation for the encoded time-step or the encoding error bar is ≈4, allowing the distinct encoding of eight levels. The mapping precision for spike-count and spike-timing based encoding can be increased by increasing the total number of sampling points.

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.

Received: 9 September 2020; Accepted: 9 March 2021;
Published online: 09 April 2021

References

1. Rudnicki, M., Schoppe, O., Isik, M., Völk, F. & Hemmert, W. Modeling auditory coding: from sound to spikes. *Cell Tissue Res.* **361**, 159–175 (2015).
2. Reiter, S., Rodriguez, C. C., Sun, K. & Stopfer, M. Spatiotemporal coding of individual chemicals by the gustatory system. *J. Neurosci.* **35**, 12309–12321 (2015).
3. Hallock, R. M. & Di Lorenzo, P. M. Temporal coding in the gustatory system. *Neurosci. Biobehav. Rev.* **30**, 1145–1160 (2006).
4. Meister, M., Lagnado, L. & Baylor, D. A. Concerted signaling by retinal ganglion cells. *Science* **270**, 1207–1210 (1995).
5. Choi, S.-Y. et al. Encoding light intensity by the cone photoreceptor synapse. *Neuron* **48**, 555–562 (2005).
6. McDonnell, M. D. & Ward, L. M. The benefits of noise in neural systems: bridging theory and experiment. *Nat. Rev. Neurosci.* **12**, 415–425 (2011).
7. Jaramillo, F. & Wiesenfeld, K. Mechanoelectrical transduction assisted by Brownian motion: a role for noise in the auditory system. *Nat. Neurosci.* **1**, 384–388 (1998).
8. Mead, C. Neuromorphic electronic systems. *Proc. IEEE* **78**, 1629–1636 (1990).
9. Douglas, R., Mahowald, M. & Mead, C. Neuromorphic analogue VLSI. *Annu. Rev. Neurosci.* **18**, 255–281 (1995).
10. Furber, S. Large-scale neuromorphic computing systems. *J. Neural Eng.* **13**, 051001 (2016).
11. Boybat, I. et al. Neuromorphic computing with multi-memristive synapses. *Nat. Commun.* **9**, 1–12 (2018).
12. Sangwan, V. K. & Hersam, M. C. Neuromorphic nanoelectronic materials. *Nat. Nanotechnol.* **15**, 1–12 (2020).
13. Hassoun, M. H. *Fundamentals of Artificial Neural Networks* (MIT Press, 1995).
14. Schmidhuber, J. Deep learning in neural networks: an overview. *Neural Netw.* **61**, 85–117 (2015).
15. Ghosh-Dastidar, S. & Adeli, H. Spiking neural networks. *Int. J. Neural Syst.* **19**, 295–308 (2009).
16. Ponulak, F. & Kasinski, A. Introduction to spiking neural networks: Information processing, learning and applications. *Acta Neurobiol. Exp.* **71**, 409–433 (2011).
17. Han, B., Sengupta, A. & Roy, K. On the energy benefits of spiking deep neural networks: a case study. *International Joint Conference on Neural Networks (IJCNN)*, 971–976 Vancouver, Canada (2016).
18. Butler, S. Z. et al. Progress, challenges, and opportunities in two-dimensional materials beyond graphene. *ACS Nano* **7**, 2898–2926 (2013).
19. Bhimanapati, G. R. et al. Recent advances in two-dimensional materials beyond graphene. *ACS Nano* **9**, 11509–11539 (2015).
20. Das, S., Robinson, J. A., Dubey, M., Terrones, H. & Terrones, M. Beyond graphene: progress in novel two-dimensional materials and van der Waals solids. *Annu. Rev. Mater. Res.* **45**, 1–27 (2015).
21. Nasr, J. R. & Das, S. Seamless fabrication and threshold engineering in monolayer MoS2 dual-gated transistors via hydrogen silsesquioxane. *Adv. Electron. Mater.* **5**, 1800888 (2019).
22. Das, S., Dodda, A. & Das, S. A biomimetic 2D transistor for audiomorphic computing. *Nat. Commun.* **10**, 3450 (2019).
23. Sebastian, A., Pannone, A., Radhakrishnan, S. S. & Das, S. Gaussian synapses for probabilistic neural networks. *Nat. Commun.* **10**, 1–11 (2019).
24. Dodda, A. et al. Stochastic resonance in MoS2 photodetector. *Nat. Commun.* **11**, 4406 (2020).
25. Schranghamer, T. F., Oberoi, A. & Das, S. Graphene memristive synapses for high precision neuromorphic computing. *Nat. Commun.* **11**, 5474 (2020).
26. Yin, Z. et al. Single-layer MoS2 phototransistors. *ACS Nano* **6**, 74–80 (2012).
27. Wang, L. et al. Functionalized MoS2 nanosheet-based field-effect biosensor for label-free sensitive detection of cancer marker proteins in solution. *Small* **10**, 1101–1105 (2014).
28. Park, M. et al. MoS2-based tactile sensor for electronic skin applications. *Adv. Mater.* **28**, 2556–2562 (2016).
29. Arnold, A. J., Shi, T., Jovanovic, I. & Das, S. Extraordinary radiation hardness of atomically thin MoS2. *ACS Appl. Mater. Interfaces* **11**, 8391–8399 (2019).
30. Das, S., Chen, H. Y., Penumatcha, A. V. & Appenzeller, J. High performance multilayer MoS2 transistors with scandium contacts. *Nano Lett.* **13**, 100–105 (2013).
31. Schulman, D. S., Arnold, A. J. & Das, S. Contact engineering for 2D materials and devices. *Chem. Soc. Rev.* **47**, 3037–3058 (2018).
32. Adrian, E. D. & Zotterman, Y. The impulses produced by sensory nerve-endings: Part II. The response of a single end-organ. *J. Physiol.* **61**, 151 (1926).
33. Arabzadeh, E., Panzeri, S. & Diamond, M. E. Deciphering the spike train of a sensory neuron: counts and temporal patterns in the rat whisker pathway. *J. Neurosci.* **26**, 9216–9226 (2006).
34. Luna, R., Hernández, A., Brody, C. D. & Romo, R. Neural codes for perceptual discrimination in primary somatosensory cortex. *Nat. Neurosci.* **8**, 1210–1219 (2005).

35. Perrett, D. I., Rolls, E. T. & Caan, W. Visual neurones responsive to faces in the monkey temporal cortex. *Exp. Brain Res.* **47**, 329–342 (1982).
36. Thorpe, S. J. Spike arrival times: A highly efficient coding scheme for neural networks. *Parallel Process. Neural Syst.* 91–94 (1990).
37. Butts, D. A. et al. Temporal precision in the neural code and the timescales of natural vision. *Nature* **449**, 92–95 (2007).
38. Chase, S. M. & Young, E. D. Spike-timing codes enhance the representation of multiple simultaneous sound-localization cues in the inferior colliculus. *J. Neurosci.* **26**, 3889–3898 (2006).
39. Reich, D. S., Mechler, F. & Victor, J. D. Temporal coding of contrast in primary visual cortex: when, what, and why. *J. Neurophysiol.* **85**, 1039–1050 (2001).
40. Zaghloul, K. A., Boahen, K. & Demb, J. B. Contrast adaptation in subthreshold and spiking responses of mammalian Y-type retinal ganglion cells. *J. Neurosci.* **25**, 860–868 (2005).
41. Zeng, F.-G., Fu, Q.-J. & Morse, R. Human hearing enhanced by noise. *Brain Res.* **869**, 251–255 (2000).
42. Wen, B., Wang, G. I., Dean, I. & Delgutte, B. Dynamic range adaptation to sound level statistics in the auditory nerve. *J. Neurosci.* **29**, 13797–13808 (2009).
43. Viana, R. et al. Dynamic range in a neuron network with electrical and chemical synapses. *Commun. Nonlinear Sci. Numer. Simul.* **19**, 164–172 (2014).
44. Wachowiak, M. & Cohen, L. B. Representation of odorants by receptor neuron input to the mouse olfactory bulb. *Neuron* **32**, 723–735 (2001).
45. Sengupta, A., Ye, Y., Wang, R., Liu, C. & Roy, K. Going deeper in spiking neural networks: VGG and residual architectures. *Front. Neurosci.* **13**, 95 (2019).
46. Diehl, P. U. & Cook, M. Unsupervised learning of digit recognition using spike-timing-dependent plasticity. *Front. Comput. Neurosci.* **9**, 99 (2015).
47. Zhang, W. & Li, P. Spike-train level backpropagation for training deep recurrent spiking neural networks. https://arxiv.org/1908.06378 (2019).
48. Brader, J. M., Senn, W. & Fusi, S. Learning real-world stimuli in a neural network with spike-driven synaptic dynamics. *Neural Comput.* **19**, 2881–2912 (2007).
49. Querlioz, D., Bichler, O., Dollfus, P. & Gamrat, C. Immunity to device variations in a spiking neural network with memristive nanodevices. *IEEE Trans. Nanotechnol.* **12**, 288–295 (2013).
50. Pfeiffer, M. & Pfeil, T. Deep learning with spiking neurons: opportunities and challenges. *Front. Neurosci.* **12**, 774 (2018).
51. Diehl, P. U. et al. Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing. *International Joint Conference on Neural Networks (IJCNN)*, 1–8 Killarney, Ireland (2015).
52. Cao, Y., Chen, Y. & Khosla, D. Spiking deep convolutional neural networks for energy-efficient object recognition. *Int. J. Comput. Vis.* **113**, 54–66 (2014).

Acknowledgements

The work was supported by Army Research Office (ARO) through Contract Number W911NF1920338. Figure 1a was designed using resources from Freepik.com "image: Freepik.com".

Author contributions

Saptarshi Das and S.S. conceived the idea and designed the experiments. Saptarshi Das wrote the paper. Sarbashis Das fabricated the devices. A.S. performed SNN simulations. Saptarshi Das, S.S., A.S., A.O., and Sarbashis Das analyzed the data, discussed the results, agreed on their implications, and contributed to the preparation of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-22332-8.

Correspondence and requests for materials should be addressed to S.D.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021