nature materials

Article

https://doi.org/10.1038/s41563-023-01704-z

Monolithic 3D integration of 2D materials-based electronics towards ultimate edge computing solutions

Received: 6 June 2023

Accepted: 27 September 2023

Published online: 27 November 2023

Check for updates

Ji-Hoon Kang[1,2,3,18], Heechang Shin[4,18], Ki Seok Kim[1,2,18], Min-Kyu Song[1,2,18], Doyoon Lee[1,2], Yuan Meng[5], Chanyeol Choi[2,6], Jun Min Suh[1,2], Beom Jin Kim[4], Hyunseok Kim[1,2], Anh Tuan Hoang[4], Bo-In Park[1,2], Guanyu Zhou[7], Suresh Sundaram[8,9], Phuong Vuong[9], Jiho Shin[1,2], Jinyeong Choe[4], Zhihao Xu[10], Rehan Younas[7], Justin S. Kim[10], Sangmoon Han[5], Sangho Lee[1,2], Sun Ok Kim[5], Beomseok Kang[1,2], Seungju Seo[1,2], Hyojung Ahn[11,12], Seunghwan Seo[1,2], Kate Reidy[13], Eugene Park[13], Sungchul Mun[14,15], Min-Chul Park[16], Suyoun Lee[16], Hyung-Jun Kim[16], Hyun S. Kum[4], Peng Lin[1,2,17], Christopher Hinkle[7], Abdallah Ougazzaden[8,9], Jong-Hyun Ahn[4] ✉, Jeehwan Kim[1,2,13] ✉ & Sang-Hoon Bae[5,10] ✉

Three-dimensional (3D) hetero-integration technology is poised to revolutionize the field of electronics by stacking functional layers vertically, thereby creating novel 3D circuity architectures with high integration density and unparalleled multifunctionality. However, the conventional 3D integration technique involves complex wafer processing and intricate interlayer wiring. Here we demonstrate monolithic 3D integration of two-dimensional, material-based artificial intelligence (AI)-processing hardware with ultimate integrability and multifunctionality. A total of six layers of transistor and memristor arrays were vertically integrated into a 3D nanosystem to perform AI tasks, by peeling and stacking of AI processing layers made from bottom-up synthesized two-dimensional materials. This fully monolithic-3D-integrated AI system substantially reduces processing time, voltage drops, latency and footprint due to its densely packed AI processing layers with dense interlayer connectivity. The successful demonstration of this monolithic-3D-integrated AI system will not only provide a material-level solution for hetero-integration of electronics, but also pave the way for unprecedented multifunctional computing hardware with ultimate parallelism.

The development of a system on a chip led to innovation in the realm of the integrated chip by providing several advantages, including flexibility in interfacing, better power efficiency, the ability to reconfigure the hardware and miniaturization of integrated chips[1,2]. However, because its lateral integration nature fundamentally limits further downscaling and miniaturization of integrated systems, there is a pressing need for a new integration strategy. A three-dimensional (3D) heterogeneous integration (3DHI) technology has become established as a promising

A full list of affiliations appears at the end of the paper. ✉e-mail: ahnj@yonsei.ac.kr; jeehwan@mit.edu; sbae22@wustl.edu

candidate to tackle these limitations of the system on a chip[3–5]. This 3DHI is a technology that allows stacking of different types of semiconductor device wafer on top of each other in 3D. Various electronic components such as memory, logic and opto-electronics can thus be vertically combined into a single unit to create smaller and more effective electronic devices[5,6]. In addition, the integration of different technologies, including complementary metal-oxide semiconductor (CMOS) circuits and microelectromechanical systems (MEMS), could lead to the creation of new functionalities such as the integration of sensors and actuators with digital logic[7,8]. However, connecting active device layers to each other also creates an extremely high technical barrier[9,10]: it requires precise hole drilling through the wafers, so-called through-silicon-via and solder bump bonding of each wafer die. Thus, conventional 3DHI requires extremely complicated wafer fabrication and bonding processing, which severely constrain chip integrability. Monolithic 3D (M3D) integration is regarded as an alternative solution for more efficient chip connection because all functional device layers are directly connected without wafers[11–13]. Nevertheless, the required removal of device layers from the substrate sets another technical challenge in regard to practical applicability on account of their intrinsically brittle nature and high internal stress level, and thus handling such layers could easily result in mechanical device failure. The emergence of two-dimensional (2D) materials-based electronics, in contrast, highlights their considerable potential in overcoming the above issues[14–18]. Due to the atomically thin nature of 2D materials, these possess intrinsically extremely low stiffness and almost zero internal stress. Accordingly, the physical constraints of M3D integration imposed by conventional rigid 3D materials can be completely overcome with 2D material-based electronics that perform on a par with their conventional, silicon-based counterparts.

Here we demonstrate the M3D integration of fully 2D material-based electronics to highlight wafer-free device stacking. Given the limited capability of academic fabrication settings, we attempted to mimic the vertical heterogenous integration of logic and memory by monolithic stacking of 2D material-based transistors and memristors, respectively. These stacked structures finally function in regard to artificial intelligence (AI) hardware in edge computing applications. As shown in Fig. 1a, an M3D-integrated system enables the hardware implementation of AI processors by stacking 2D material-based multifunctional layers including sensory layers, signal-processing layers and AI computing layers. Depending on the purpose of the application, different combinations of layers can be designed. Various sensors integrated into a sensory layer can provide redundant and complementary information that compensates for errors and increases accuracy through sensor fusion[19–22]. Different filters and amplifiers can be implemented on signal-processing layers for input data enhancement. Finally, multiple AI computing layers perform AI computations according to different AI applications. The configuration of the AI computing layer can be readily modified, taking into account the balance between AI computing complexity and power consumption. This study demonstrates a fully stackable, non-von Neumann architecture-based, two-tier, AI computing layer consisting of memristors and transistors. The combination of non-von Neumann architecture and M3D integration enables near/in-sensor computing that improves processing time latency and power consumption by reducing the physical distance between the sensor and processing unit[23]. Due to the outstanding intrinsic properties of 2D materials, we succeeded in securing a high degree of integrability that facilitated a high degree of freedom in vertical integration, allowing vertical integration of a total of six layers of 2D material-based electronic devices on a single chip. First we obtained large-area uniform 2D heterostructures by leveraging our layer-resolved splitting technique and semidry transfer method[24]. Despite the outstanding mechanical advantages of 2D materials in regard to M3D integration, existing methods have limitations in maintaining pristine interfaces in 2D material stacks because the interface is affected by mechanical deformations and process residues (mostly polymer). This severely limits the large-scale realization of M3D integration of devices based on multiple layers of 2D materials[25–29]. However, using our splitting technique and semidry transfer method[24], large-area stacked 2D heterostructures were secured with ultraclean interfaces. Based on this, 2D material-based memristors operating at large scale were implemented. Second, 2D transistors were monolithically integrated with 2D memristor arrays as a supporting circuitry. Last, we successfully achieved M3D integration of such device layers, implementing more efficient AI hardware. Combining M3D integration technology with non-von Neumann-based, in/near-sensor computing architecture enables efficient edge computing by direct processing of data from adjacent layers, minimizing redundant data transmission. The fully M3D-integrated AI system has thus demonstrated much faster processing time, lower voltage drops, lower latency and a smaller footprint. We strongly believe that the successful demonstration of M3D integration-based 2D material electronics will not only improve computing performance but also open up new possibilities for advanced integration of electronics toward ultimate area-effective and multifunctional systems.

To develop monolithically stackable AI hardware we developed 2D material-based memristors that store and compute information contemporally and are readily stackable. Among various types of memristor we chose to construct conductive bridge random access memory, because resistive switching can be naturally achieved almost regardless of the switching medium providing that defects are vertically aligned through the electrodes. In the same manner as the typical utilization of a metal oxide layer as an insulating switching medium for conductive bridge random access memory, an *h*-BN layer—an insulating 2D material with a bandgap of ~5.9 eV—was tested as a switching medium. *h*-BN film was transferred on patterned Pt/Cr layers on SiO_2/Si substrates. Next, Ag was deposited and patterned using conventional photolithography. The Ag electrode served as a reactive electrode for electrochemical metallization (Supplementary Note 1). As expected, the *h*-BN-based memristors showed resistive switching behaviour with a good on:off ratio of around 10^3 (Supplementary Fig. 1). However, unstable and non-uniform switching was observed in set and reset processes. Previous findings have shown that integration of semiconducting and insulating layers can result in enhanced electrical performance, including endurance, on:off ratio and uniformity[30–33]. Thus we configured double layers of WSe_2 and *h*-BN by transfer of WSe_2 film on *h*-BN (Supplementary Fig. 2). This approach led to the development of a reliable memristor with excellent electrical properties, achieved by engineering the semiconducting/insulating double-layer configuration—Ag/WSe_2/*h*-BN/Pt/Cr/SiO_2/Si. As shown in Fig. 1b,c and Supplementary Fig. 3, double-layer (WSe_2/*h*-BN)-based memristors showed good set–reset behaviour at low set voltage with sufficient on:off ratio and stable multistate retention for 100 s, which represents an enhanced performance compared with that of WSe_2-based memristors. (Supplementary Figs. 3–5). Moreover, good endurance under >1,000 cycles of set–reset processes was also confirmed (Fig. 1d and Supplementary Fig. 4a), which is mandatory for computing operation. Based on comparative study, layer configuration using 2D materials enabled optimization of electrical performance by taking advantage of both layers. Fine-tuning of switching performance was achieved by implementation of further studies on the ideal combination and thickness of 2D materials. It should be noted that, for wafer-scale fabrication of neuromorphic computing hardware on 2D materials, it is crucial to secure appropriate transfer methods that do not contaminate the interfaces of the heterostructures because minor contaminants can easily affect ion migration for resistive switching. Thus our semidry transfer, termed layer-resolved transfer[24], was utilized to maintain clean interfaces. Two-dimensional memristor heterostructures created by a conventional wet-transfer process failed to switch memristor devices because of processing

https://doi.org/10.1038/s41563-023-01704-z

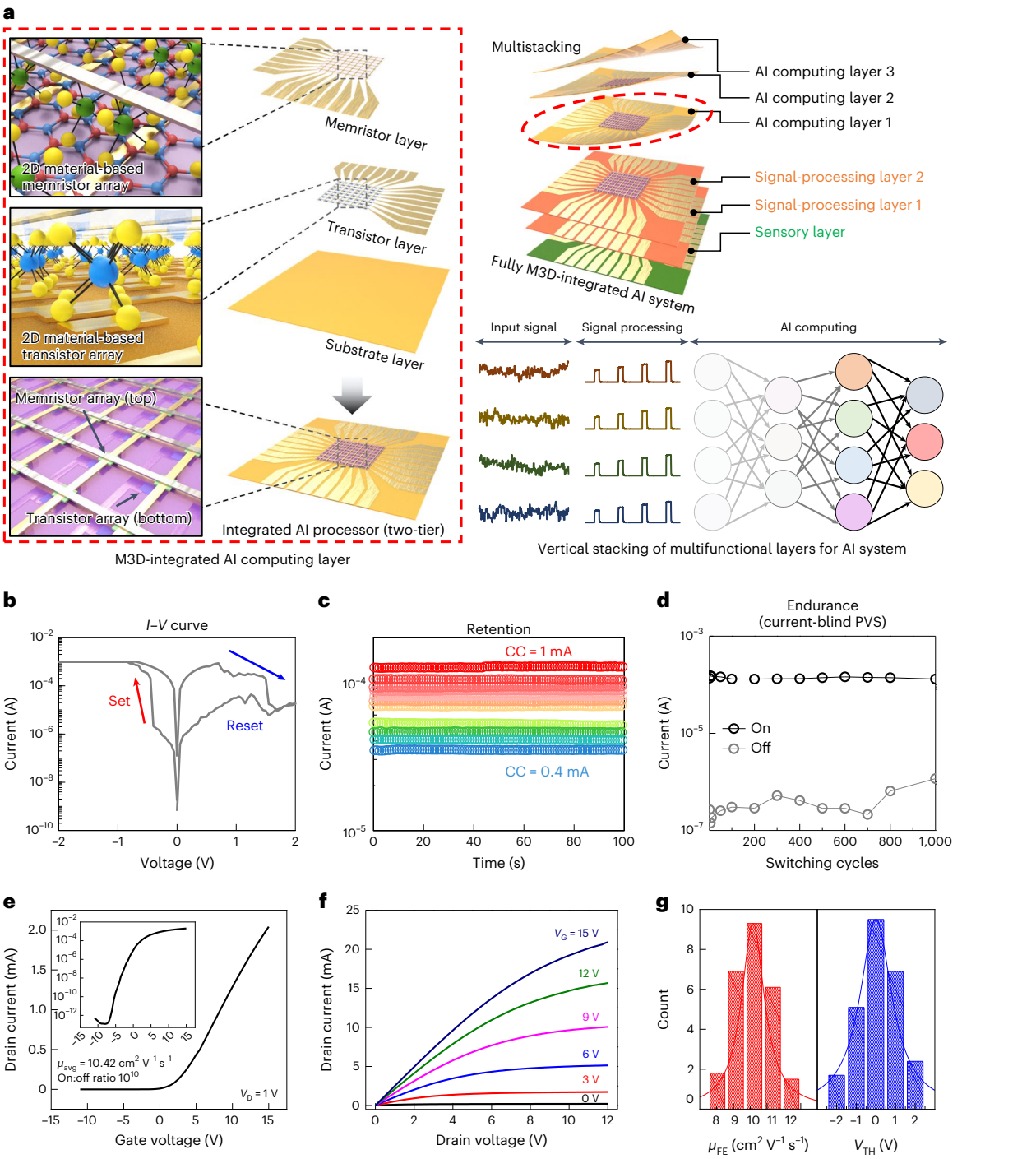

Fig. 1 | M3D integration of 2D material-based memristors and transistors. **a**, Schematic illustration of ultimate edge computing system based on M3D-integrated, 2D material-based electronics. The schematic of the M3D-integrated system (top) illustrates multistacking of different functional layers including AI computing layers, signal-processing layers and a sensory layer. All layers can be monolithically integrated into the 3D heterostructure. As shown in the inset on the left, an AI computing layer consists of a 2D material-based memristor array (top) and 2D material-based transistor array (middle) to construct an integrated, 2D-based AI processor (bottom). Both memristor and transistor arrays were M3D integrated. The schematic illustration of the computing system (bottom) shows the sequence of data processing throughout the system. Input signal is detected by the sensory layer and transferred to the signal-processing layers, which convert the signal to an appropriate configuration as an input to AI computing layers. Finally, AI computing layers play a role in cognitive computing by utilization of each layer in parallel. **b**–**d**, Electrical performance of double-layer (WSe_2/*h*-BN)-based memristors. Excellent memristor performance was confirmed by direct current (DC) switching performance of set and reset processes (**b**), multistate data retention for 100 s programmed under various current compliance levels (**c**) and an endurance test under 1,000 cycles of set and reset pulsed voltage stresses (PVS) (**d**). **e**,**f**, Electrical performance of bilayer MoS_2 TFTs. High on-current and on:off ratio with high device-to-device uniformity were confirmed by transfer characteristics (I_D–V_G) under V_D = 1 V (**e**) and output characteristics (I_D–V_D) under various V_G (**f**). **e**, Inset shows the transfer curve in log scale. **g**, Histograms of μ_{FE} and V_{TH}. CC, compliance current; V_D, drain voltage. μ_{avg}, average field-effect mobility.

https://doi.org/10.1038/s41563-023-01704-z

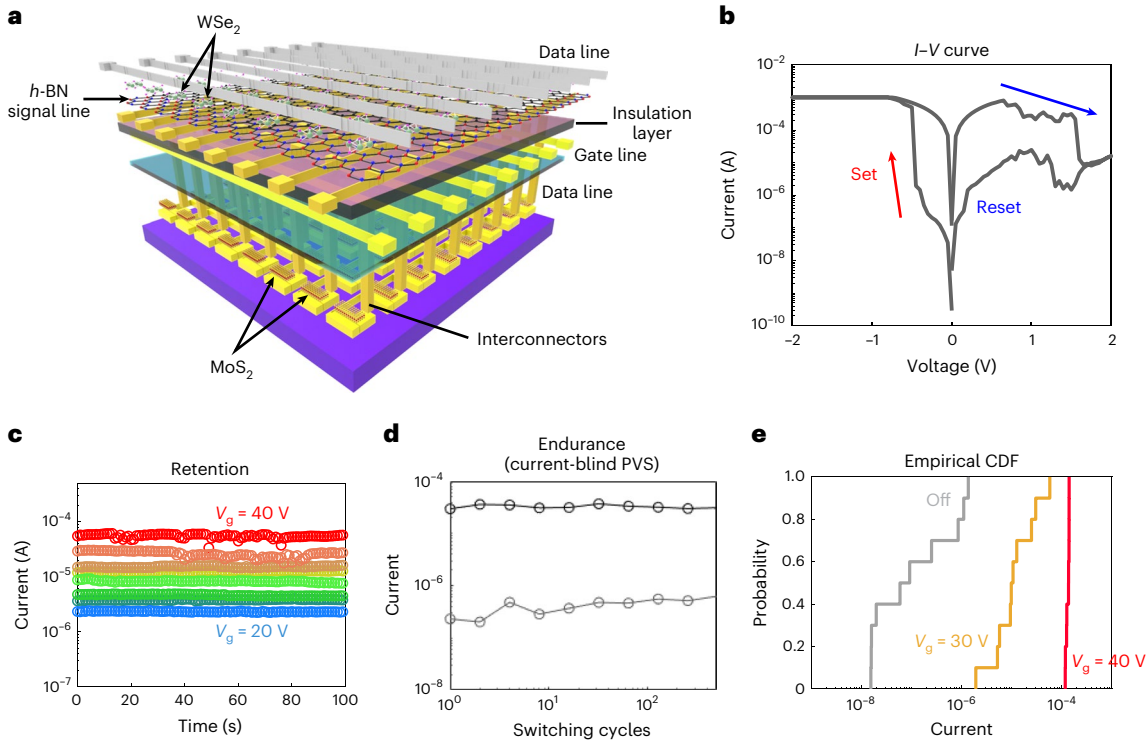

Fig. 2 | M3D-integrated 2D material-based 1 transistor-1 memristor arrays for AI processing. a, Schematic diagram of M3D-integrated AI processor comprising WSe_2/*h*-BN-based memristors and MoS_2-based transistors. The drain electrode line of the transistors is connected to the bottom electrode of the memristors. **b**, DC switching performance of the M3D-integrated device. Red and blue arrows indicate set and reset processes, respectively. **c**, Multistate retention properties of the M3D-integrated device array over 100 s. Multiple states were controlled by gate bias on MoS_2-based transistors. **d**, Endurance test of the M3D-integrated device array. **e**, Empirical CDF of the M3D-integrated device array with various values of V_G.

residue, which impacts ion diffusion kinetics in 2D double layers (Supplementary Fig. 6).

Next we fabricated 2D transistor arrays for monolithic connection to 2D memristor arrays as a driving circuitry to control switching behaviour. We attempted to modulate the device performance of both transistor and memristor arrays to match the current density of both devices for seamless monolithic integration. We tuned 2D transistors to have a higher on-current and lower off-current than those of 2D memristors[32]. To ensure that the transistor footprint matched with that of the memristors, we attempted to match the current requirement by maximizing the performance of 2D transistors rather than increasing channel dimension. We mainly engineered the interface to modulate the energy barrier arising from metal-induced gap states[34] and Coulombic scattering[35,36], by deposition of a 10-nm-thick Al_2O_3 planarization layer on the substrates and MoS_2 films, which led to high on-current (Supplementary Fig. 7). Metal-organic chemical vapour deposition (MOCVD)-grown MoS_2 was transferred to the source/drain metal contacts and subsequently a top-gate dielectric Al_2O_3 layer was deposited on the MoS_2 film. The transfer characteristics (drain current-gate voltage, I_D–V_G) of bilayer MoS_2 thin-film transistors (TFTs) exhibited field-effect mobility (μ_{FE}) of 10.42 ± 2.1 $cm^2\ V s^{-1}$, an on:off ratio of 10^9, a subthreshold swing of 0.612 ± 0.05 V dec^{-1} and threshold voltage (V_{TH}) of 1.7 ± 0.8 V under a drain voltage of 1 V (Fig. 1e,f and Supplementary Figs. 8 and 9). The average values of on and off currents were 2.02 mA and 1.80 pA, respectively. Histograms of the evaluated μ_{FE} and V_{TH} of the MoS_2 transistor array represent uniform characteristics with average values of 10.17 $cm^2\ V s^{-1}$ and 0.87 V, respectively (Fig. 1g). Detailed fabrication processes and characterization conditions are discussed in Methods.

Given the separately optimized stackable AI hardware and driving circuitry based on 2D materials, we attempted to integrate these monolithically. We chose to integrate WSe_2/*h*-BN-based neuromorphic computing arrays on top of the MoS_2-based transistors because construction of neuromorphic computing arrays can be completed by stacking and depositing each component at room temperature. Following integration of transistor arrays on the substrate, we deposited Al_2O_3 layers as an isolation layer followed by the formation of via-holes on the source region of the underlying MoS_2 transistors. It is important to note that electronic disconnection can easily happen when processing residue, such as when photoresist residue is present on the drain electrode of the driving circuitry. To address this issue we implemented a two-step etching process. Initially, reactive ion etching was utilized to create via-holes until reaching the etch stop layers, which is a drain electrode. Next, wet chemical etching using buffered oxide etchant was performed to ensure complete removal of any remaining residue. Even after thorough cleaning of the top surface of the driving circuitry, the deposition of solely Au as an interconnector still resulted in open circuitry due to the weak adhesion of Au layers. We therefore first deposited 10 nm of Cu, a material well known for its strong soldering properties, and then deposited Au on top of the Cu layer. The word lines of the memristor crossbar arrays were formed across the Au fills, which provided a direct connection between transistors and memristors. We constructed WSe_2/*h*-BN-based memristor arrays on top of the underlying word lines by transferring WSe_2 and *h*-BN layers. Finally, memristor-based AI processors were fully integrated by the formation of bit-line metal arrays on the memristors as depicted in Fig. 2a and Supplementary Fig. 10. It is critical to perform semidry transfer of WSe_2 and *h*-BN to construct memristors to maintain the performance of memristors (Methods). As shown in Fig. 2b, the performance of memristors integrated on the transistors was comparable to that of those on a plane substrate. When correct integration is compromised it disrupts the seamless flow of electrical signals, impeding the normal operation and performance of the system. These issues can result in undesirable consequences such as an open circuit or incorrect set–reset behaviour

https://doi.org/10.1038/s41563-023-01704-z

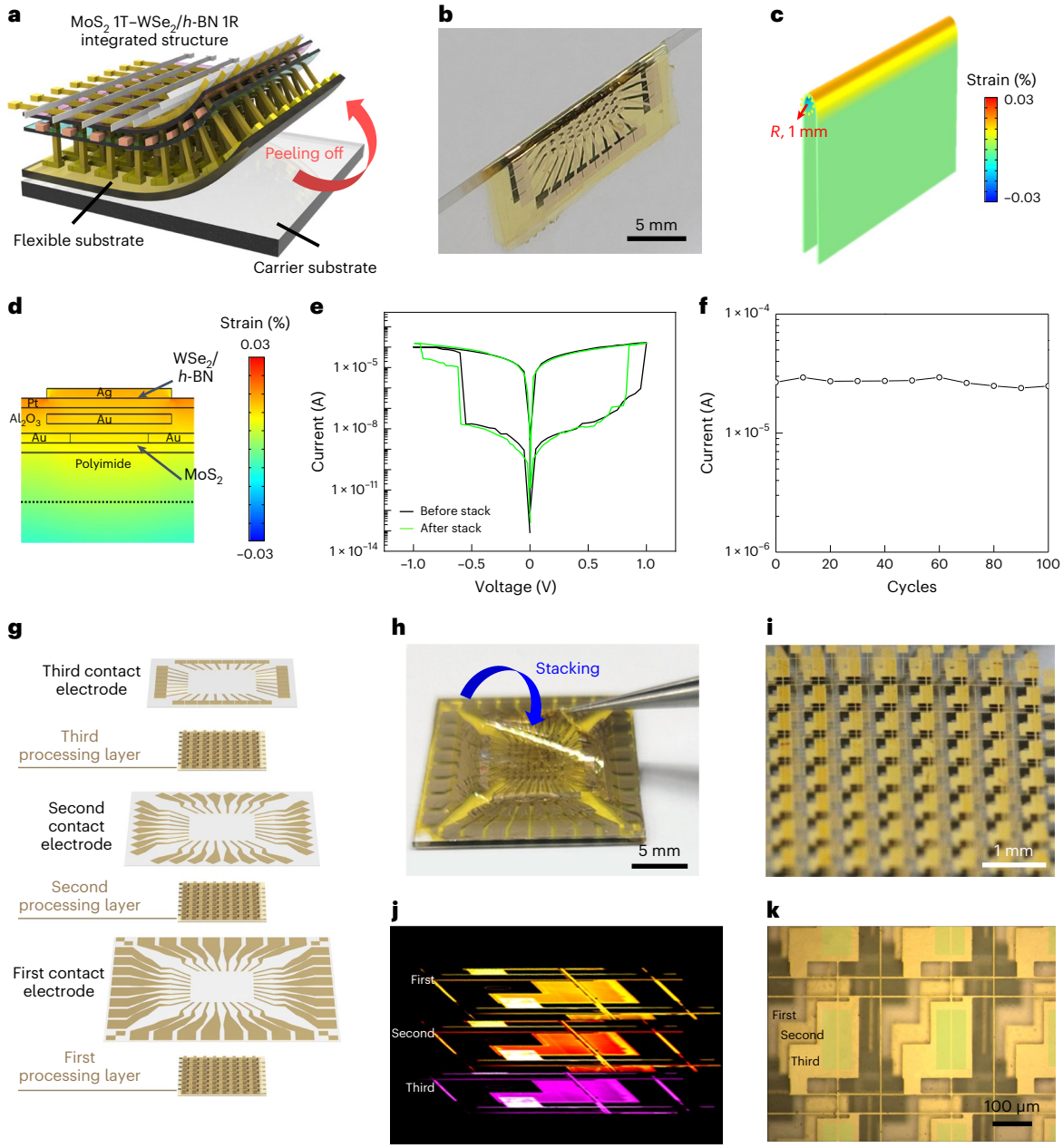

Fig. 3 | Peeling and stacking of M3D-integrated, 2D material-based devices. **a**, Schematic diagram of peelable M3D-integrated devices. The devices were precisely peeled off whole for M3D integration following fabrication. **b**, Optical image of M3D-integrated device array. **c**,**d**, Distribution of strain under bending on a substrate (**c**) and in an M3D-integrated device (**d**). **e**, DC switching performance of the M3D-integrated device before peeling and after transfer. No degradation in electrical performance was observed following transfer. **f**, Mechanical robustness of M3D-integrated device array as measured by repetitive bending test in up to 100 cycles. **g**, Multiple stacking of M3D-integrated AI processors. **h**–**k**, Photographs showing low resolution (**h**) and high resolution (**i**) of three stacks of AI processors (**j**) and an optical microscopy image showing overlap of multiple layers (**k**). Three AI processing layers were M3D integrated into the AI processing system.

(Supplementary Fig. 11). Figure 2c shows successful hybridization of the function via monolithic integration, where multiple conductance states were precisely controlled by modulation of the gate voltages of MoS_2-based transistors. Limiting transistor current by gate bias through the memristors during set operation enables precise weight programming[37,38]. Following successful integration, excellent conductance retention was observed similar to that of control devices on plane wafers. We statistically confirmed the operation of M3D-integrated AI hardware (Supplementary Figs. 12 and 13). In addition this showed good endurance behaviour over >300 cycles of switching, and empirical cumulative distribution function (CDF) exhibited readily controllable multiple resistance states, as shown in Fig. 2d,e. These 2D material-based memristors and transistors could be further improved by direct growth of 2D materials and further optimization of 2D layer configuration in terms of electrical performance and device yield[17,39].

Owing to their extreme flexibility and stackability, and while maintaining their device functionality, 2D material-based devices can take full advantage of monolithic 3D integration. Because the wafer-based bonding process, including through-silicon-via and bonding via solder bumps, can be avoided in M3D, the number of stacked layers is not limited in principle. This ultimate degree of freedom in layer stacking with our 2D material-based M3D strategy would allow unprecedented levels of integration and functionality in electronic systems. To demonstrate this potential we attempted to stack three layers of M3D-integrated

https://doi.org/10.1038/s41563-023-01704-z

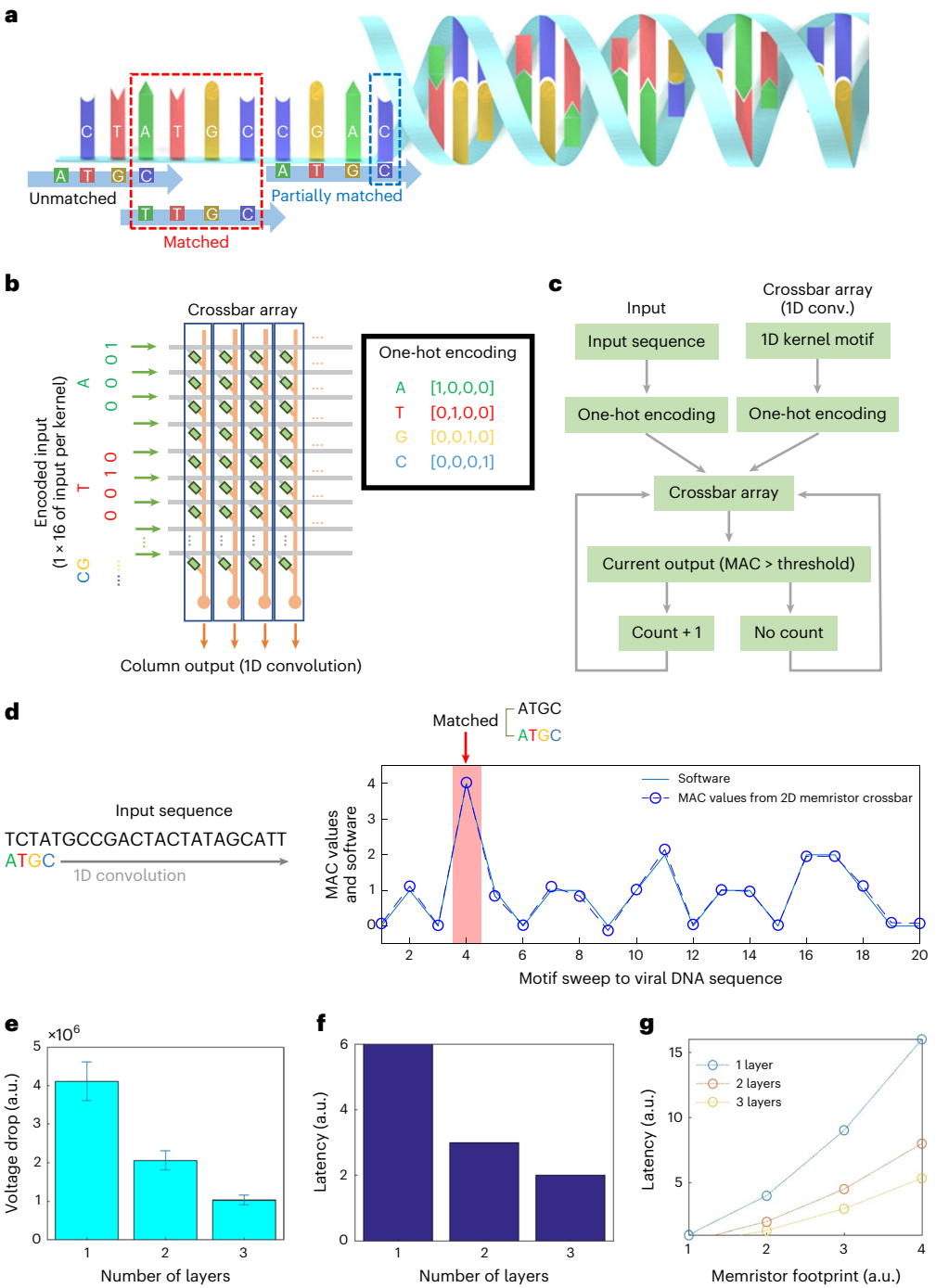

Fig. 4 | DNA motif discovery using the M3D-integrated, 2D material-based AI system. a, Schematic illustration of DNA motif discovery by 1D convolution. **b**, Implementation of 1D kernel pattern of four DNA bases (A, T, C and G) by one-hot encoding. **c**, Flow chart of DNA motif discovery by 1D convolution. **d**, Output values of MAC and software results of DNA motif sweep by 1D convolution. **e**,**f**, Voltage drop across computing devices (**e**), and computing latency as a function of the number of multistacks of the M3D-integrated, 2D material-based AI hardware (**f**). Error bars represent s.d. ($n = 100$). **g**, Computing latency as a function of M3D-integrated, 2D material-based AI system footprint.

AI processors with driving circuits, resulting in stacking a total of six layers of device arrays (three transistor and three memristor arrays) as shown in Fig. 3g–k, where reduction in both voltage drops and latency were expected while performing computations. Moreover, heavier machine learning tasks could be implemented with smaller-sized artificial synapse arrays because series vertical connection of artificial synapses is facilitated by stacking (detailed discussion below). To construct such multiple M3D-integrated layers we must ensure the mechanical robustness of the stacked device arrays. As shown in Fig. 3a and Supplementary Fig. 14, we first performed M3D integration of a transistor–memristor array on an ultrathin flexible PI substrate that is readily peelable and stackable. The multistack of the M3D-integrated, 2D-based AI processor shows outstanding mechanical performance (Fig. 3b), with a bending radius of 1 mm. The originated tensile strain was evaluated as 0.027% at the top bent surface by finite element analysis—much lower than the fracture stress level of any materials used in M3D-integrated AI processors (Fig. 3c,d). This ensures high device yields following 3D integration stacking processes. We confirmed that

the performances of 2D material-based memristors and transistors were maintained even after experiencing such a small bending radius (Fig. 3e and Supplementary Fig. 15). Their mechanical robustness was measured by repetitive bending test in up to 100 cycles (Fig. 3f) and, due to their extremely thin nature and low stiffness, critical adhesion energy substantially reduces, allowing multiple stacking of M3D-integrated AI processors.

The high yield of M3D-integrated device circuitry was further evaluated by performing computing tasks utilizing our 2D-based AI processors. First we demonstrated DNA motif discovery by one-dimensional (1D) convolution, as shown in Fig. 4a. The purpose of DNA motif discovery is to identify a particular DNA sequence within an input DNA sequence. As a 1D kernel, the target DNA pattern can be made in any way desired and the motif 'ATGC' was used as an example. The 1D kernel pattern was created by one-hot encoding and programming of four DNA bases (A, T, G and C) into memristor crossbar arrays, as shown in Fig. 4b. One-hot encoding is among techniques commonly used in machine learning and data processing, especially when utilized for categorical data such as DNA motifs. Using one-hot encoding, each DNA base in the input DNA sequences was converted to a 1×4 input (1×16 inputs per kernel in this example (four DNAs based per kernel)). There are three possible scenarios for discovering DNA motifs, as shown in Fig. 4a: (1) unmatched, (2) partially matched and (3) fully matched. The multiply–accumulate (MAC) operation produces maximum current when input sequence is fully matched to a 1D kernel. Figure 4c shows the flow chart of DNA motif discovery by 1D convolution. The relative current amplitudes of a 16-row/one-column memristor array are used to detect a target genetic sequence in a randomly produced genetic sequence with the one-hot encoding approach when the target genetic sequence is detected (target sequence in this example is ATGC) (Fig. 4d). In this experiment, cumulative currents from memristor arrays (MAC values) are normalized between 0 and 4 and a software simulation displays the number of genetic sequences that fit ATGC. In Fig. 4d the MAC values from memristor arrays and software simulation are shown as matching. Because the results from software simulation represent the ground truth of DNA motif discovery, the more similar the MAC values to software simulation results the more accurate the computing performance of the M3D-integrated AI device becomes. It has been discovered that the genetic sequence ATGC is detected using 1D convolution programmed in memristor arrays, which results in the maximum cumulative current value. This eventually matches values obtained from software simulations of the sequence. In the fourth sweep the input sequence perfectly matched the 1D convolutional kernel (ATGC), resulting in maximum current in both software-simulated results and MAC values from the 2D memristor crossbar (Fig. 4d and Supplementary Fig. 16). More information on our measurement set-up and kernel operations for crossbar arrays can be found in Methods. More importantly, we investigated the improvement in performance yielded by the fully M3D-integrated AI system. The effect of multistack was demonstrated by the excellent performance of vertically stacked memristor arrays in the following areas: (1) processing time, (2) device footprint and (3) voltage drop, among others. The stacked memristor arrays were used to achieve this acceleration, which was accomplished through parallel processing. In the present emerging AI environment, a large quantity of data must be handled at the same time; our fully M3D-integrated AI system will provide excellent parallel computing capability while also reducing latency. Figure 4e–g illustrates how these allow for a small footprint yet provide high performance. Using the same number of devices we investigated voltage drop as a function of the number of layers, as shown in Fig. 4e. Voltage drops across computing devices decrease as the numbers of multistacks of M3D-integrated devices increase, because the number of devices required for each row under a given number of total devices is reduced as a result of multistack, which makes it possible to reduce voltage drop across the array. It is worth noting that predicted processing time and latency also decreased as the number of vertically stacked memristor layers increased, by reduction in the length of routing paths, word lines and bit lines (Fig. 4f,g). To estimate voltage drop and latency as a function of the number of layers, simulations were conducted based on the measurement data obtained from our fabricated devices. As mentioned above, multistack reduces the number of devices per single layer when the total number of devices is constant, and allows multiple layers to handle multiple inputs in parallel. For upscaling of the device array along the same row/column, in the simulation part the calculation is based on a straightforward assumption that both voltage drop and latency are linearly proportional to the number of devices in a single layer. The results we have obtained thus far are highly promising, indicating that the M3D integration of our AI processor substantially enhances its computing performance.

In conclusion, we successfully demonstrated 2D material-based M3D-integrated electronics, leveraging the fabricated WSe_2/*h*-BN-based memristors and MoS_2-based transistors with excellent performance. The M3D integration of each layer was experimentally demonstrated to verify reliable and uniform operation of AI processors. Owing to the extremely low stiffness and internal stress of 2D materials, we successfully realized multistack of the M3D-integrated devices—a total of six layers. The multistack of M3D-integrated AI processing layers was also verified by improved latency, voltage drop and footprint. M3D integration allows integration of different functional layers with high density and reduced surface area. It can handle large volumes of data from different sensors with high bandwidth and low latency. It is expected that this sensor fusion approach can reduce errors and improve accuracy by providing redundant and complementary sensing information. The combination of M3D integration and near/in-sensor computing architectures enables power-efficient edge computing solutions. Furthermore, because M3D-integrated devices on an ultrathin flexible PI substrate exhibited outstanding mechanical properties, these can be utilized in next-generation wearable AI platforms for various applications including real-time health and fitness monitoring, personalized medicine, emotional and cognitive monitoring, augmented and virtual reality interaction and even soft robotics. We envision that the proposed M3D integration strategy, based on 2D materials, will bring considerable innovation in integrated chip applications and lead to the next generation of integration.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41563-023-01704-z.

References

1. Saleh, R. et al. System-on-chip: reuse and integration. *Proc. IEEE* **94**, 1050–1069 (2006).
2. Wolf, W., Jerraya, A. A. & Martin, G. Multiprocessor system-on-chip (MPSoC) technology. *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.* **27**, 1701–1713 (2008).
3. Patti, R. S. Three-dimensional integrated circuits and the future of system-on-chip designs. *Proc. IEEE* **94**, 1214–1224 (2006).
4. Shulaker, M. M. et al. Three-dimensional integration of nanotechnologies for computing and data storage on a single chip. *Nature* **547**, 74–78 (2017).
5. Choi, C. et al. Reconfigurable heterogeneous integration using stackable chips with embedded artificial intelligence. *Nat. Electron.* **5**, 386–393 (2022).
6. Lin, P. et al. Three-dimensional memristor circuits as complex neural networks. *Nat. Electron.* **3**, 225–232 (2020).
7. Zhou, F. et al. Optoelectronic resistive random access memory for neuromorphic vision sensors. *Nat. Nanotechnol.* **14**, 776–782 (2019).

8. Mennel, L. et al. Ultrafast machine vision with 2D material neural network image sensors. *Nature* **579**, 62–66 (2020).
9. Tu, K.-N. Reliability challenges in 3D IC packaging technology. *Microelectron. Reliab.* **51**, 517–523 (2011).
10. Lau, J. H. Evolution, challenge, and outlook of TSV, 3D IC integration and 3D silicon integration. In *2011 International Symposium on Advanced Packaging Materials (APM)*, 462–488 (IEEE, 2011).
11. Shulaker, M. M. et al. Monolithic 3D integration: a path from concept to reality. In *2015 Design, Automation & Test in Europe Conference & Exhibition (DATE)* 1197–1202 (IEEE, 2015).
12. Wong, S. et al. Monolithic 3D integrated circuits. In *2007 International Symposium on VLSI Technology, Systems and Applications (VLSI-TSA)* 1–4 (IEEE, 2007).
13. Bishop, M. D., Wong, H.-S. P., Mitra, S. & Shulaker, M. M. Monolithic 3-D integration. *IEEE Micro* **39**, 16–27 (2019).
14. Liu, Y., Huang, Y. & Duan, X. Van der Waals integration before and beyond two-dimensional materials. *Nature* **567**, 323–333 (2019).
15. Geim, A. K. & Grigorieva, I. V. Van der Waals heterostructures. *Nature* **499**, 419–425 (2013).
16. Novoselov, K. S., Mishchenko, A., Carvalho, O. A. & Castro Neto, A. 2D materials and van der Waals heterostructures. *Science* **353**, aac9439 (2016).
17. Kim, K. S. et al. Non-epitaxial single-crystal 2D material growth by geometric confinement. *Nature* **614**, 88–94 (2023).
18. Li, M. et al. Imperfection-enabled memristive switching in van der Waals materials. *Nat. Electron.* **6**, 491–505 (2023).
19. Yang, S., Jiang, C. & Wei, S.-H. Gas sensing in 2D materials. *Appl. Phys. Rev.* **4**, 021304 (2017).
20. Long, M., Wang, P., Fang, H. & Hu, W. Progress, challenges, and opportunities for 2D material based photodetectors. *Adv. Funct. Mater.* **29**, 1803807 (2019).
21. Pang, Y., Yang, Z., Yang, Y. & Ren, T. L. Wearable electronics based on 2D materials for human physiological information detection. *Small* **16**, 1901124 (2020).
22. Li, T. et al. Reconfigurable, non-volatile neuromorphic photovoltaics. *Nat. Nanotechnol*. https://doi.org/10.1038/s41565-023-01446-8 (2023).
23. Song, M.-K. et al. Recent advances and future prospects for memristive materials, devices, and systems. *ACS Nano* **17**, 11994–12039 (2023).
24. Shim, J. et al. Controlled crack propagation for atomic precision handling of wafer-scale two-dimensional materials. *Science* **362**, 665–670 (2018).
25. Jia, X. et al. High-performance CMOS inverter array with monolithic 3D architecture based on CVD-grown n-MoS_2 and p-$MoTe_2$. *Small* https://doi.org/10.1002/smll.202207927 (2023).
26. Jayachandran, D. et al. Monolithic three-dimensional (3D) integration of two-dimensional (2D) field effect transistors. Preprint at https://doi.org/10.21203/rs.3.rs-2512945/v1 (2023).
27. Guan, S.-X. et al. Monolithic 3D integration of back-end compatible 2D material FET on Si FinFET. *NPJ 2D Mater. Appl.* **7**, 9 (2023).
28. Wang, C.-H. et al. 3D monolithic stacked 1T1R cells using monolayer MoS_2 FET and hBN RRAM fabricated at low (150 °C) temperature. In *2018 IEEE International Electron Devices Meeting (IEDM)* 22.25.21–22.25.24 (IEEE, 2018).
29. Tang, B. et al. Wafer-scale solution-processed 2D material analog resistive memory array for memory-based computing. *Nat. Commun.* **13**, 3037 (2022).
30. Kumar, D., Aluguri, R., Chand, U. & Tseng, T.-Y. Enhancement of resistive switching properties in nitride based CBRAM device by inserting an Al_2O_3 thin layer. *Appl. Phys. Lett.* **110**, 203102 (2017).
31. Sun, T. et al. Stable resistive switching in ZnO/PVA: MoS_2 nilayer memristor. *Nanomaterials* **12**, 1977 (2022).
32. Tsai, T.-L., Jiang, F.-S., Ho, C.-H., Lin, C.-H. & Tseng, T.-Y. Enhanced properties in conductive-bridge resistive switching memory with oxide-nitride bilayer structure. *IEEE Electron Device. Lett.* **37**, 1284–1287 (2016).
33. Wu, F. et al. Interface engineering via MoS_2 insertion layer for improving resistive switching of conductive-bridging random access memory. *Adv. Electron. Mater.* **5**, 1800747 (2019).
34. Shen, P.-C. et al. Ultralow contact resistance between semimetal and monolayer semiconductors. *Nature* **593**, 211–217 (2021).
35. Das, S. et al. Transistors based on two-dimensional materials for future integrated circuits. *Nat. Electron.* **4**, 786–799 (2021).
36. Wang, Y. et al. Van der Waals contacts between three-dimensional metals and two-dimensional semiconductors. *Nature* **568**, 70–74 (2019).
37. Kim, T.-H. et al. Multilevel switching memristor by compliance current adjustment for off-chip training of neuromorphic system. *Chaos Solitons Fractals* **153**, 111587 (2021).
38. Wan, H. et al. In situ observation of compliance-current overshoot and its effect on resistive switching. *IEEE Electron Device. Lett.* **31**, 246–248 (2010).
39. Zhu, J. et al. Low-thermal-budget synthesis of monolayer molybdenum disulfide for silicon back-end-of-line integration on a 200 mm platform. *Nat. Nanotechnol.* **18**, 456–463 (2023).

[1]Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, USA. [3]Department of Electronic Engineering, Inha University, Incheon, Republic of Korea. [4]School of Electrical and Electronic Engineering, Yonsei University, Seoul, Republic of Korea. [5]Department of Mechanical Engineering and Materials Science, Washington University in Saint Louis, Saint Louis, MO, USA. [6]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA. [7]Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN, USA. [8]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA. [9]CNRS, Georgia Tech – CNRS IRL 2958, GT-Europe, Metz, France. [10]Institute of Materials Science and Engineering, Washington University in Saint Louis, Saint Louis, MO, USA. [11]Future Innovation Research Center, Korea Aerospace Research Institute, Daejeon, Republic of Korea. [12]Aerospace System Engineering, University of Science and Technology, Daejeon, Republic of Korea. [13]Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. [14]Department of Industrial Engineering, Jeonju University, Jeonju, Republic of Korea. [15]Convergence Institute of Human Data Technology, Jeonju University, Jeonju, Republic of Korea. [16]Post-Silicon Semiconductor Institute, Korea Institute of Science and Technology, Seoul, Republic of Korea. [17]Present address: College of Computer Science and Technology, Zhejiang University, Hangzhou, China. [18]These authors contributed equally: Ji-Hoon Kang, Heechang Shin, Ki Seok Kim, Min-Kyu Song. ✉e-mail: ahnj@yonsei.ac.kr; jeehwan@mit.edu; sbae22@wustl.edu

Methods

Growth of MoS_2

Bilayer MoS_2 thin film was grown using the MOCVD system. A thermally grown, 300-nm-thick SiO_2 on a 4 inch Si wafer was placed in a quartz tube following cleaning with acetone, isopropanol and water. Molybdenum hexacarbonyl (Sigma-Aldrich, purity ≥99.9%) and anhydrous dimethyl sulfide (Sigma-Aldrich, purity ≥99.0%) were used as precursors for Mo and S, respectively, and were introduced into the quartz tube using 1.0 standard cubic centimetres min^{-1} (sccm) for H_2 and 310 sccm for Ar as carrier gases. The injected amounts of precursor were precisely controlled by mass flow controllers, with NaCl plates loaded upstream of the furnace. Optimized conditions for synthesis of bilayer MoS_2 included pressure of 10.0 Torr, growth temperature of 580 °C, growth time of 22 h, molybdenum hexacarbonyl flow of 1.0 sccm and dimethyl sulfide flow of 0.6 sccm.

Growth of WSe_2

Growth of WSe_2 was performed using a seed layer growth method on a 2 inch, single-side-polished sapphire (0001) wafer using molecular beam epitaxy. The sapphire was first annealed at 900 °C for 60 min and then cooled to 20 °C. The growth process started with deposition of a seed layer equivalent to 0.5-monolayer-thick WSe_2 at 20 °C, by codeposition of tungsten (W) and evaporation using a multipocket e-beam evaporator and elemental selenium (Se) with a cracker source. A tungsten:selenium flux ratio of 1:200 was used with a W flux of ~5×10^{-9} mbar and Se flux of 1×10^{-6} mbar. After deposition of the seed layer, the sample was then annealed at 900 °C under a Se background only for 1 h and then ramped down to the growth temperature of 600 °C at 0 °C min^{-1}. The main WSe_2 was then grown using the same W and Se fluxes as for the seed layer, at a growth rate of 1 layer 5 h^{-1}. Total growth time was 25 h, corresponding to a thickness of ~3 nm.

Growth of *h*-BN

Growth of 3 nm *h*-BN was performed in an Aixtron MOVPE close-coupled showerhead reactor on the 2 inch sapphire substrate at 1,280 °C and 90 mbar pressure. Triethylboron and ammonia (NH_3) were used as B and N precursors, respectively, and hydrogen was used as carrier gas; growth rate was 15 nm h^{-1}. Further information on specific growth conditions can be found in ref. 40.

Fabrication and characterization of MoS_2 TFT

A 10-nm-thick Al_2O_3 layer was deposited on a 300-nm-thick SiO_2 wafer as a planarization layer using atomic layer deposition (ALD). Source and drain electrodes (Cr/Au, 3/30 nm) were patterned on the Al_2O_3/SiO_2 wafer using photolithography, with channel dimensions designed as 750 and 5 µm for width and length, respectively. The bilayer MoS_2 film was transferred onto the wafer and patterned as an isolated channel via reactive ion etching using CHF_3/O_2 plasma. Subsequently, a 30-nm-thick Al_2O_3 gate dielectric layer was deposited on MoS_2.

To inhibit trapping of H_2O molecules and improve the interface between Al_2O_3 and MoS_2, the device was annealed overnight at 120 °C under vacuum conditions. The top-gate electrode (Cr/Au, 3/30 nm) was formed using photolithography and a lift-off process. MoS_2 TFT was characterized using a source measure unit (Keithley 4200 SCS parameter analyser, Keithley Instruments, Inc.).

The MoS_2 film grown using the hot-wall MOCVD system was transferred to the bottom-contact source and drain (S/D) structure, which is designed for high drain current (channel width/channel length = 150) (Supplementary Note 2 and Supplementary Figs. 17 and 18). Subsequently, a top-gate dielectric Al_2O_3 layer was deposited on the MoS_2 film. The Al_2O_3 layer involved in S/D electrodes can effectively reduce contact resistance due to screening effects from the high-*k* top dielectric overlayer (*k* = ~7.2 in ALD Al_2O_3), resulting in decoupling of the metal–semiconductor interaction. Moreover, the top-gate insulator layer enables n-type doping of the MoS_2 layer due to its oxygen-deficient surface and interfacial oxygen vacancies, lowering the conduction band edge below Fermi level[41]. These effects allow charge carriers to fill lower-conduction band states at the semiconductor–insulator interface, resulting in n-type carrier injection in MoS_2. N-type doped MoS_2 in the channel and contact regions resulted in increased electron concentration at the interface of S/D contact regions, reducing Schottky barrier width and thus decreasing contact resistance[42]. Reduction in contact resistance enhances both carrier mobility and high on-current level, which facilitate matching of the current with memristor operation for seamless monolithic integration. The transfer characteristics (I_D–V_G) of bilayer MoS_2 TFTs exhibited μ_{FE} of 10.42 ± 2.1 cm^2 Vs^{-1}, an on:off ratio of 10^9, a subthreshold swing of 0.612 ± 0.05 V dec^{-1} and V_{TH} of 1.7 ± 0.8 V under a drain voltage of 1 V (Fig. 1e and Supplementary Figs. 8 and 9). In particular, V_{TH} of the TFTs indicates a normally off operation, which enables lower power consumption following integration with WSe_2/*h*-BN-based memristors. In particular, the output characteristics (I_D–V_D) of the MoS_2 TFT indicate that drain current reached 10 mA at a gate voltage of 9.0 V with ohmic behaviour (Fig. 1f). Furthermore, histograms of the evaluated μ_{FE} and V_{TH} of the MoS_2 transistor array represent uniform characteristics with average values of 10.17 cm^2 Vs^{-1} and 0.87 V, respectively (Fig. 1g).

3D monolithic integration

Following fabrication of bilayer MoS_2 transistors, a 50-nm-thick Al_2O_3 layer was then deposited by ALD as a passivation and insulation layer between the TFTs and WSe_2/*h*-BN-based memristors. For construction of the interconnections between the drain electrode of MoS_2 TFTs and the bottom electrode of the WSe_2/*h*-BN-based memristor, via-holes were patterned and dry-etched by reactive ion gas plasma (CHF_3/O_2, 40/10 sccm) with precise control of etching rate (3.7 Å s^{-1}) followed, after a few seconds, by treatment by buffered oxide etchant to create a clean interface. Interconnections (Cu/Au, 10/50 nm) were made through the via-holes with bottom-electrode lines (Ti/Pt, 5/30 nm) for the WSe_2/*h*-BN-based memristor. Electrical connections through interconnections were identified by probe station. The integrated circuit was placed in the central region (7 × 7 mm^2) connected with four outer source lines. Gate, source and drain lines were used for driving MoS_2 transistors, the drain line was simultaneously operated as a bottom electrode and a further top-electrode line for WSe_2/*h*-BN-based memristors was constructed. Following fabrication of the first layer of the integrated arrays, the second and third were peeled off precisely. Stacking of the second and third layers of the integrated arrays was controlled by an alignment mark from the first layer, sharing four outer source lines soldered with Ag paste. We designed the upper layers to be smaller than those underneath, to expose contact pads on the edges of each layer for convenience. By stacking a smaller layer on top we could maintain exposure of the contact pads on the lower layers.

Implementation of 1D kernel motif on memristor crossbar array for DNA motif discovery

The 1D kernel motif used in this study consisted of four DNA bases. First, each DNA base (A, T, G and C) was one-hot encoded as shown in Fig. 4b. Because each one-hot-encoded DNA base consists of four digits, 4 × 1 memristor cells are required for one DNA base and thus the 1D kernel motif with four DNA bases was programmed on 16 × 1 memristor cells using a pulse generator unit (Keysight, 33622 A) and a digital storage oscilloscope (Keysight, DSOX3024T).

Electrical measurement set-up for DNA motif discovery

For the measurement of crossbar arrays, a pulse generator unit was used to generate input pulse strain and a digital storage oscilloscope was used to measure output currents calculated by MAC operation. Memristors were programmed to the binary state for DNA motif discovery, to minimize errors from interdevice variability. This was further improved by optimization of array programming into multistate conductance.

As shown in Fig. 4a,b, four input DNA bases were selected from the input sequence and one-hot encoded. Next, 16 one-hot encoded inputs (four inputs × four DNA bases) were applied as voltage pulses by the pulse generator unit to the 1D kernel motif implemented on the crossbar array. The current measured in each cell in the memristor array is equal to the product of input voltage and cell conductance by Ohm's law, and the accumulation of current in the output is equal to the sum of the currents measured in each cell by Kirchhoff's current law. The sum of output currents calculated by MAC operation was measured by digital storage oscilloscope. Having determined the measured value, the next four DNA bases were selected from the input sequence and the same process repeated until the end of the input sequence.

Data availability

All data are available in the main text or Supplementary Information. All relevant data are available from the corresponding authors upon reasonable request.

References

40. Li, X. et al. Large-area two-dimensional layered hexagonal boron nitride grown on sapphire by metalorganic vapor phase epitaxy. *Cryst. Growth Des.* **16**, 3409–3415 (2016).
41. Wang, Y. & Chhowalla, M. Making clean electrical contacts on 2D transition metal dichalcogenides. *Nat. Rev. Phys.* **4**, 101–112 (2022).
42. Choi, M. et al. Flexible active-matrix organic light-emitting diode display enabled by MoS_2 thin-film transistor. *Sci. Adv.* **4**, eaas8721 (2018).

Acknowledgements

The team at Massachusetts Institute of Technology acknowledges support from the Korea Institute of Science and Technology (nos. 2E32260 and 2E32242). J.-H.A. acknowledges support from the National Research Foundation of Korea (no. NRF-2015R1A3A2066337). C.H. acknowledges support from the National Science Foundation (no. DMR-1921818) and SUPREME, one of seven centres in JUMP 2.0, a Semiconductor Research Corporation programme sponsored by Defense Advanced Research Projects Agency (DARPA). S.-H.B. acknowledges financial support from Washington University in St. Louis and the institute of Materials Science and Engineering for the use of instruments and staff assistance. S.-H.B. also acknowledges that this work was partially supported by Samsung Electronics Co., Ltd. (IO221219-04250-01). This work was carried out in part through the use of MIT.nano's facilities. The authors would like to acknowledge Dr. Baoming Wang for assistance in focused ion beam (FIB) sample preparation. M.-K.S. acknowledges support from the National Research Foundation of Korea (no. NRF-2021R1A6A3A14044297). A.O. acknowledges financial support from Georgia Tech Europe in Metz-France.

Author contributions

J.K. and S.-H.B. conceived the idea and led the research. J.-H.K., S.-H.B. and J.K. designed experiments. J.-H.K., H.S., K.S.K., M.-K.S., J.-H.A., S.-H.B. and J.K. prepared the manuscript. H.S., A.T.H., K.S.K., D.L., R.Y. and G.Z. grew 2D thin films. P.V., S. Sundaram and A.O. worked on the growth of *h*-BN on sapphire. J.-H.K., H.S. and M.-K.S. performed device fabrication. H.S., J.-H.K., Y.M. and C.C. designed and conducted computing simulations. B.K., H.K., J.C., B.-I.P., J.S., J.S.K., S.H., Sangho Lee, B.K., Seungju Seo and Seunghwan Seo conducted film transfer and characterization. K.R. and E.P. conducted STEM measurements. D.L., Y.M., J.M.S., B.J.K., S.L., S.O.K., S.M., M.-C.P., S.L., H.-J.K., G.Z., S. Sundaram, A.T.H., Z.X., R.Y., H.A., H.S.K., P.L., C.H., A.O. and J.-H.A. provided feedback throughout the experiments and data analysis. The manuscript was written by J.-H.K., S.-H.B. and J.K. with input from all authors. All authors contributed to the analysis and discussion of the results leading to 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/s41563-023-01704-z.

Correspondence and requests for materials should be addressed to Jong-Hyun Ahn, Jeehwan Kim or Sang-Hoon Bae.

Peer review information *Nature Materials* thanks Weida Hu, Tianyou Zhai and Ilia Valov for their contribution to the peer review of this work.