Supplementary Information of

Physically unclonable memristor-based compute-in-memory chip for secure AI

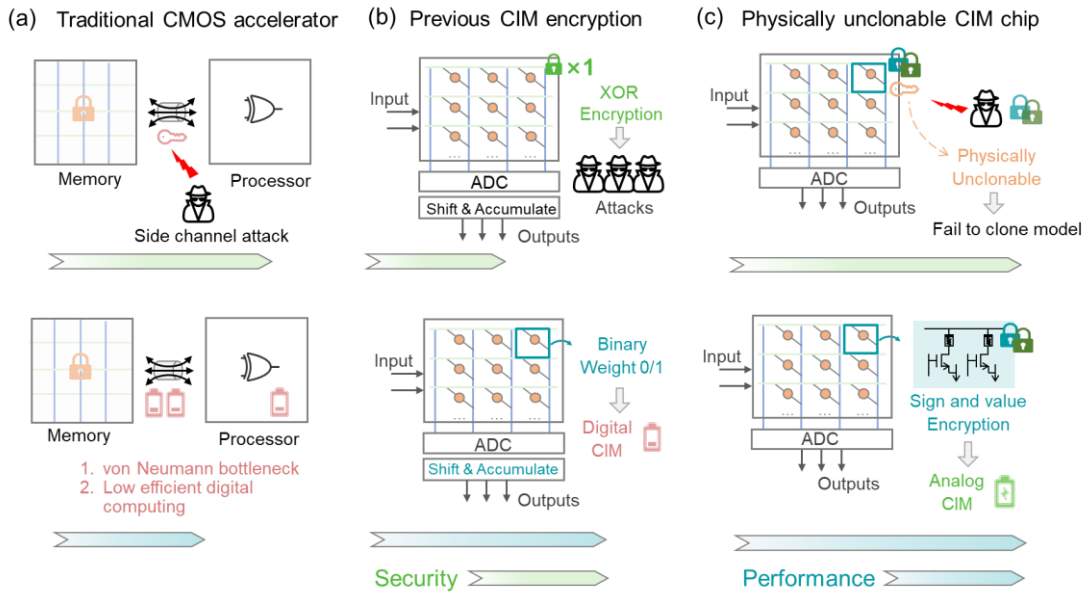
Yiyang Chen[1,2], Lixia Han[1,2,3], Ao Shi[1,2], Lianliang Wu[1,2], Hairuo Lu[1,2], Kexun Li[1,2], Haozhang Yang[1,2], Jiaqi Li[1,2], Zheng Zhou[1,2], Lifeng Liu[1,2*], Jinfeng Kang[1,2] and Peng Huang[1,2*]

[1]School of Integrated Circuits, Peking University, Beijing 100871, China;

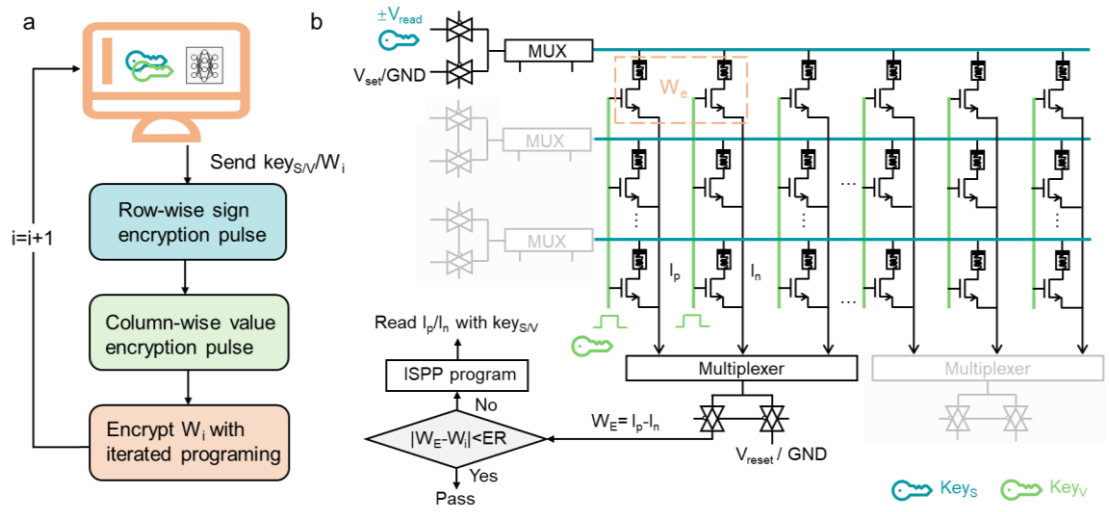
[2]Beijing Advanced Innovation Center for Integrated Circuits, Beijing 100871, China;

[3]School of Integrated Circuits, Nanjing University of Aeronautics and Astronautics, Nanjing Jiangsu, 211106, China;

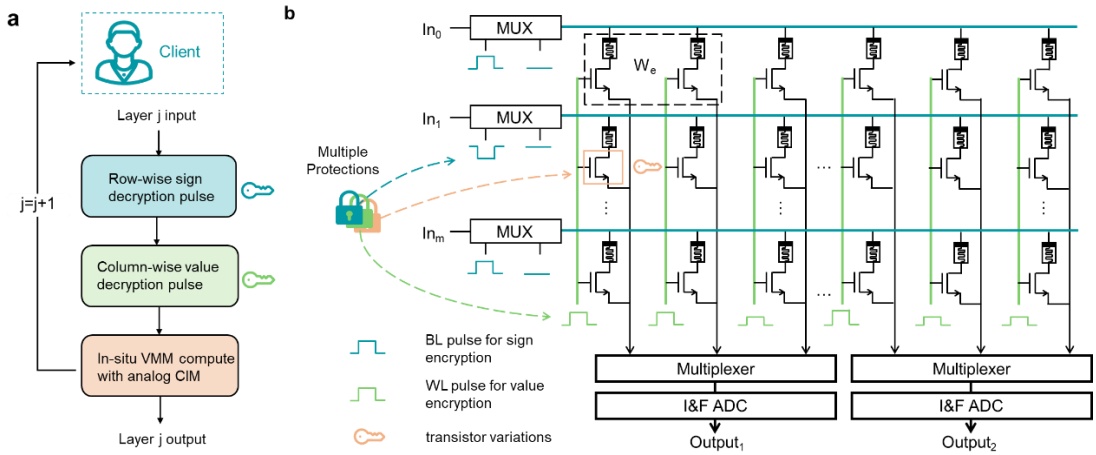
[*]Emails: lfliu@pku.edu.cn; phwang@pku.edu.cn



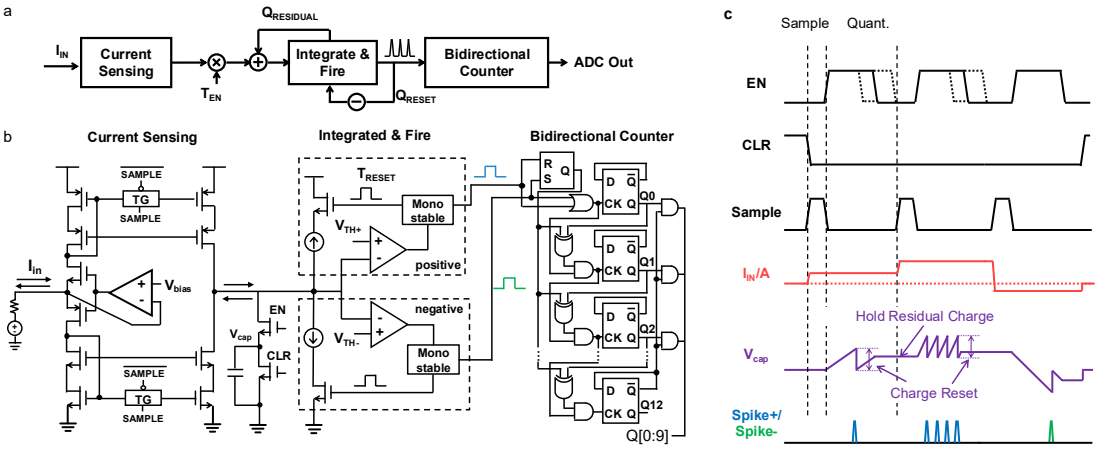
Supplementary Figure 1. The comparison of previous accelerators and physically unclonable CIM chip. a, Traditional CMOS accelerators face challenges in terms of low energy efficiency caused by von Neumann bottleneck and side channel attacks. **b,** While prior CIM encryption accelerators trade security with energy efficiency by using digital in-memory XOR encryption for model protection. **c**, In this work, we propose sign and value encryption to protect model parameters while leveraging high energy efficiency analog CIM. Furthermore, to mitigate risks from side-channel attacks, we embed a physically unclonable analog key within the memristor array, thereby eliminating the need for key transmission during computation.



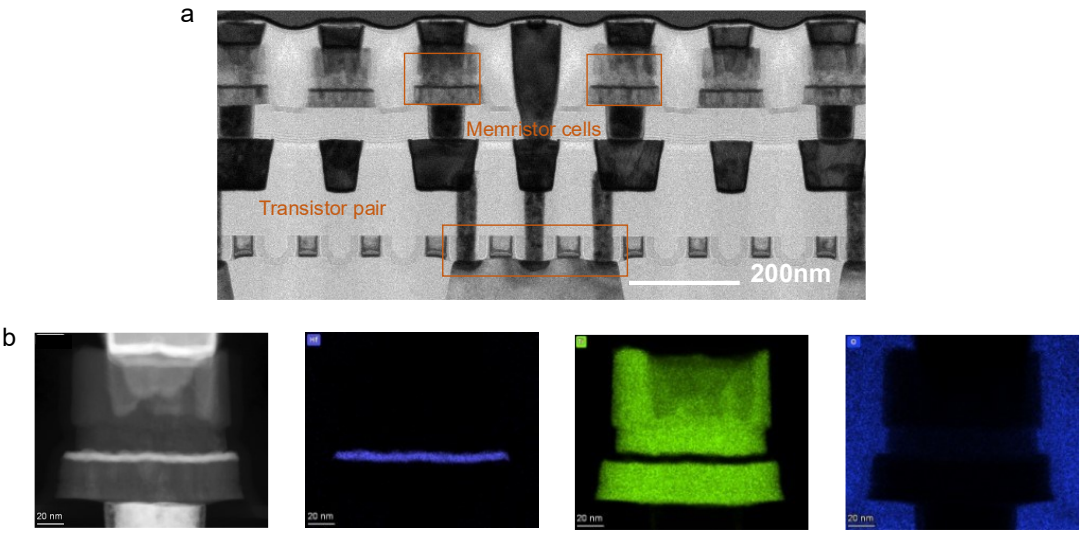
Supplementary Figure 2. The weight mapping process for sign and value encryption. a, Flow chat of sign and value encryption. To map the weight into differential cell (Two 1T1R structure), the PC transfers key$_S$ and key$_V$ with W_i to CIM chip. The row-wise and column-wise pulses are generated according to these keys. The W_i is programed with write-verify ISPP methods[1]. **b**, The schematic for W_i mapping with key$_S$ and key$_V$. During each iteration, the W_E is measured with pulses encoded by key$_S$ and key$_V$. We compare W_E with W_i to generate proper programing pulse for next iteration. The encrypted weights are stored in memristor devices to protect valuable AI models.



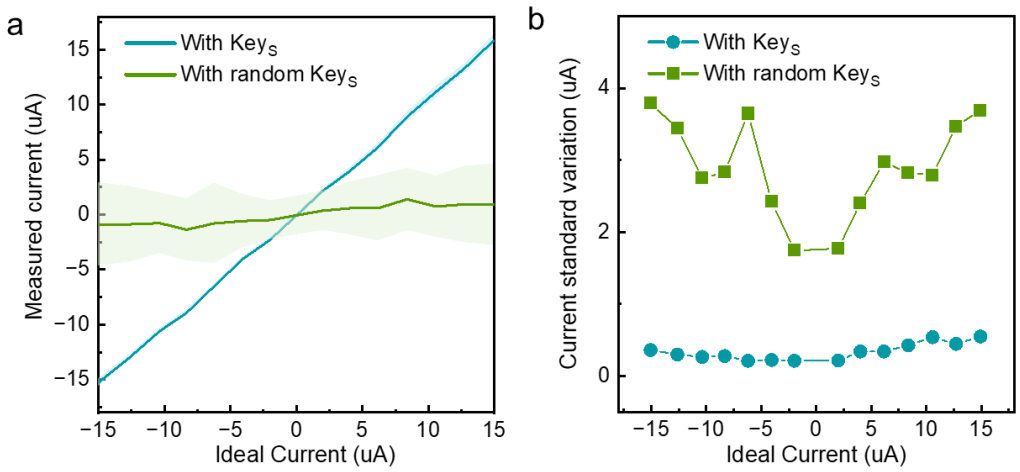
Supplementary Figure 3. The schematic of model decryption and in-situ vector matrix multiplication in the chip. a, The flow chart of model decryption process. For each layer i in neural network, the row-wise pulses for sign decryption and column-wise pulses for value decryption are generated in parallel to achieve in-situ model decryption and model computations. **b**, By configuring key_S as voltage pulses along bit lines (BLs) and key_V across word lines (WLs), our approach enables simultaneous sign decryption and value decryption for W_E. Moreover, the physically unclonable key is integrated in series with the memristor device, enabling parallel decryption without the need for external key transmission. In the meantime, to compute VMM in crossbar, the input vectors are encoded as bit-wise pulses in BL direction, which is identical to other compute in-memory accelerators. The accumulated outcome is obtained by subtracting positive and negative currents from source lines (SLs) that are quantified by on-chip I&F ADC. Our chip shows concurrent decryption with the above secure techniques, along with in-memory computation.



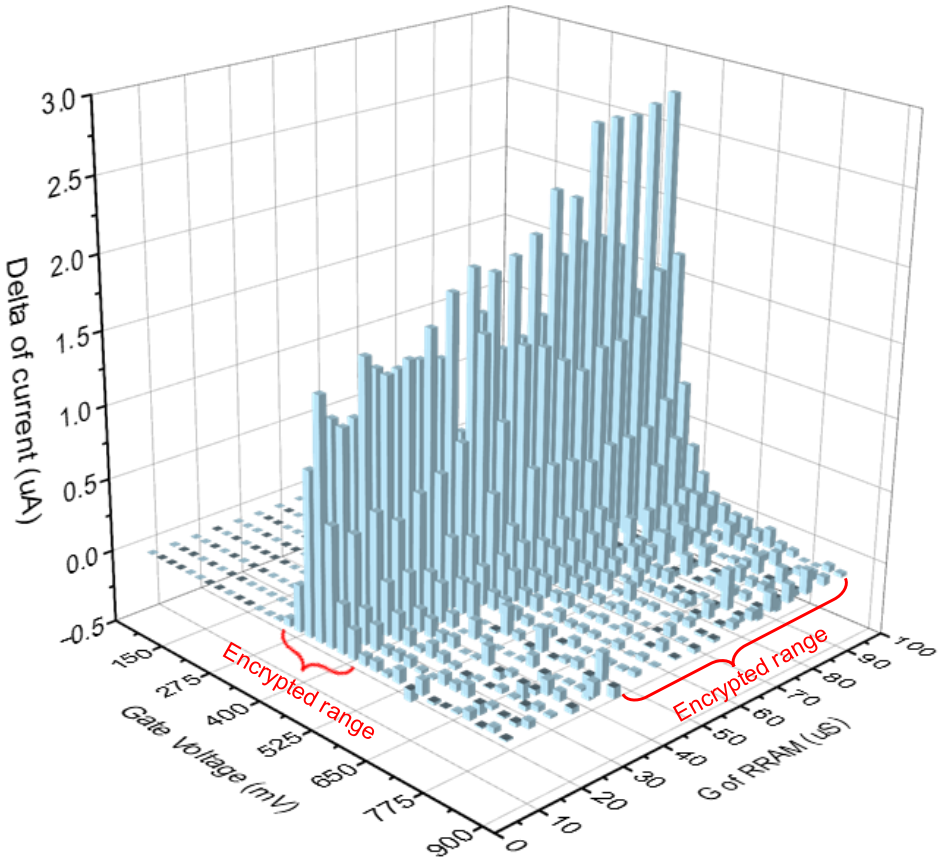
Supplementary Figure 4. Schematic of Integrated & Fire (I&F) ADC. a, The work principle of I&F ADC. The I_{IN} from memristor array is first sensed and accumulated by I&F units to generate pulses counting by a bidirectional counter. **b**, The I&F ADC consists of current sensing, integrated & fire, and bidirectional counter units. The bidirectional current mirror is adopted to linearly scale down bidirectional current from array. The charge accumulated in V_{cap} is quantized to bidirectional pulses with positive and negative threshold fire circuits. Current sources are utilized to reset the V_{cap} to improve ADC linearity. A bidirectional counter is designed to count pulse numbers from a positive or negative fire circuit. **c**, The timing diagram of I&F ADC. The V_{cap} can hold the residual charge during each reset operation to reduce quantization errors. The quantization precision can be modified with EN signal duration (Integrate time) to achieve better system level performance.



Supplementary Figure 5. The 1T1R cell and memristor device analysis results. a, Transmission electron microscopy (TEM) of 1T1R cell in 28nm CMOS technology. The memristor cells are fabricated in VIA of M1 layer. **b**, Energy dispersive spectroscopy (EDS) mapping results of the memristor device showing Hf (purple), Ti (green), and O (blue).

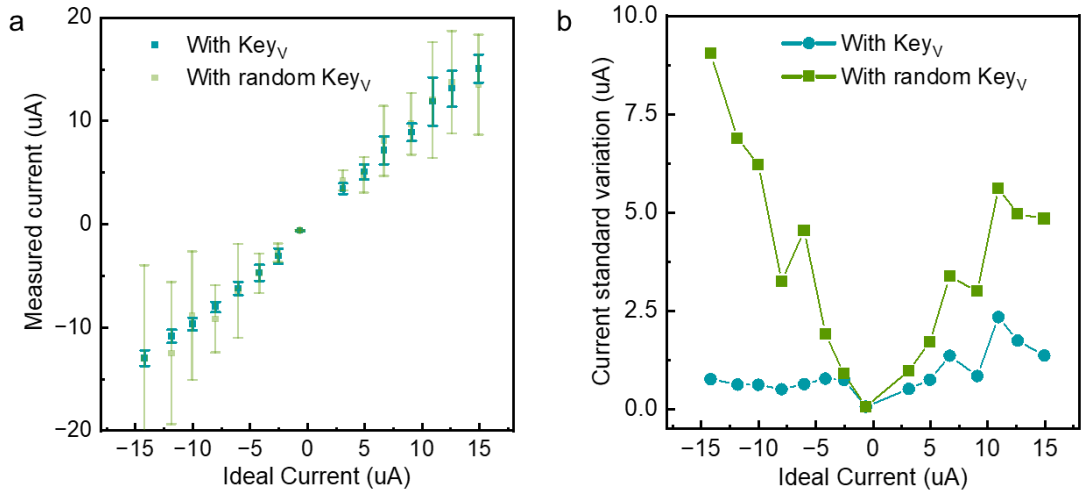


Supplementary Figure 6. a, The measured VMM computing results under the right and random key$_S$. **b**, The comparison of VMM errors under different output currents. Without the correct key$_S$ for sign encryption, the sign of decrypted weight is obfuscated by a one-bit key$_S$, leading to accumulated errors in the VMM computations.

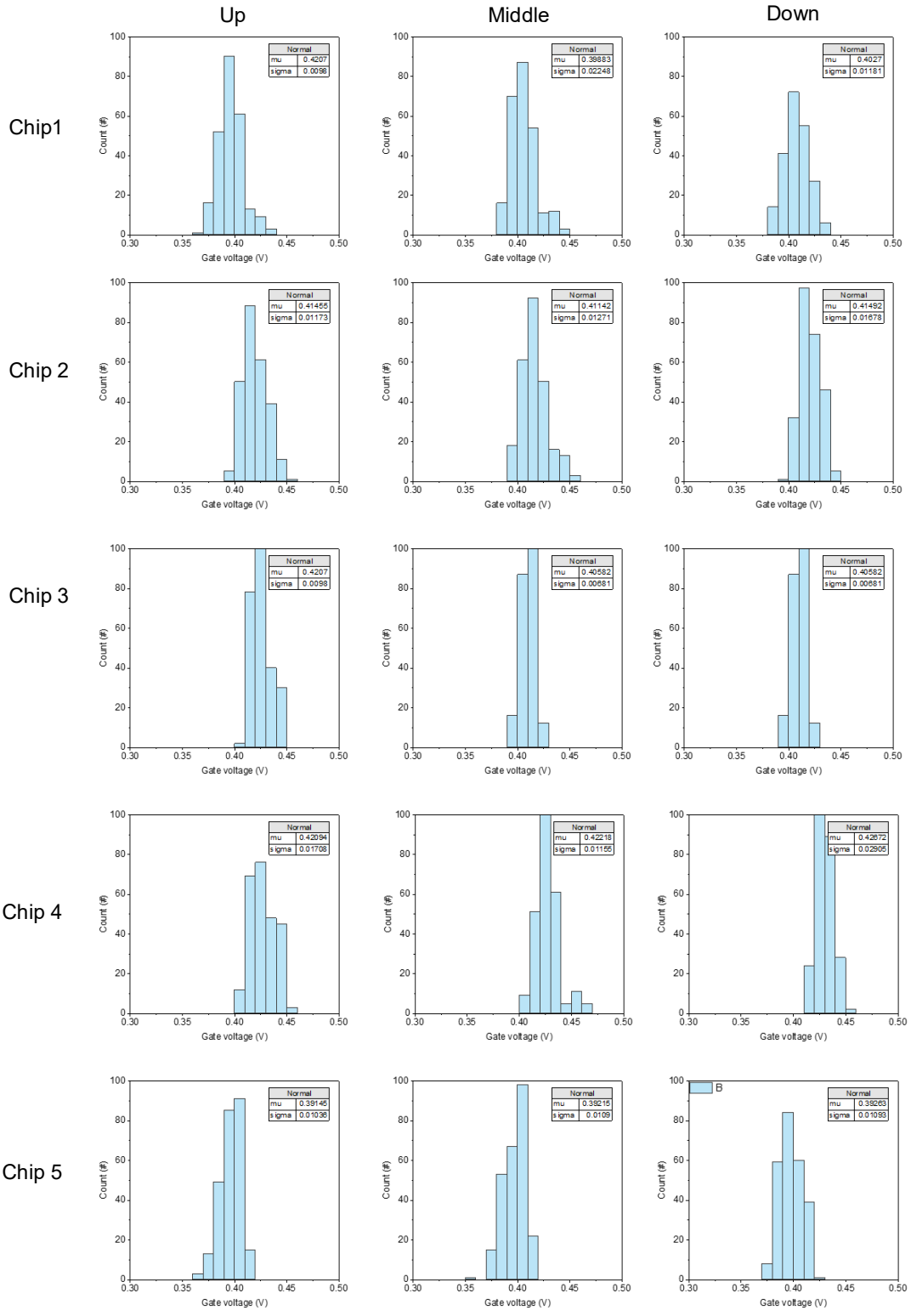


77

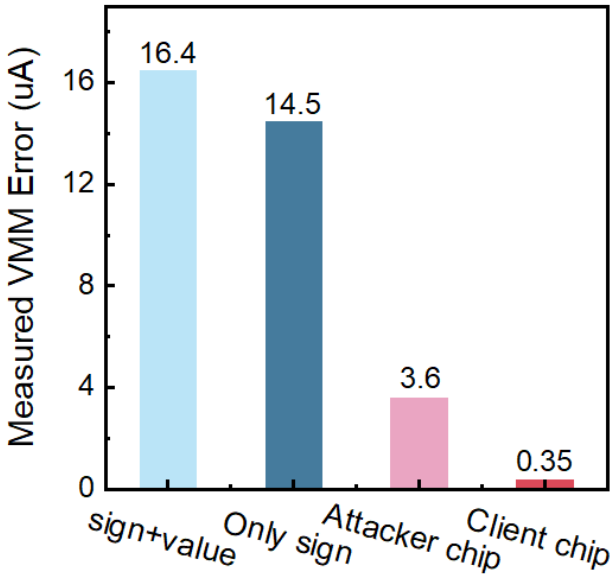
78 **Supplementary Figure 7** The hot map of ΔI under different conductance of memristor
79 and V_g. High ΔI range of V_g and memristor conductance is used to encrypt weight value.
80 Value encryption for weight protection is only effective when ΔI is sufficiently large.
81 Consequently, we restrict value encryption to regions by exhibiting large ΔI value.
82



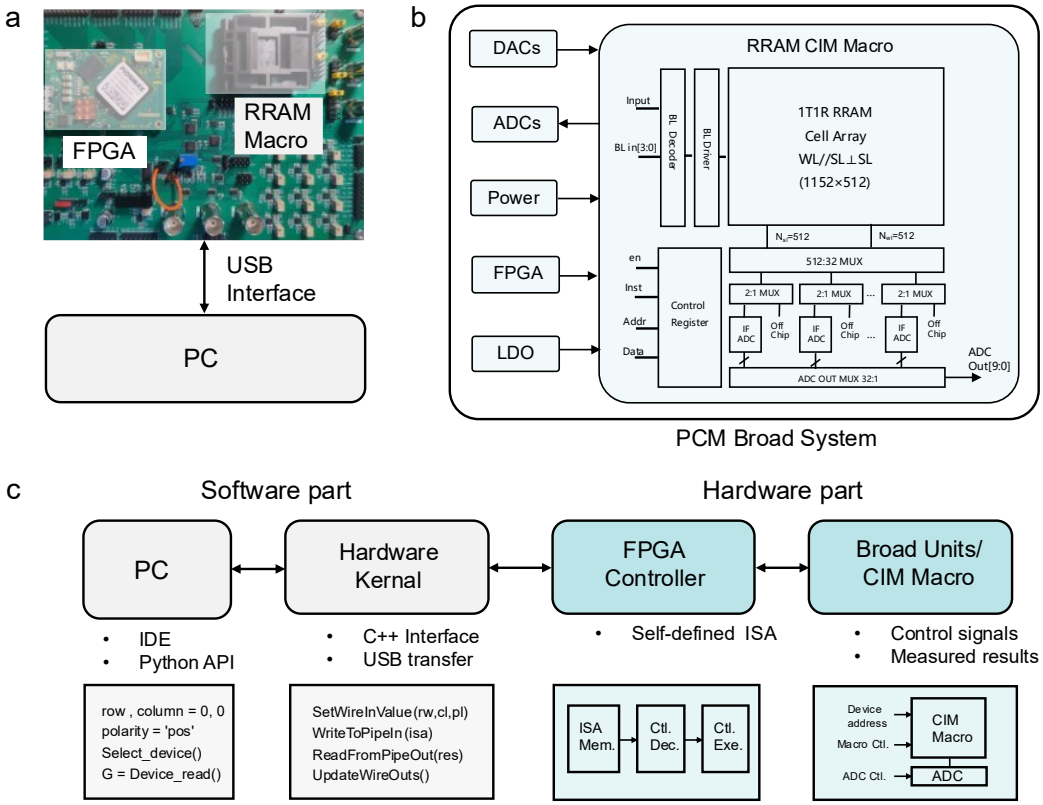
Supplementary Figure 8. a, The measured VMM computing results under the right and random key$_V$. **b**, The comparison of VMM errors under different output currents. With value encryption, the value of weight is obfuscated, and the accumulated VMM results show large errors caused by value encryption.



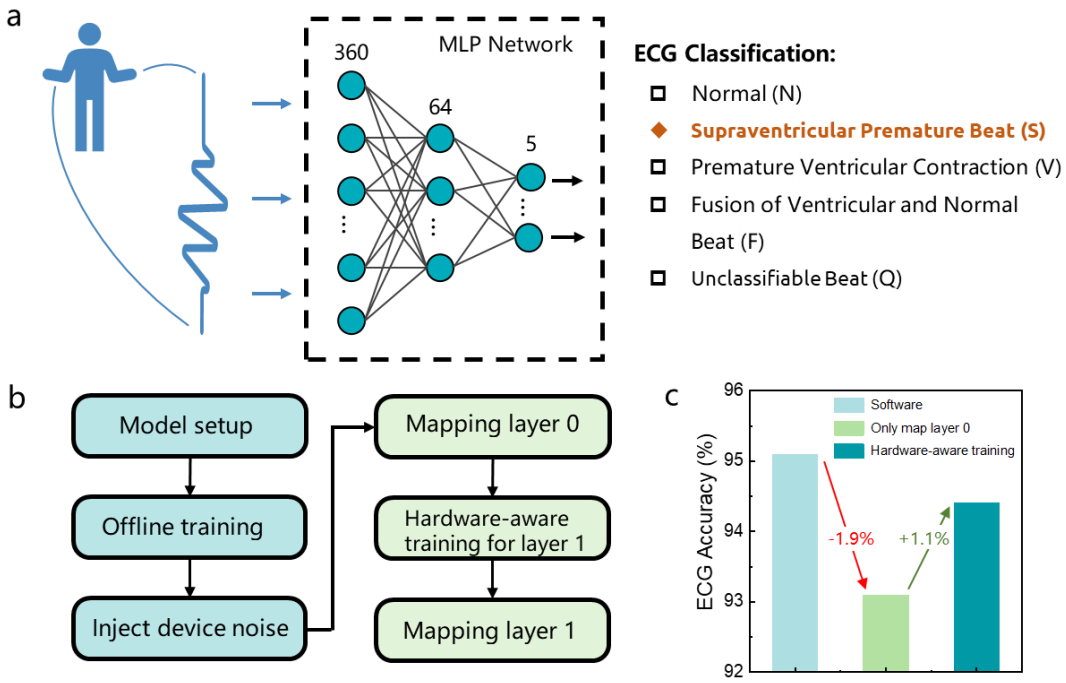
Supplementary Figure 9. The statistical results of threshold voltage distributions across different chips and different positions inside a chip. Each distribution is derived from measurements of 256 transistors. Physical variations arise due to inherent process fluctuations during semiconductor fabrication.



Supplementary Figure 10. The comparison of VMM computing errors in various conditions. Large VMM errors are observed without keys for sign and value encryption. Additionally, with the help of physically unclonable analog key, the attacker chip shows large computation errors caused by mismatched transistors.



Supplementary Figure 11. Demonstration system design for ECG task. a, The system consists of a custom printed circuit broad (PCB) with FPGA and PC. The PC communicates with PCB through USB interface. **b**, The architecture of custom PCB. All necessary units including DAC, ADC, and LDO are integrated for the demonstration. **c**. The software and hardware tool chain for demonstration system. The FPGA controller facilitates control and data flow through self-defined ISA.



Supplementary Figure 12. Neural network structure and training details for ECG demonstration. a, The ECG signals are sampled and subsequently fed into a two-layer Multilayer Perceptron (MLP) for classification into five categories. **b**, We employ offline training with device noise to train the model with robustness for device non-idealities. After that, we implement a layer-by-layer weight mapping scheme to facilitate hardware-aware online training. This approach could further improve task accuracy. **c**, Accuracy comparison with hardware-aware training. The accuracy loss due to device non-idealities can be effectively mitigated.

Supplementary Discussion 1. The Energy and Latency Estimation for ECG demonstration

To precisely estimate the energy and latency consumption for ECG demonstration, we experimentally measure the energy efficiency and latency of vector matrix multiplication (VMM) with the same setting of demonstration (See Supplementary Table I for more details). We measure latency by considering computing latency and data transfer latency. The 6-bit input vectors are fed into CIM chip in bit-serial fashion. Thus, the total computing latency is computed by:

$$Compute\ Latency = \sum_{i=1}^{6} T_{ADC}^{i} = (100 + 200) \times 3 = 900ns \quad (1)$$

For latency of data transfer, our CIM chip could operate maximum in 50MHz clock (20ns) for loading data to row register. During each cycle, 8-bit independent input bits can be stored in parallel. As a result, 8 cycles are required to send 64 bits data to registers. To sum up, the latency for VMM operation is conducted by: *VMM latency = Compute latency + Transfer latency=900+20ns ×6 ×8=1860ns.*The energy efficiency (EE) is measured by the average power of computing 10,000 randomized vectors with 10% sparsity. We adapt the integration time of I&F in the high and low 3 bits to enable balance energy and accuracy trade-off. During the VMM computation process, we activate 64 rows simultaneously. The 32 I&F ADCs in the CIM chip could operate in parallel to improve system throughput. Thus, the EE is deduced by: *EE=VMM Ops/(Average power ×total latency)=64 ×32 ×2/((5.44uA ×0.9V ×960ns +787.52uA × 1.8V ×900ns)= 3.20TOPS/W.* The VMM Ops means the total operation number in each VMM computing.

We estimate system-level performance based on measured results. We deploy two-layer MLP in the CIM chip to classify the ECG signals. The detailed neural network structure is shown in Supplementary Figure 11. As a result, the total operation number for inference is conducted by: *Total Ops=2 ×(Input_size ×hidden_size+Hidden_size × output_size)=46,720.* The rectified linear unit (ReLU) activation function is employed in the hidden layers of the neural network. Therefore, the energy cost of activation function is negligible because low cost of ReLU implementation in hardware. The energy cost of inference is conducted by: *Total energy = Total Ops/EE=14.6nJ.* Similar to energy cost, the total latency for inference is calculated as: *Total latency = Total Ops/VMM Ops ×VMM latency=22.78us.* The energy and latency overhead can be minimized through high-bandwidth on-chip data transfer or co-optimization of circuits and technology to reduce memristor current.

Hardware Specification	Parameter
Computing matrix size	64×32
Input precision	6 bit
Weight precision	3 bit
High 3bit quantization precision	3 bit
Low 3bit quantization precision	6 bit
Digital/ Analog power voltage	0.9V/1.8V
Average digital current	5.44 uA
Average analog current	787.52uA
Average Energy efficiency	**3.20 TOPS/W**
Average Latency	**1860ns**

Supplementary Table 1. Hardware specifications for CIM chip during inference and the measured performance results. Energy and latency of VMM operations is conducted with the same parameters utilized in our ECG task demonstration. We modify I&F ADC precision during the high and low 3bit weight computations to improve energy efficiency. We generate 10,000 randomized input vectors and fed into CIM chip to measure the average digital and analog currents through Keithley 2450 Source Meter. The analog current encompasses I&F ADC and memristor power consumption while the digital current contains energy cost of MUX and DFF inside the Chip.

	NE 2023[2]	CICC 2021[3]	IEDM 2022[4]	NC 2023[5]	NC 2025[6]	This work
Cell Type	MRAM	RRAM	FeTFET	AND FeFET	RRAM	RRAM
Bit cell	1T1MTJ	2T2R	1T	2T	1T1R	2T2R
Cryptography	SHA256 &2DHC	XOR	XOR	XOR	Bipartite-sort+PUF	Sign and value encryption, Analog key
Encrypted Weight	Single bit	Single bit	Only sign	Multi bit	Multi bit	Multi bit
Technology	22nm	40nm	14nm	28nm	40nm	28nm
Array size	6.6Mb	16kb	N/A	8×7	16kb	576kb
Code length	288b	128bit	1bit	8bit	128bit	1088bit
Complexity	2^{288} 5.0×10^{86}	2^{128}= 3.3×10^{327}	2^{1}	2^{8}	C_{128}^{64}=2.4 $\times 10^{37}$	2^{1088}= 3.3×10^{327}
Key transfer	Yes	Yes	Yes	Yes	Yes	Only key_S and key_V, not analog key

Hardware Level	Chip	Chip	Device	Array	Array	Chip

Supplementary Table 2. Quantitative comparison of this work and other CIM encryption.

References

1. Liu, J.-C., Wu, T.-Y. & Hou, T.-H. Optimizing incremental step pulse programming for RRAM through device–circuit co-design. *IEEE Transactions on Circuits and Systems II: Express Briefs* **65**, 617–621 (2018).
2. Chiu, Y.-C. *et al.* A CMOS-integrated spintronic compute-in-memory macro for secure AI edge devices. *Nature Electronics* **6**, 534–543 (2023).
3. Li, W., Huang, S., Sun, X., Jiang, H. & Yu, S. Secure-RRAM: A 40nm 16kb compute-in-memory macro with reconfigurability, sparsity control, and embedded security. In *2021 IEEE Custom Integrated Circuits Conference (CICC)* 15.4.1-15.4.2 (IEEE, 2021).
4. Luo, J. *et al.* Novel Ferroelectric Tunnel FinFET based Encryption-embedded Computing-in-Memory for Secure AI with High Area-and Energy-Efficiency. In *2022 International Electron Devices Meeting (IEDM)* 36.5.1-36.5.4 (IEEE, 2022).
5. Xu, Y. *et al.* Embedding security into ferroelectric FET array via in situ memory operation. *Nature communications* **14**, 8287 (2023).
6. Yue, W. *et al.* Physical unclonable in-memory computing for simultaneous protecting private data and deep learning models. *Nat Commun* **16**, 1031 (2025).