Artificial Working Memory Constructed by Planar 2D Channel Memristors Enabling Brain-Inspired Hierarchical Memory Systems

Xinglong Ji, Song Hao, Kian Guan Lim, Shuai Zhong, and Rong Zhao*

In human hierarchical memory structures, working memory functions as a temporal interface to integrate multimodal information from sensory memory and long-term memory, underpinning cognition development. Therefore, an artificial working memory is highly anticipated to endow computing systems with advanced capabilities. Herein, an artificial working memory that can synthesize and process multimodal information to enable a brain-inspired hierarchical memory system is developed. To emulate the characteristic synaptic behavior of working memory, a planar 2D channel memristor (PTCM) with stable short-term frequency-dependent plasticity is designed. An array of PTCMs on a monolayer MoS2 grain are assembled and the information encoding, forgetting, retrieval, and updating functions of working memory are experimentally demonstrated. A hierarchical memory system through simulation, which integrates working memory network, sensory memory, and long-term memory, is built, emulating the information synthesizing and processing procedures of the human brain. The proposed system provides a promising way to enhance the cognitive capability of intelligence machines.

1. Introduction

Memory is a vital part in the development of human learning, cognition, perception, and other behaviors. Besides carrying information, it also plays a crucial role in biological neurocomputing and multimodal information synthesizing/processing.[1-2] Proposed by Atkinson-Shiffrin’s multistore model, the hierarchical structure of human memory is the biological substrate for these complex functions.[3] It functionally disassembles human memory into three hierarchical levels across multiple time spans and different cerebral cortex regions, including sensory memory, short-term memory, and long-term memory (Figure 1a). As an important subdivision of short-term memory, working memory can efficiently integrate information from other memory hierarchies, which provides a multimodal information interface to acquire and filter key information from redundant environmental information flow.[4,5] It is composed of four main components, including an attention control system that drives the whole system and deals with cognitive tasks such as mental arithmetic and problem-solving, the central executive control, and three subsidiary storage systems, namely, the phonological loop that deals with spoken and written materials, the visuospatial sketchpad that stores and processes information in a visual or spatial form, and episodic buffer that acts as a “backup” store to communicate with both long-term memory and the subcomponents of working memory (Figure 1b).[6] The four parts are responsible for different forms of multimodal information storage and processing, especially visual information and auditory information. They work together as a functional working memory to perform a wide range of coding and manipulation activities that are essential for generating advanced cognitive behaviors, such as planning, reasoning, and decision-making.

Inspired by the hierarchical human memory structure, solid-state memories are expected to undertake more computing and information processing functions rather than mere data storing. An artificial working memory is anticipated to endow intelligence machines with enhanced cognitive learning capability and therefore, make up the lack of computing capability of existing memory architecture.[7,8] Such artificial working memory needs to assemble various complex functions, especially the multimodal information synthesizing and processing in a single memory device, which is challenging and yet to be explored.
In this work, we exploit an artificial working memory exhibiting multimodal information synthesizing and processing capabilities to enable a brain-inspired hierarchical memory system. According to the proposed synaptic theory, the main functions of working memory originate from billions of synapses with stable short-term memory characteristic and excellent frequency-dependent plasticity.[9] In our design, such a specific synapse is emulated by a planar memristor with a cation reservoir and monolayer 2D channel design (planar 2D channel memristor, PTCM). The cation reservoir can finely control Ag⁺ ions' release and recycle under e-field, therefore, mimicking the Ca²⁺-controlled neurotransmitter migration and diffusion dynamics in biological neuron cells. The monolayer 2D channel limits the diffusion degree of freedom to make the cation diffusion highly controllable. On this basis, we construct a functional working memory with an array of PTCM on a monolayer 2D material grain (MoS₂) and experimentally demonstrate various basic functions, including information encoding (acquisition and consolidation), forgetting, retrieval, and updating. We further conducted a simulation to integrate our working memory in a brain-inspired hierarchical memory system and demonstrated complex information processing and encoding for multimodal inputs. Our working memory network receives multimodal inputs (digit, image, and voice) from sensory memory and processes it based on the existing “knowledge” or “experience” from long-term memory, which well emulates the information synthesizing function in biological working memory. The coexistence of short-term information encoding and multimodal information processing/synthesizing capabilities makes our artificial working memory an important complement to the brain-inspired hierarchical memory system, which can potentially endow neuromorphic systems with cognitive computing capability.

2. Using PTCM Device to Emulate the Synaptic Characteristics of Working Memory

Working memory is a subdivision of short-term memory, but it has a broader connotation than the basic short-term information storage devices.[10,11] Besides information storage function, working memory plays a critical role in the execution of a wide range of learning and cognitive tasks. There are two electrophysiological explanations for working memory. One theory gives an explanation from a neuron coding perspective and proposes that temporal information encoded in working memory might reside in stimulus-specific spiking activity.[12] However, holding information in a spiking form is energetically expensive because of the high metabolic cost of action potentials.[6,13] Another theory explains the short-term time span of working memory from intrinsic neuron cell properties. This explanation proposes a synaptic theory,[9] that the short-term information maintaining is attributed to the continuous accumulating Ca²⁺ levels in the synapse. As shown in Figure 1c,d, the continuous input action potentials from axon open the ionic gates of the synapse of presynaptic neuron. Ca²⁺ ions flow into the cell and increase the release probability of neuron transmitter in cleft, therefore,
enhancing the connection between pre- and postsynaptic neurons, leading to the increasing excitatory postsynaptic current (EPSC). Removal of residual calcium from presynaptic terminals is a relatively slow process. Therefore, the information can be transiently held for 10–15 s even without enhanced spiking activity. The synaptic theory reduces the need for metabolically costly action potentials, providing a more reasonable explanation for working memory.

Based on the biological fundamental, we propose a PTCM device with controllable ionic doping channels to mimic the Ca$^{2+}$ regulation and intracellular control in the synaptic theory of working memory, as shown in Figure 2a. In conventional electrochemical metallization devices, metal ions (like Ag$^+$ and Cu$^{2+}$) are electrically driven into a functional layer, which is a one-way doping process. The metal electrode could hardly recycle the cation from the functional layer. The continuous release of metallic ions over time will cause the formation of voids on electrode/dielectric interface and the drift of electrical performance.[14] We introduce an ion reservoir into a planar 2D device, which can release and recycle cations in the monolayer channel (MoS$_2$) under e-field. As schematically shown in Figure 2a, the Ag:GeSbTe (Ag:GST) mixed ionic–electronic conductor (MIEC)$^{[15,16]}$ acts as the cation reservoir and forms a heterojunction with monolayer MoS$_2$. The monolayer MoS$_2$ serves as the 2D semiconductor channel, which is deposited by chemical vapor deposition with high quality and large grain size (60 μm grain diameter), as shown in Figure S1, Supporting Information. The thickness of the MoS$_2$ single crystal is around 1 nm (Figure S2, Supporting Information), corresponding to the monolayer on the basis of 0.65 nm per layer.$^{[17]}$ The length of MoS$_2$ channel between cation reservoir (presynaptic terminal) and inert electrode (postsynaptic terminal) is ≈200 nm. It is worth pointing out that, compared with memristors with conventional metal–insulator–metal (MIM) vertical structures, the monolayer MoS$_2$ planner channel provides less spatial degrees of freedom for Ag$^+$ ions migrating and diffusing, which is more controllable for fine tuning cation concentration.

When we apply positive bias to the presynaptic terminal, Ag in MIEC will be ionized and driven into the MoS$_2$ channel. The injection of Ag$^+$ ions would significantly change the conductance of MoS$_2$ channel. As shown in Figure 2b, PTCM shows a rectifying analog switching behavior. It is worth noting that the rectifying characteristic of our device is induced by the injection and extraction of Ag$^+$ ions in the MoS$_2$ channel, which is different

![Figure 2. Device structure and electrical performance of the PTCM.](image-url)
from the Schottky-based rectifying characteristic in memristors with conventional MIM structure.\cite{18} The positive bias (from pre- to postsynaptic terminal) drives Ag$^+$ ions into MoS$_2$ channel and leads to the increase in channel conductance, whereas the negative bias (from post- to presynaptic terminal) drives the Ag$^+$ ions out of MoS$_2$ channel and leads to rapid drop of channel conductance. It is worth noting that Ag atoms have a small solid solubility in MoS$_2$ channel and a large solid solubility in MIEC. This difference in solid solubility results in the slow release of Ag in MoS$_2$ channel and rapid absorption by the MIEC reservoir once negative bias is applied. However, from the enlarged J–V characteristic for the negative bias sweep, we can still see a gradually decreasing process of the conductance, indicating the extracting process of Ag from MoS$_2$ channel. To support this discussion, we fabricated a series of devices with different structure configurations, as discussed in Figure S3, Supporting Information. The distinct switching behaviors of various interface combinations differentiate our PTCM from the Schottky-based, electrochemistry metallization and threshold switching memristors. The coexistence of analog switching behavior and rectifying characteristic in the PTCM benefits the high-density integration of passive computational array because the intrinsic one-way selectivity can help to suppress the leakage current and avoid the need of additional selector devices.\cite{19} Therefore, we can realize passive PTCM arrays without worrying about the sneak path issue.

The injection and diffusion dynamics of Ag$^+$ ions under e-field well emulate the migration of Ca$^{2+}$ ions in biological neurons. Therefore, the conductance change of MoS$_2$ channel shows obvious short-term memory characteristics under the continuous stimuli of pulse train. As shown in Figure 2c, when applying positive pulses to the presynaptic terminal, the conductance is continuously increased by sixfold. After potentiation, the high conductance state spontaneously relaxes to the initial conductance state in $\approx$50 s which emulates forgetting. This is because the Ag atoms have a small solid solubility in monolayer MoS$_2$. When the electric field force is removed or gradually decreases, the Ag atoms can hardly be maintained in MoS$_2$ channel and will be reabsorbed by the MIEC reservoir, resulting in the decrease in channel conductivity. The characteristic time (the time when conductance decreases by 50%) of the forgetting curve is around 10 s in line with the time span of biological working memory. As the conductance switching mechanism of the PTCM device is based on ion concentration modulation instead of stochastic filament bridging, it shows better yield and performance uniformity (Figure 2c), which help to improve the computing accuracy.\cite{20}

We apply pulse sequences with different pulse amplitudes to trigger the potentiation of the PTCM. As shown in Figure 2d, the conductance can be finely tuned to different final states. Interestingly, the relaxing times remain same. We also capture the pulse number-dependent forgetting time curves (in Figure S4, Supporting Information), and the high conductance state can fully relax to initial value in 20 s. This feature bioplusibly emulates the certain time span of working memory (normally 10–15 s).\cite{21} We also demonstrate several spike-based learning rules based on PTCM, including post-tetanic potentiation (Figure S5a, Supporting Information), paired pulse facilitation (Figure S5b, Supporting Information), and spiking timing-dependent plasticity (Figure S5c, Supporting Information). These results indicate that our device with rectifying analog switching has the biorealistic short-term synaptic learning/computing capability and can serve as the basic component for artificial working memory.\cite{22}

### 3. Frequency-Dependent Plasticity for Working Memory

In human working memory, the communication between neurons relies on sending and receiving spike (or action potential) sequences. The information is encoded in the form of timing and rating. Therefore, the frequency-dependent plasticity is essential for information encoding and processing. Our PTCM shows obvious frequency-dependent plasticity, as shown in Figure 3a. The final conductance state highly relates to the frequency of spike sequence. This is attributed to the finely controllable cation doping in the 2D MoS$_2$ channel. Similar to the Ca$^{2+}$ migration and diffusion mechanism in biological neurons, Ag$^+$ ions are driven into MoS$_2$ channel in pulse duration and spontaneously diffuse in the pulse interval. If the pulse interval is too short for Ag$^+$ ions to diffuse, Ag$^+$ concentration will accumulate in the channel and continuously enhance the connection strength between pre- and postsynaptic terminals, resulting in the facilitation of EPSC. However, for low-frequency stimulation condition (with long pulse intervals), the Ag$^+$ ions have longer time to diffuse, and the concentration gradually reaches a lower equilibrium value in the MoS$_2$ channel, resulting in the decrease in channel conductance. Therefore, whether the PTCM exhibits facilitation or depression under pulse sequence is determined by the competition between Ag$^+$ ions’ accumulation and diffusion in MoS$_2$ channel.

With the excellent ion-regulation capability, the conductance of PTCM can continuously change under the pulse sequence with varying frequency. As shown in Figure 3b, we alternately switch the frequency between 1000 and 400 Hz, and the device exhibits alternating facilitation and depression correspondingly. Under high-frequency stimuli, the EPSC increases to 1.3 μA; whereas under low-frequency stimuli, the EPSC relaxes to 0.6 μA.

The synaptic theory of working memory requires the device to present the short-term information encoding and retrieval functions under different neural population activities.\cite{9} Here, by applying well-designed spike sequences, we use the PTCM to emulate the key synaptic features of working memory. First, a high-frequency pulse sequence is adopted to approximate the intense neuronal activity, and a low-frequency pulse sequence is used to emulate the background noise from inactive neurons. Each high-frequency pulse sequence segment represents an encoding event. In Figure 3c, the multiple high-frequency spike sequences with short intervals simulate the multiple encoding enhancement and frequent retrieval of information in working memory. It can be clearly seen that the EPSC gradually increases as the effective population stimuli accumulate. We also demonstrate retrieval after a longer time span, which reflects another key feature of working memory.\cite{9} As shown in Figure 3d, the first encoding event with high-frequency pulse sequence causes the significant enhancement of EPSC. Then, in 0.18 s, the surrounding neurons maintain low activity until a retrieval event.
occurs. The output EPSC of the retrieval is higher than the floor EPSC triggered by the very first encoding event, which means the device still “remembers” the previous encoding event. The above results indicate that our PTCM is frequency programmable and is competent for the unique short-term synapse in working memory.

4. Information Encoding, Processing, and Synthesizing Demonstrations for Working Memory

Through the cooperation of four main parts (central executive, visuospatial sketchpad, phonological loop, and episodic buffer), working memory dynamically processes vision and auditory information in real time and provides a high-speed interface between brain and surrounding environment.[23]

As a bridging component between memory hierarchies, working memory mainly provides three functions: encoding, to acquire information from the environment and consolidate it as short-term memory; processing, to identify and extract key features from the redundant information by adopting knowledges and experiences from long-term memory, select useful memory fragment, and form sparse information flow; code converting, to process and convert the redundant information flow in a sparse spike form and transmit it to the higher cerebral cortex for further processing, as shown in Figure 4a. These high-level information processing functions make working memory superior to the basic short-term information storage concept. Here, we design experiments and demonstrate the three main functions one by one.

We fabricate a crossbar array on a monolayer MoS2 grain, which contains 36 PTCMs (6 × 6), as shown in Figure 4b. The individual device structure in the array is shown in Figure S6, Supporting Information. As shown in Figure 4c, we demonstrate the short-term encoding function by writing “S” character into the 6 × 6 memristor array. We use 1000 Hz pulse sequences as programming input for the high-weighted pixels and 400 Hz pulse sequences as the programming input for the low-weighted pixels. The comparation between original images and programmed images is shown in Figure S7, Supporting Information. The conductance color map over programming time with a sampling read operation every 0.02 s is shown in Figure 4c. The “S” pattern becomes clearer as the input stimuli accumulates, indicating that the information is successfully written and consolidated into the crossbar array. Then we keep reading the conductance states of the cells in crossbar array and find that the pattern gradually disappears in 10 s (as shown in Figure 4d), indicating forgetting. When we write the same pattern into the crossbar array (Retrieval), the conductance of programmed devices rapidly reaches the highest value in 0.1 s, which is much faster than the initial encoding. When encoding a new character “U” into the crossbar array (Update), we can see that the pixels that overlapped with the previous pattern reach the higher conductance state faster, as shown in Figure 4d. This means that although the pattern almost fully decays in 10 s, the operation history is still remembered in PTCM, and the programmed pixels are easier to be recalled during retrieval. Till now, the short-term information encoding (acquisition and
consolidation), forgetting, retrieval, and update of working memory are well demonstrated in our PTCM array.

In human memory hierarchy, working memory can temporarily load the knowledge (or experience) from long-term memory and perform processing and encoding for the multimodal signals from sensory memory. As shown in Figure 5a, we proposed a brain-inspired memory hierarchical system design to perform multimodal information synthesizing and processing. In this system, working memory act as a configurable neural network to bridge the sensory memory and long-term memory (Figure 5a). To process inputs like 2D image and audio waveform, larger PTCM arrays are needed to form a deep neural network. However in practical, it is still challenging to grow wafer-scale high-quality monolayer MoS$_2$.[25] Therefore, we perform a system-level simulation instead of experimental implementation to verify the main functions of working memory. We build a compact model for the PTCM device to enable network simulation. The details of our device and array models are presented in Section 3, Supporting information. The simulation results are in good agreement with experimental results, as shown in Figure S10, Supporting Information. For the network-level simulation, as the conductance of our PTCM device only has positive value, we use the conductance difference between two memristor devices to represent one weight, as shown in Figure 5a.[26,27] A three-layer full-connection network configuration is adopted, with each hidden layer containing 100 Rectified Linear Unit (ReLU) neurons and the output layer containing ten sigmoid neurons, representing 10 classes.

In our simulation, three datasets with different features are used to train the working memory network and form the “knowledges” in long-term memory, which are MNIST dataset for Handwritten digits recognition,[28] Fashion-MNIST dataset for fashion image classification,[29] and AudioMNIST for voice recognition.[30] We built a simulator according to the devices’ parameters and conducted off-chip training in our simulator. In our design, the pretrained weight matrices for different tasks are stored in long-term memory. The parameters of neural network are trained using a hardware-friendly stochastic gradient descent algorithm (as described in Section 4, Supporting information). The image and audio inputs are stored in sensory memory before being fed into the working memory network. A preprocessing and spike sequence encoding layer is inserted between

Figure 4. Basic functions’ demonstration of working memory in a PTCM array. a) Schematic to show the three main functions of working memory. The coexistence of information encoding, processing, and code converting make working memory different from conventional short-term memory. b) Optical image of a 6 × 6 PTCM array integrated on a single MoS$_2$ grain, with the scale bar of 15 μm. c) Information encoding demonstration for working memory. An “S” character is parallelly written into a 6 × 6 PTCM crossbar array by a series of spike sequences. d) Forgetting followed by reprogramming (Retrieval and update) demonstration for working memory.
sensory memory and working memory, which can convert the 2D image and audio waveform into a 1D pulse sequence vector according to the value of the signal intensity. It should be noted that for calculating convenience, the audio waveforms are converted to the spectrum format first, which is done in the preprocessing layer. When conducting information processing, the working memory loads pretrained weights from long-term memory and accepts the input from sensory memory to conduct in-memory computing. After forward transmission, the output vector (translated to voltage form by sensing amplifiers) can be collected from the output layer. By comparing the actual output vector with the expected output vector, the weight increments could be calculated by applying the delta rule. Then, the weight increments are translated into pulse sequence and written back to the weight matrixes. After 30 epochs training (the training batch sizes are 60,000 for MNIST, 60,000 for fashion-MNIST,

Figure 5. A brain-inspired hierarchical memory system integrated with a working memory. a) Schematic diagram to show a brain-inspired hierarchical memory system, in which a working memory network bridges sensory memory and long-term memory for multimodal processing demonstration. b) Multimodal task processing demonstration. With different weight matrices loaded from long-term memory, the working memory network can sequentially deal with three different inputs, including handwritten digits, fashion images, and audio inputs. c) By replacing the output sigmoid neurons with stochastic neurons, the working memory network decodes an audio input of a postcode (487372) and translates it into a spike sequence output, which demonstrates the code-converting function.
and 24,000 for AudioMNIST, respectively), the training accuracies of the three datasets reach 96.10, 93.50, and 99.90%, respectively.

Then, we conduct multimodal information processing demonstrations based on the brain-inspired hierarchical memory system, as shown in Figure 5b. Three test datasets, that contain samples that never appeared in training sets, are prepared and fed into sensory memory, including 10,000 handwritten digit samples, 10,000 fashion image samples, and 6,000 audio samples. For task 1, we load handwritten weight matrix pairs into our working memory network. From the confusion matrix results, we can see that the working memory can well recognize the handwritten digit inputs, and the output values are in good agreement with the expected values for digit classification. At this moment, the network cannot recognize image and audio waveform inputs. Next, we leave the network unstimulated for 10 s, so it can forget the previously loaded weight matrix pairs. For task 2 and task 3, we sequentially load the fashion image classification and audio recognition weight matrices into the sensory memory. As we expected, for task 2, the fashion image test input gets the best classification result, and for task 3, the audio test input gets the best classification result. The recognition accuracies of the three tasks are 94.58, 90.71, and 99.55%, respectively. The above results indicate that the brain-inspired hierarchical memory system has the ability to process multimodal information in different contexts.

In Figure 5c, we also demonstrate the code-converting function in the working memory network. To enable stochastic spike encoding, we use ten stochastic neurons to replace the sigmoid neurons in output layer. The stochastic neuron model is described in Figure S12, Supporting Information. A 6-digit post-code voice signal (487 372) is encoded and sequentially input into the working memory network. With weight matrices trained by AudioMNIST dataset loaded, the ten stochastic neurons show different levels of activity and produce featured spike sequences with spiking rate changing over time. Each data point in Figure 5c (bottom figure) represents a firing event of stochastic neuron. By directly reading from the spiking rates to identify active neurons, we can easily recover the postcode sequence.

5. Conclusion

In summary, this work first proposes an artificial working memory based on a PTCM memristive array. With the ionic reservoir and 2D channel design, the PTCM presents excellent conductance-controlling capability. The frequency-dependent plasticity and rectified analog switching behavior make it successfully reproduce several synaptic characteristics of working memory. By integrating 36 PTCM devices into a 6 × 6 crossbar array on a monolayer MoS2 grain, the information encoding (acquisition and consolidation), forgetting, retrieval, and update operation schemes are verified in array level. A compact model of the PTCM device is also established. Based on it, we conduct system-level simulation and demonstrate the information processing and code-converting functions based on a brain-inspired hierarchical memory system, in which the working memory network acts as a bridging part to process and synthesize information from sensory memory and long-term memory.

6. Experimental Section

Monolayer MoS2 Synthesis: The synthesis of monolayer MoS2 was conducted in a CVD system with double-heating zones. Sulfur (Sigma-Aldrich, 5N) and MoO3 (Sigma-Aldrich, 5N) powders were used as the reactant precursors to synthesize MoS2 on a SiO2/Si substrate. The argon gas acted as the carrier gas to carry sulfur and MoO3 to the substrate. In a vacuum environment, the heating zone was heated up to 800 °C in 30 min and maintained for 5 min. Sulfur powders were heated up to 180 °C during the reaction process. Subsequently, the substrates carrying the monolayer MoS2 were naturally cooled to room temperature under the protection of argon gas environment.

Device and Crossbar Array Fabrication: The single cell and crossbar array of the PTCM device followed the same fabrication process. First, 50 nm TiW was deposited on as-grown monolayer MoS2 using a magnetron sputtering system, followed by e-beam lithography and liftoff process, serving as inert electrode (or postsynaptic terminal). Subsequently, 60 nm SiN was deposited as dielectric layer, which could fully cover the inert electrode to isolate it from other active layers. Then, Ag and CST were cosputtered to form 50 nm MIEC ion reservoir layer. The distance between the ion reservoir and inert electrode was precisely controlled to 200 nm by e-beam lithography method, forming the channel. Finally, another 50 nm TiW was formed on the ion reservoir layer and served as the presynaptic terminal electrode.

Electrical Characterization: The I–V characteristic of the PTCM device was measured using a Keithley 4200 semiconductor parameter analyzer (4200-SCS) with a Cascade Microtech Summit 11 000 probe station. All the direct current (DC) sweeping curves are collected at a same sweeping rate. The pulse mode measurement of the PTCM device was carried out using a Keithley 4225-PUM (Pulse Measurement Unit). The 6 × 6 PTCM device array was bonded to a home-made testing board. The input voltage pulse sequence and output current were applied and collected parallelly by the semiconductor parameter analyzer.

Working Memory Network Simulation: The single device and network-level simulation in this article was executed in Spyder 4 using Python 3.7 programming language. The hardware training method is discussed in Supporting Information. The datasets used for training were MINIST,[28] Fashion-MNIST,[29] and AudioMNIST[30] for handwritten digit recognition, fashion image classification, and audio signal recognition, respectively. The audio samples for training and testing were decoded using librosa package.[31] The 1D waveform signal was transformed to 2D spectrum format. Then, we down scaled the spectrum image to 28 × 28. In this way, the spectrum image can be flattened into spike train and fed into the neural network to conduct classification.

Supporting Information

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

Acknowledgements

This work was supported by National Key R&D Program of China (No. 2018YFE0200200), National Nature Science Foundation of China (No. 61836004), Brain-Science Special Program of Beijing under grant Z181100001518006, The Suzhou-Tsinghua Innovation Leading Program (No. 2016SZ0102), and CETC Haikang Group-Brain Inspired Computing Joint Research Center.

Conflict of Interest

The authors declare no conflict of interest.
Data Availability Statement

Data available on request from the authors.

Keywords

artificial working memory, brain-inspired hierarchical memory, planar 2D channel memristors, short-term synapses

Received: June 14, 2021
Revised: August 12, 2021
Published online: October 7, 2021

[1] M. W. Brown, J. P. Aggleton, Nat. Rev. Neurosci. 2001, 2, 51.
[2] H. E. Lee, J. H. Park, T. J. Kim, D. Im, J. H. Shin, D. H. Kim, B. Mohammad, I.-S. Kang, K. J. Lee, Adv. Funct. Mater. 2018, 28, 1801690.
[3] K. J. Malmberg, J. G. W. Raaijmakers, R. M. Shiffrin, Mem. Cognit. 2019, 47, 561.
[4] J. S. Nairne, Annu. Rev. Psychol. 2002, 53, 53.
[5] T. S. Braver, J. D. Cohen, Cognit. Process. 2001, 2.
[6] A. Baddeley, Curr. Biol. 2010, 20, R136.
[7] M. Persiani, A. M. Franchi, G. Gini, Cognit. Syst. Res. 2018, 51, 1.
[8] R. C. O’reilly, M. J. Frank, Neural Comput. 2006, 18, 283.
[9] G. Mongillo, O. Barak, M. Tsodyks, Science 2008, 319, 1543.
[10] T. Chang, S.-H. Jo, W. Lu ACS Nano 2011, 5, 7669.
[11] Y. Park, J.-S. Lee, ACS Nano 2017, 11, 8962.
[12] E. K. Miller, C. A. Erickson, R. Desimone, J. Neurosci. 1996, 16, 5154.
[13] D. Attwell, S. B. Laughlin, J. Cereb. Blood Flow Metab. 2001, 21, 1133.
[14] Y. Yang, P. Gao, L. Li, X. Pan, S. Tappertzhofen, S. Choi, R. Waser, I. Valov, W. D. Lu, Nat. Commun. 2014, 5, 4232.
[15] X. Ji, C. Wang, K. G. Lim, C. C. Tan, T. C. Chong, R. Zhao, ACS Appl. Mater. Interfaces 2019, 11, 20965.
[16] X. Ji, L. Song, W. He, K. Huang, Z. Yan, S. Zhong, Y. Zhang, R. Zhao, ACS Appl. Mater. Interfaces 2018, 10, 10165.
[17] B. Radisavljevic, A. Radenovic, J. Brivio, V. Giacometti, A. Kis, Nat. Nanotechnol. 2011, 6, 147.
[18] D. Li, B. Wu, X. Zhu, J. Wang, B. Ryu, W. D. Lu, W. Lu, X. Liang, ACS Nano 2018, 12, 9240.
[19] B. H. Mun, B. K. You, S. R. Yang, H. G. Yoo, J. M. Kim, W. I. Park, Y. Yin, M. Byun, Y. S. Jung, K. J. Lee, ACS Nano 2015, 9, 4120.
[20] J. D. Greenlee, W. L. Calley, M. W. Moseley, W. A. Doolittle, IEEE Trans. Electron Devices 2013, 60, 427.
[21] B. E. Goldstein, J. Chem. Inf. Model. 2013. https://doi.org/10.1017/CBO9781107415324.004.
[22] J. Pei, L. Deng, S. Song, M. Zhao, Y. Zhang, S. Wu, G. Wang, Z. Zou, Z. Wu, W. He, F. Chen, N. Deng, S. Wu, Y. Wang, Y. Wu, Z. Yang, C. Ma, G. Li, W. Han, H. Li, H. Wu, R. Zhao, Y. Xie, L. Shi, Nature 2019, 572, 106.
[23] A. D. Baddeley, G. Hitch, in Exploring Working Memory, Routledge, England 2017, pp. 43–79.
[24] Y. LeCun, in 2019 IEEE Inter. Solid-State Circuits Conf.- (ISSCC), IEEE, Piscataway, NJ 2019, pp. 12–19.
[25] S. Chen, M. R. Mahmoodi, Y. Shi, C. Mahata, B. Yuan, X. Liang, C. Wen, F. Hui, D. Akinwande, D. B. Strukov, M. Lanza. Nat. Electron. 2020, 3, 638.
[26] M. Prezioso, F. Merrikh-Bayat, B. D. Hoskins, G. C. Adam, K. K. Likharev, D. B. Strukov, Nature 2015, 521, 61.
[27] F. M. Bayat, M. Prezioso, B. Chakrabarti, H. Nili, I. Kataeva, D. Strukov, Nat. Commun. 2018, 9, 2331.
[28] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Proc. IEEE 1998, 86, 2278.
[29] H. Xiao, K. Rasul, R. Vollgraf, arXiv Prepr. arXiv1708.07747, 2017.
[30] S. Becker, M. Ackermann, S. Lapuschkin, K.-R. Müller, W. Samek, arXiv Prepr. arXiv1807.03418, 2018.
[31] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, O. Nieto, in Proc. of the 14th Python in Science Conf., Austin, TX, USA 2015, pp. 18–24. https://doi.org/10.25080/majora-7b98e3ed-003.