nature electronics

Article

https://doi.org/10.1038/s41928-023-01064-1

A large-scale integrated vector–matrix multiplication processor based on monolayer molybdenum disulfide memories

In the format provided by the authors and unedited

Table of Contents

1. Material Characterization **2**

2. Device Characterisation **3**

3. LSI Methodology and Experimental Setup **6**

4. Yield Criteria **9**

5. Variation Analysis **10**

6. Single-device Characterization **11**

7. Open-Loop Compensation and Reliability Analysis **15**
7.1. Modeling of Parasitics 16
7.2. Open-loop Compensation 17
7.3. Reliability Analysis 20

References **23**

1. Material Characterization

To characterize the quality and type of the materials we perform basic Raman and STEM characterisation. Figure S1 shows the Raman spectra of the grown material. The difference between the E_{2g} and A_{1g} peaks is 19.5 cm^{-1}, confirming the monolayer nature. Figure S2 shows the STEM image which also confirms the atomic nature.

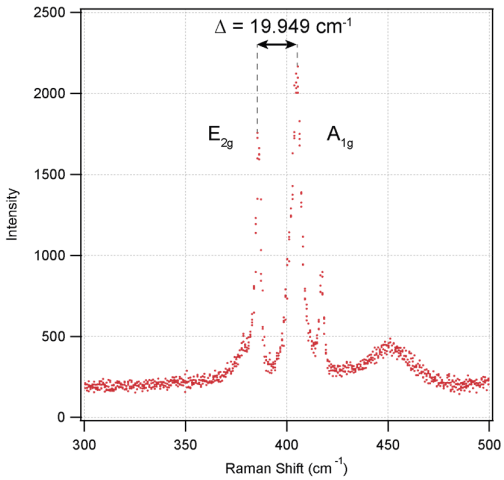

Figure S1. Material characterisation. Raman spectra of MoS_2 transferred onto a SiO_2 substrate. We used a 523 nm laser excitation and a 3000 line mm-1 grating with 10 s acquisition time and averaged from 10 acquisitions. The observed difference between E_{2g} and A_{1g} Raman modes (19.5 cm^{-1}) of MoS_2 is consistent with a monolayer.

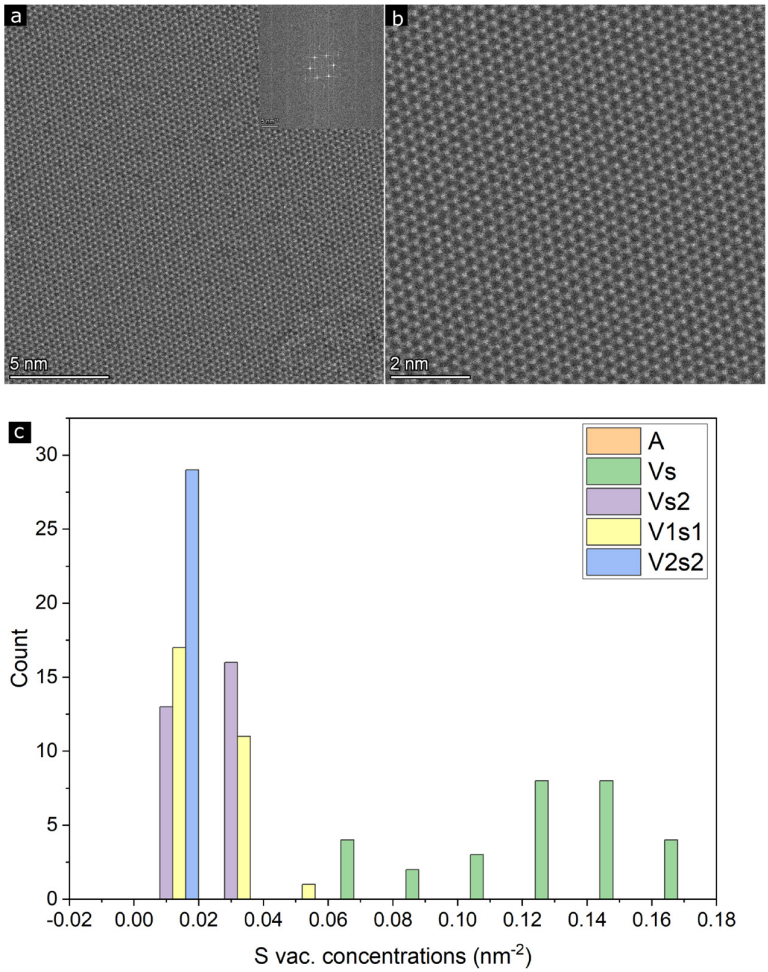

Figure S2. a and b, STEM images of MOCVD grown monolayer MoS_2. STEM imaging of MoS_2. Annular-dark field scanning transmission electron microscopy (ADF-STEM) image shows the hexagonal lattice of single-layer MoS_2. The ADF-STEM image intensity is directly proportional to the atomic number thus the Mo (Z = 42) atoms give brighter contrast than the S atoms (Z=16). **c**, Point defect concentration analysis.

2. Device Characterisation

After device fabrication, the FGFETs were analyzed using different characterization methods in order to assess the integrity and quality of the MoS_2 film and the deposition of oxides and metals. In Figure S3 to Figure S7, we show the AFM, photoluminescence, EDX spectroscopy, SEM, and cross-sectional STEM image of a FGFET fabricated on the same wafer as the circuit presented in the main text. These characterization methods demonstrate that the device corresponds with the expectations and show that the MoS_2 quality is preserved during fabrication.

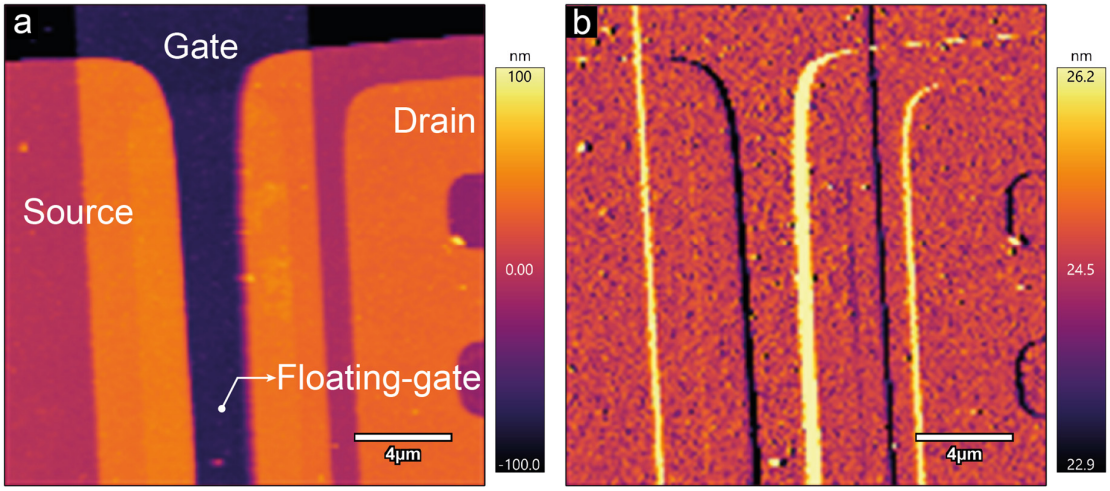

Figure S3. AFM image of the FGFET gate channel, showing different parts of the device. **a,** Height image showing the different device parts. **b,** Amplitude image.

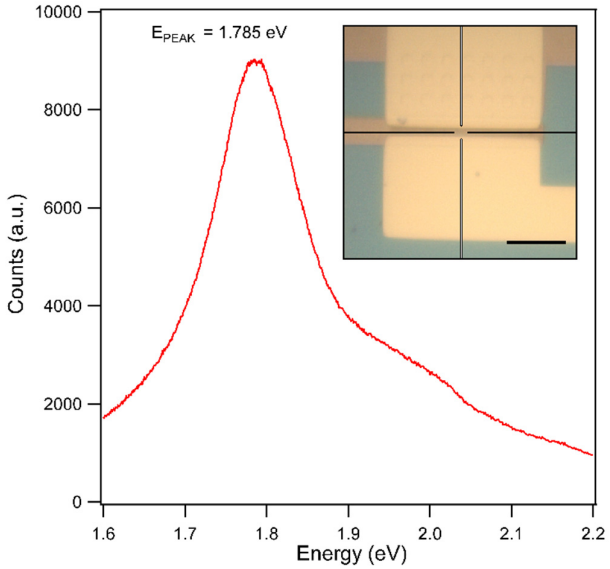

Figure S4. Photoluminescence spectrum acquired on MoS_2 from one of the completed devices.

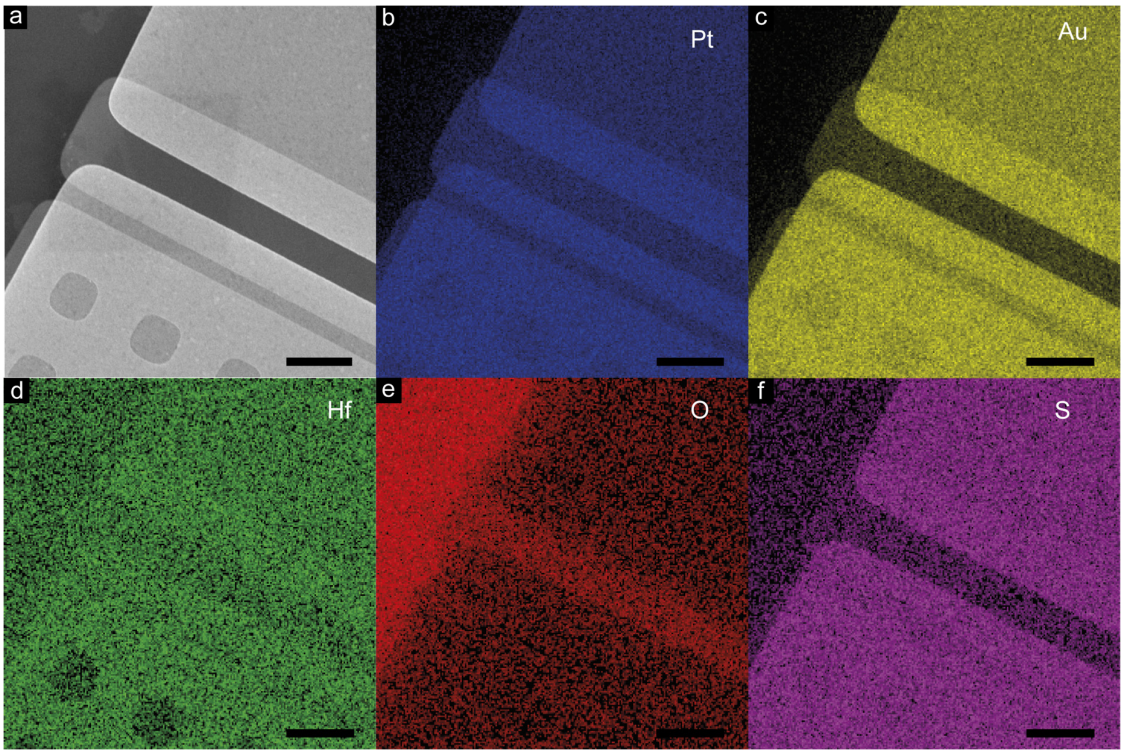

Figure S5. EDX maps of a FGFET fabricated on the same wafer. **a,** SEM image of zoom region in the channel region of the FGFET devices. **b,** Platinum (Pt) EDX map of the SEM region. **c,** Gold (Au) EDX map of the SEM region. **d,** Hafnium (Hf) EDX map of the SEM region. **e,** Oxygen (O) EDX map of the SEM region. **f,** Sulfur (S) EDX mapping of the SEM region. Scale bar for all images: 5 μm.

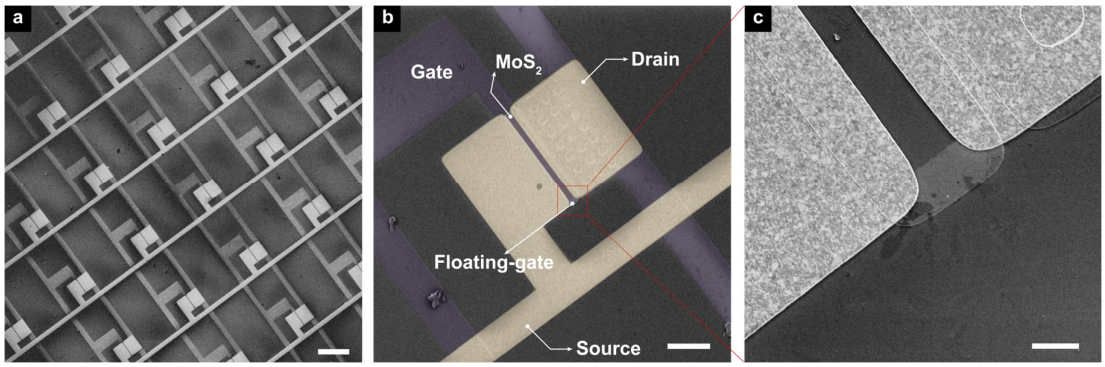

Figure S6. Colored SEM image of a FGFET fabricated on the same wafer as the circuit presented in the main text. a, Wide-view of many devices organized in an array, scale bar: 100 µm. **b,** Fake colored image of a single device, scale bar: 20 µm. **c,** Zoom in the channel region of the FGFET devices, highlighting, drain-source contacts, MoS_2 channel, floating-gate and gate, scale bar: 3 µm.

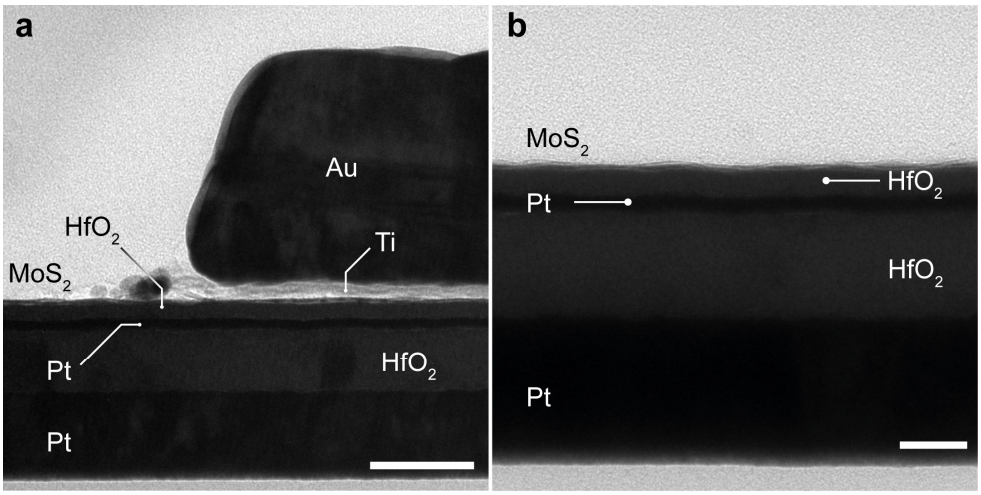

Figure S7. FGFET TEM device cross-section. a, Magnified view in the device contact region, scale bar: 50 nm. **b,** Cross-section image of the gate stack consisting of (from bottom to top) Pt bottom gate, HfO_2 blocking oxide, Pt floating gate and HfO_2 tunnel oxide. The MoS_2 2D channel is on top of the gate stack, scale bar: 20 nm.

3. LSI Methodology and Experimental Setup

To achieve a high yield and low device-to-device variation, we combine a wafer-scale fabrication process with VLSI methodology tools. To design our chips, we take advantage of the Tanner EDA© platform. First, we create the parametric cells of our devices and the chip design. Next, we carry out the fabrication process in 100 class clean-room in a 4-inch fabrication line. Figure S8a shows the image of the final wafer and, in Figure S8b, the image of the final chip after dicing.

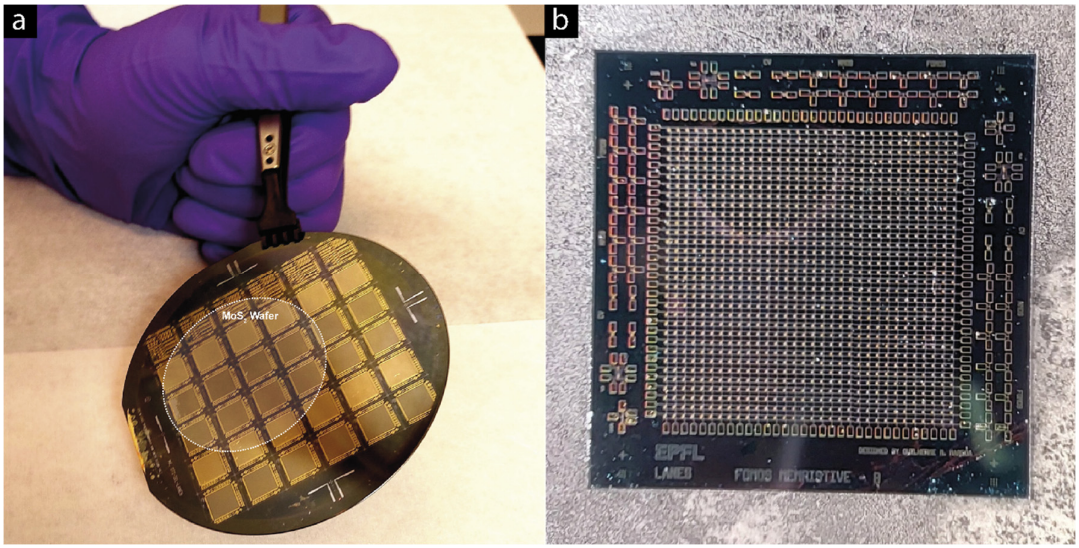

Figure S8. Wafer-scale fabrication. a, Photo of the fabricated 4-inch wafer. **b,** Micro-photograph of the fabricated FGFET matrix.

To characterise our full processor and its operation as a signal processor, we designed a full measurement setup consisting of a compactRIO running Real-Time LabView with analog input-output and digital input-output modules, a device interface board with the capability of selecting individual rows and columns or a range of lines and measuring the output current. Figure S9a shows the schematic of the measurement setup with the previously described components, and Figure S9b the optical image of the setup. Figure S10 shows the details of the board schematics used to measure our devices. The board is composed of 3 main modules: 32-SPST (Single Pole Single Throw), 32 bit serial-to-parallel converter and 32-array transimpedance amplifier. Firstly, the 32-SPST (Single Pole Single Throw) is an array of bidirectional analog switches that can be independently switched ON and OFF. Secondly, the serial-to-parallel converter that takes the input digital signal from the real-time module and converts it to the digital output signal used to select each individual SPST switch. Finally, the transimpedance amplifier is used to convert the currents from the columns in readable voltage and to provide a virtual ground for the signals. The circuit is a simple transimpedance amplifier based on an operation amplifier.

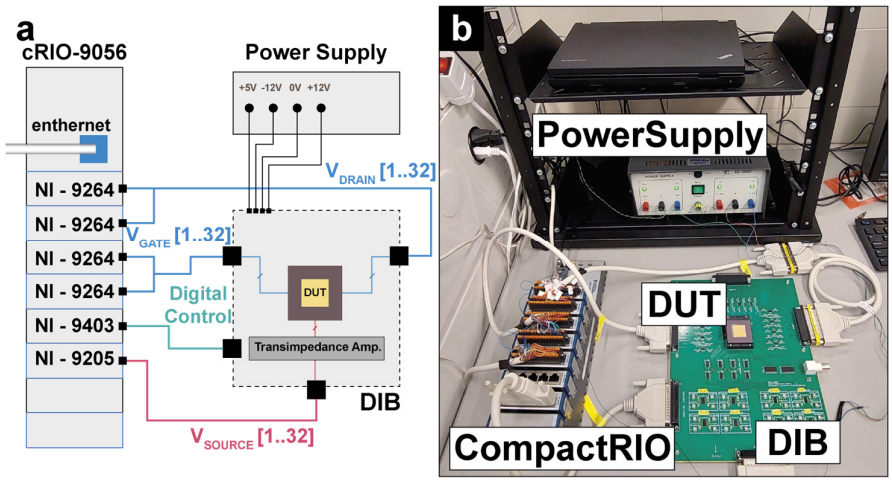

Figure S9. Experimental setup for high-speed measurements, characterisation and processing. a, Schematic of the experimental setup consisting of a power supply generating +5V, ±12V and ground signals; Device Interface Board (DIB); CompactRIO running a web server with modules for applying and reading analog and digital signals. **b,** Optical image of the previously described experimental setup.

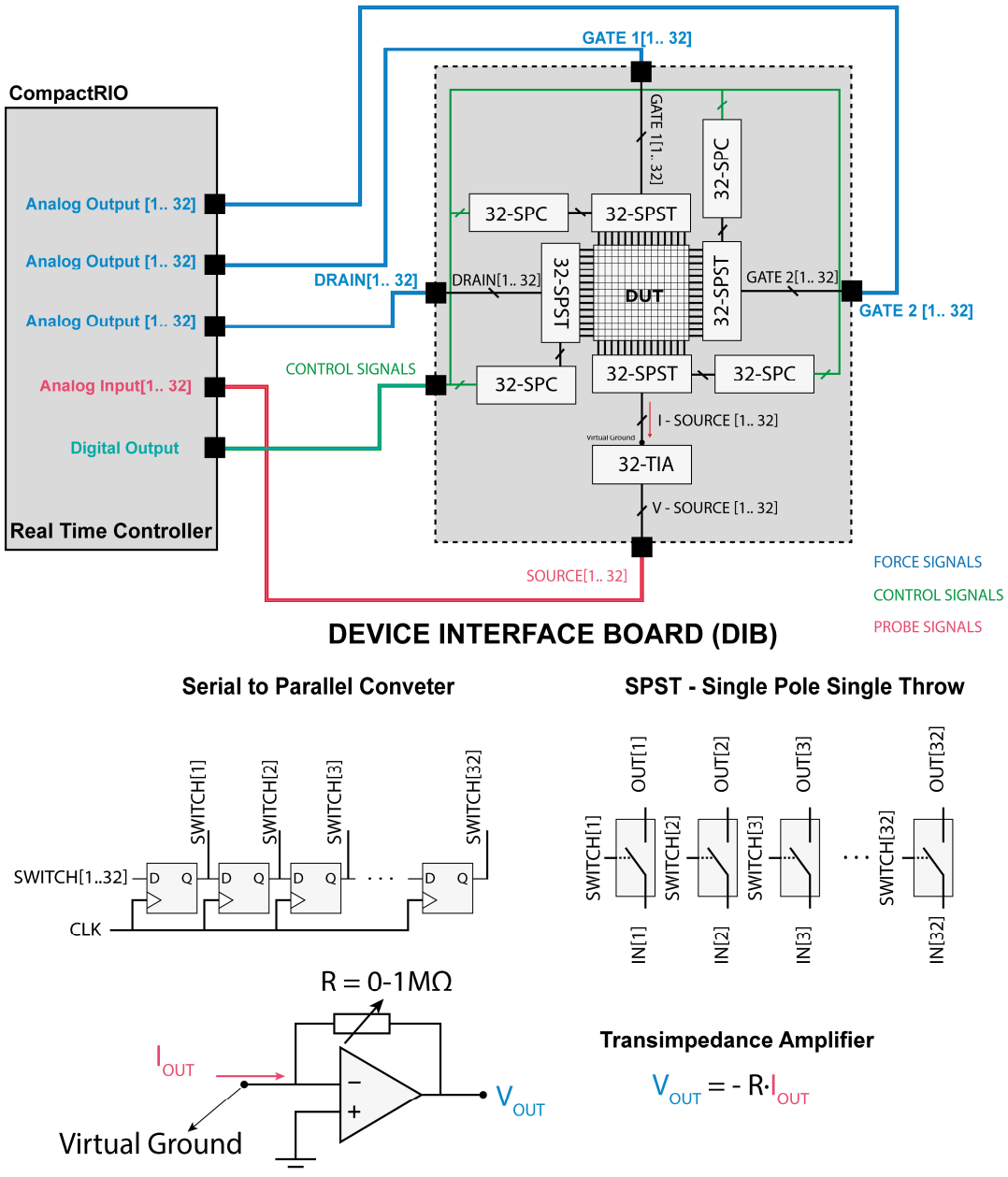

Figure S10. Device interface board schematics and details.

4. Yield Criteria

Since we have a multitude of devices (> 1000), we need to establish a set of criteria that can quantitatively define if a device is working. For us, memory must show a minimal current modulation and a memory window to be considered a working device. The following criteria are an attempt to consider many of the problems that can be encountered during device fabrication.

Criteria:

1. Oxide layers showing high leakage currents (**Criteria:** max ($I_G^{(MAX)}$, $|I_G^{(MIN)}|$) > 1μA).
2. High leakage currents in the OFF state (**Criteria:** $I_{DS}^{(MIN)} > 1\mu A$).
3. Short-circuit of the drain-source contacts (**Criteria:** $I_{DS}^{(MAX)} > 100\ \mu A$) or broken drain-source contact/line (**Criteria:** $I_{DS}^{(MAX)} < 1$ nA).
4. The modulation ratio between the OFF and ON states is very small (**Criteria:** $|I_{ON} / I_{OFF}| < 25$).
5. Large device current at $V_G^{(MIN)}$ (**Criteria:** $I_{DS}@\ V_G^{(MIN)} > 1\mu A$).
6. Memory window too small (ΔV_{TH}) or absent (**Criteria:** $\Delta V_{TH} < 2$). For the estimation of the memory window (ΔV_{TH}) we use the constant current method taken at 1μA.

The values of the reference currents ($I_G^{(MIN)}$, $I_G^{(MAX)}$, $I_{DS}^{(MAX)}$, $I_{DS}^{(MIN)}$) are higher than traditional ones because our setup has a limited current resolution while performing a full chip characterization. Use of higher resolution setups can result in lower current measurement and more precise estimations. However, in our case, the resolution of the setup is not detrimental for the estimation of the memory yield.

Applying the previously described conditions, we measure an 83.1% yield with 851 devices working. Relaxing the values of the previous criteria can result in even higher fabrication yields (≈90%). Figure S11a shows the distributions of the failed devices following each of the previously described conditions. We observe that criteria 2 and 3 are the most impactful for yield. Scratched lines and local damage in the material during the transfer are likely caused for failures of complete lines and/or columns on the chip. Figure S11b shows the distribution of the failed devices and the failure category.

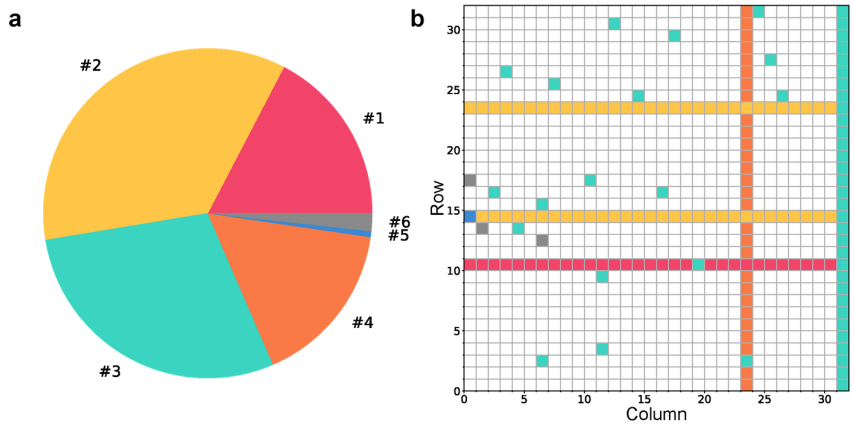

Figure S11. Failure analysis and distribution. a, Pie chart showing the distribution of failed devices that meet of the of criteria. **b,** Distribution on the chip of each categorized failed device.

5. Variation Analysis

Due to a large number of working devices, 851 floating-gate memories, we are able to perform statistical analysis of the most important device metrics. Figure S12 shows the statistical distribution of different parameters extracted from the I_{DS} *versus* V_G curves. Figure S12a shows the distributions of maximum currents (I_{MAX}) and absolute values of the minimum current (I_{MIN}) on the $\log_{10}$ scale. Both curves are fitted with a gamma function. The exponential decay seen for the minimum current can be explained by the low resolution of the measurement system where most of the readings are at the lowest end of the scale. Using a higher-resolution measurement setup would probably lead to a distribution more similar to the one seen for the maximum current which has a bell-shaped distribution. Figure S12b shows the memory window (ΔV_{TH}) measured using the constant current method at 1μA. The distribution is also fitted using a gamma function. A likely explanation is that the gamma distribution of the memory window is due to the exponential nature of the Fowler-Northeim empirical tunneling function (1).

$$i_{tunelling} = C \cdot V^{2\alpha} \cdot e^{-\frac{B}{V}} \tag{1}$$

Small variations in the parameters and voltages should lead to an exponential variation of the tunneling current. Since the memory window (ΔV_{TH}) is directly proportional to the charge stored (Q) in the trapping layer as described by equation (2):

$$\Delta V_{TH} = \frac{Q}{C_{FG-CG}} \tag{2}$$

exponential changes in the stored charge will produce an exponential change in the memory window. This would explain an intrinsic relation for using a gamma distribution as a fitting curve.

$$\Delta V_{TH} = \frac{Q}{C_{FG-CG}}$$

Finally, Figure S12c shows the I_{LRS}/I_{HRS} taken at constant voltage V_G = -3V. It can be fitted by a gamma distribution. Due to the limitation of the setup, the base current is limited to 10 nA. However, single-device measurements performed using an Agilent E5270B parameter analyzer show that the minimum current is much smaller, in the range of ~pA, which would increase the I_{LRS}/I_{HRS} ratio by a factor of 1000. If this correction were applied, the average ratio would be in the 10^5 range, common for high-quality MoS_2 devices.

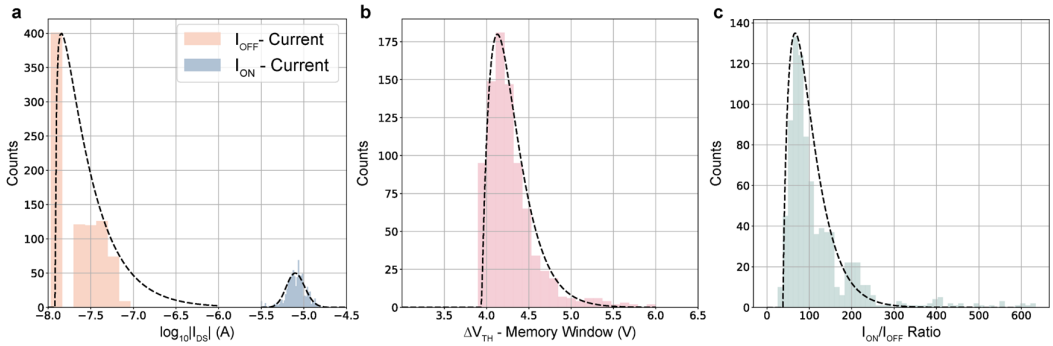

Figure S12. Distribution of extracted features of the floating-gate memories. a, $I_{OFF,}$ and I_{ON} current distributions. The curves are fitted with a gamma distribution with a mean (μ) of -7.5 and -5.09 and a variance (σ^2) of 0.1 and 0.01 **b**, ΔV_{TH} - Memory Window distribution taken with the constant current method at 1 μA. The curve is fitted with a gamma distribution with a mean (μ) of 4.30 and a variance (σ^2) of 0.06. **c**, I_{ON}/I_{OFF} ratio distribution taken with the constant voltage at -3V. The curve is fitted with a gamma distribution with a mean (μ) of 101 and a variance (σ^2) of 2184.

6. Single-device Characterization

To characterize the entire chip, we have opted to develop our own instrumentation in order to gain flexibility and speed by using a setup which is closely tailored to our chips. However, to achieve this, we had to compromise the resolution of our measurements. To have a more complete characterization and understanding of our FGFETs, we developed a secondary device interface board that consists of very low leakage multiplexers (MAX336CWI-T) that connect the devices' rows and columns to an Agilent E5270B parameter analyzer. Figure S13a shows the schematic of the system and Figure S13b shows the optical image of the board with the VMM processor.

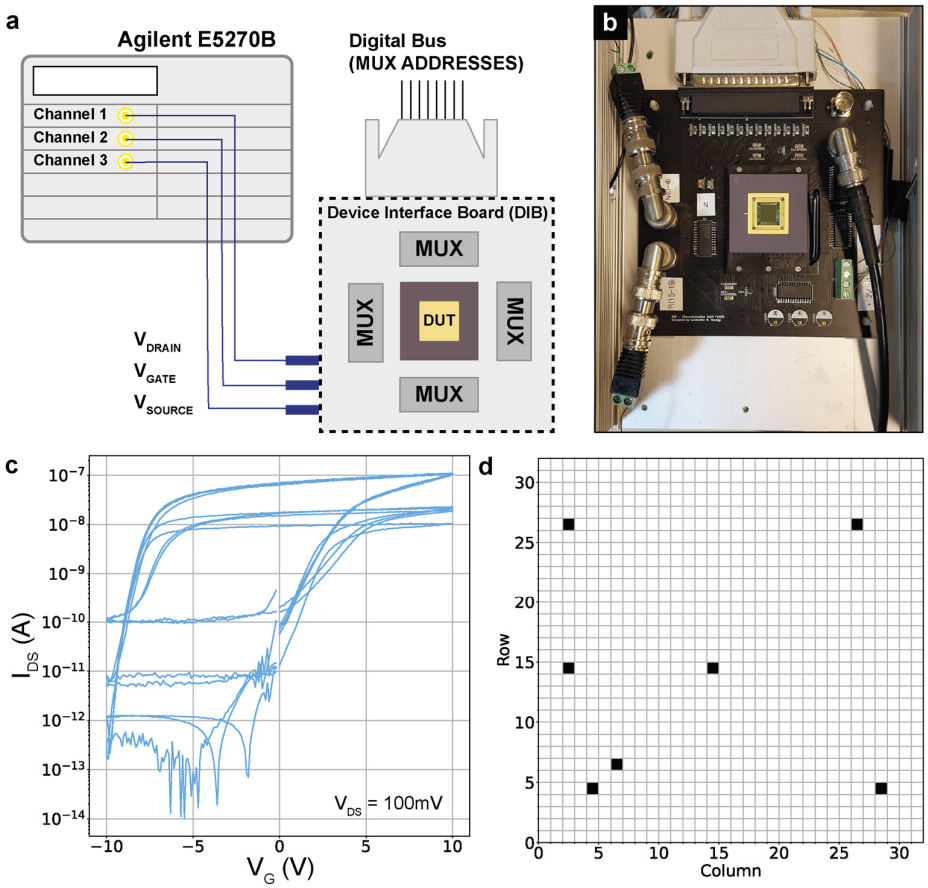

Figure S13. High-resolution hysteresis characterisation of the devices. a, Schematic of the setup with the parameter analyzer connected to the device interface board (DIB) and device under test (DUT). **b,** Optical image of the setup with the device. The board is placed into a metal container for shielding the measurements from external noise. **c,** I_{DS} *versus* V_G characterization curves for different memories on the device for V_{DS} = 100mV. **d,** Mapping of the measured memories on the device.

Figure S13c shows the I_{DS} *versus* V_G curves of our chip for V_{DS} = 100mV while Figure S13d shows the physical distribution of the devices on the chip. We confirm that the OFF current is on the order of ~pA and that the higher OFF current observed in the full chip characterization is due to a limitation of the measurement setup. Also, for single-device measurements, we observe a much larger memory window due to the slower slew-rate of the high-resolution measurements, similarly to previously realized devices[1–3].

Figure S14 shows additional memory characteristics. Figure S14a shows the I_{DS} *versus* V_{DS} for different programming voltages, read at 0 V. We observe a linear behavior near 0 V drain-source bias, confirming an ohmic trend and therefore showing that at low biases the memory acts as a programmable resistor at low bias. This is an important feature since it demonstrates the symmetry and linearity of the multiplication operation done in the physical layer while performing calculations in the context of in-memory computing. Figure S14b shows the retention of the memory for different programming voltages during a 1-hour time period and read at 0V gate voltage. Finally, Figure S14c exhibits the endurance of the memory after 500 program and erase cycles.

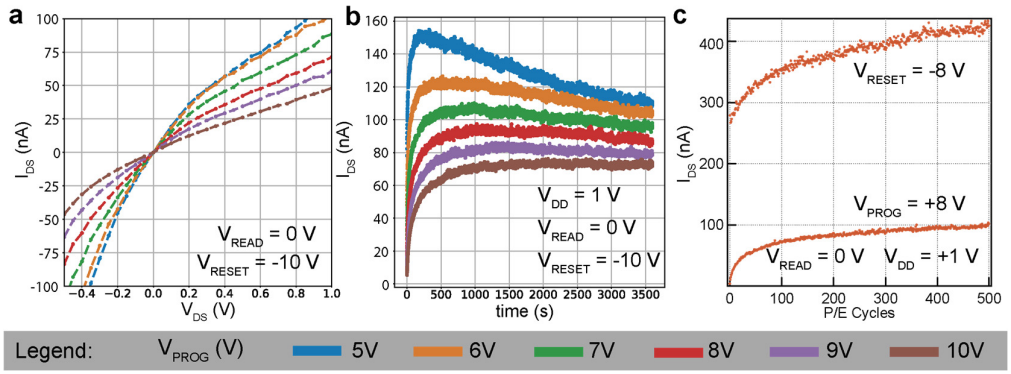

Figure S14. Memory characterization curves. a, I_{DS} *versus* V_{DS} for different programming voltages (V_{PROG}) and V_{READ} = 0V. **b,** Retention measurements – Drain-source current stability during a 1-hour time period. **c,** Endurance Measurements – Program and Erase (P/E) Cycles.

The lower performance observed for these devices can be explained by two possible reasons. Firstly, before the single device measurements, the devices have been measured continuously for a period of one month, this could have decreased their performance due to fatigue. Secondly, a possible leak in the epoxy sealing of the chip package could allow oxygen to enter inside, decreasing the device performance.

Figure S15 shows the comparison between different 2D-flash memories with regards to their I_{ON}/I_{OFF} ratio, subthreshold swing (SS), normalized memory window ($\Delta V_{TH}/V_{PROG}^{(MAX)}$) and maximum programming voltage ($V_{PROG}^{(MAX)}$). We observe that the memories presented in this work are among the best-performing 2D-based flash memories. We also stress that our memories are not based on exfoliated but on wafer-scale grown 2D materials.

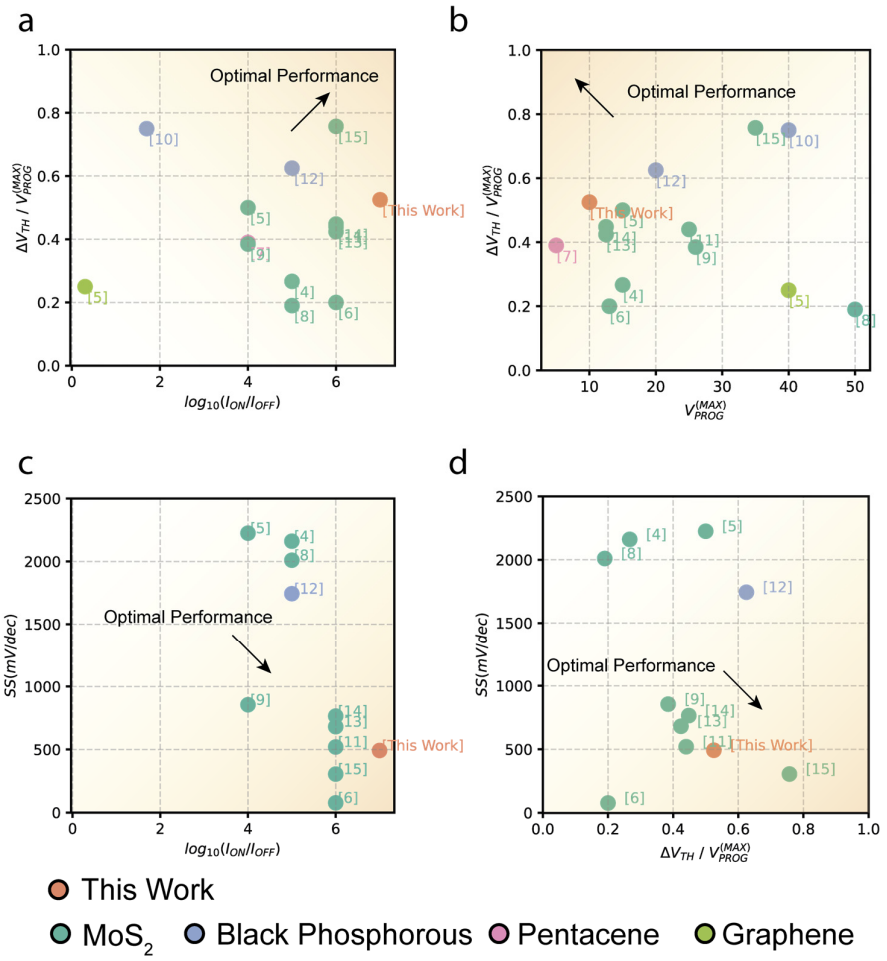

Figure S15. State-of-the-art comparison of state-of-the-art 2D flash memories from the literature. a, The normalized programming window by the maximum programming voltage ($\Delta V_{TH}/V_{PROG}^{(MAX)}$) versus I_{ON}/I_{OFF} ratio. **b,** The normalized programming window by the maximum programming voltage ($\Delta V_{TH}/V_{PROG}^{(MAX)}$) versus the maximum programming voltage ($V_{PROG}^{(MAX)}$). **c,** Subthreshold slope (mV/dec), from low to high, versus the I_{ON}/I_{OFF} ratio in $\log_{10}$ scale. **d,** Subthreshold slope (mV/dec), from low to high, versus the normalized programming window by the maximum programming voltage ($\Delta V_{TH}/V_{PROG}^{(MAX)}$). These values are extracted from papers or from the graphs in the following papers[4–15].

In order to justify our choice of flash memory devices, Table 1 presents a comparison of the different types of memories in regard to their most important metrics, extracted from Kim et al. (doi:10.1109/VLSI-TSA.2012.6210112). From this table, we observe that flash memories have comparable cell area to RRAM and PCM memories which are considered to have high cell density. Moreover, flash memories have lower programming write energy than most of the other memories, and of the same order of magnitude as PCM memories. In contrast, flash memories have a significantly slower write and erase time in the order of **ms** compared to **ns** for the other emerging memories. However, as demonstrated by L. Liu et al, MoS_2 based flash memories can achieve ultra-fast programming with 10's of ns pulses[1–3]. Hence, flash memories based on 2D materials can overcome one of the major drawbacks compared to traditional flash memories.

Table 1 Comparison of performance metrics among different types of memory devices[23].

Performance metrics	DRAM	Flash	PCM	STT-MRAM	RRAM	HDD
Feature size (nm)	36	22	45	95	9	NA
Cell Area	$6F^2$	$4F^2$	$4F^2$	$4F^2$	$4F^2$	~256*
Write/Erase Time	< 10ns	**1/0.1ms**	100ns	<10ms	<1ns	5ms
Retention	64ms	>10y	>10y	>10y	>10y	>10y
Endurance	$>10^{16}$	10^4	10^9	$>10^{12}$	10^{12}	$>10^{16}$
Nonvolatility	N	Y	Y	Y	Y	Y
Multi-level capability	N	Y	Y	N	Y	-
Write Energy (J/bit)	4×10^{-15}	$> 2\times10^{-16}$	6×10^{-16}	2.5×10^{-12}	10^{-13}	-
Standby Power (W/Gb)	10^{-1}	10^{-3}	10^{-3}	10^{-3}	10^{-3}	110

7. Open-Loop Compensation and Reliability Analysis

During the open-loop characterization, we convert the output currents of each device into voltage by means of a transimpedance amplifier. To rescale the voltage into current and then to a workable range, we apply a constant digital gain (G_{CONST}) in the output current (I_{OUT}). Therefore, we can write:

$$y_{OUT} = -\frac{G_{CONST}}{A_{TI}} V_{OUT} \tag{3}$$

The memory state can be fixed $w = y_{OUT}$ for a defined V_{DS}, in this case 1 V. Initially, we observed that the output (w_{OUT}) maps had a preferred direction as can be seen in Figure S16. A couple of hypotheses could be raised, first: the mobility or other material parameters changed linearly over the device. This would be highly unlikely since the deposition did not have a preferential flow direction that could influence the behavior in such a manner. The other hypothesis would be that the parasitic impedances on the chip related to the access line resistance produced such behavior.

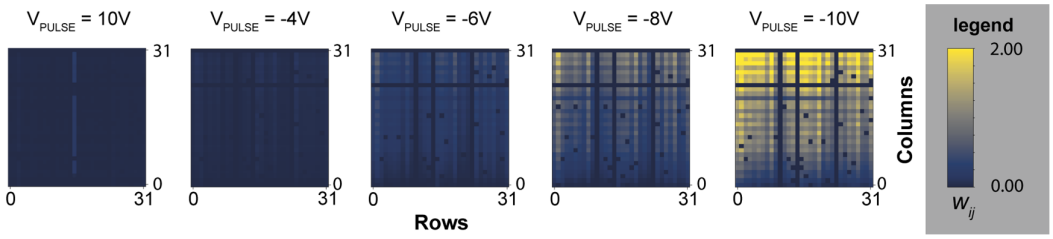

Figure S16. w_{OUT} maps for different programming voltages, using the constant compensation method.

7.1. Modeling of Parasitics

To tackle this problem, we model the parasitic resistances of the device. First, we created a linear resistance model of the drain-source lines. The gate line is not considered since the small gate currents should produce negligible voltage drops in the programming lines. The resistance of the first (M1) and second (M2) metal lines is determined by Ohm's law.

M1 – Drain Lines		**M2 – Source Lines**	
ρ_{Pt}	0.106 μΩ/m	ρ_{Au}	0.022 μΩ/m
W_{Pt}	12.4 μm	W_{Au}	14 μm
t_{Pt}	40 nm	t_{Au}	60 nm
$L_{DRAIN}^{(i)}$	250 μm	$L_{SOURCE}^{(j)}$	250 μm

Table 2. Parameters for modeling the parasitic resistances for metal 1 (drain and gate lines) and metal 2 lines (source lines).

The access resistance (i,j) can be broken down into the resistance of the row and the resistance of the column, as described in Figure S17a. Equations used for estimating the total resistance are given below:

$$R_{DRAIN}^{(j)} = \frac{\rho_{Pt}}{W_{Pt} \cdot t_{Pt}} \cdot L_{DRAIN} \cdot (31 - j) \tag{4}$$

$$R_{SOURCE}^{(i)} = \frac{\rho_{Au}}{W_{Au} \cdot t_{Au}} \cdot L_{SOURCE} \cdot (31 - i) \tag{5}$$

$$R_{PARASITICS}^{(i,j)} = R_{DRAIN}^{(j)} + R_{SOURCE}^{(i)} \tag{6}$$

Next, a resistance map is generated using a Python code. In this model, we observe that M2 lines have a negligible resistance value and therefore, most of the drop happens in the drain lines.

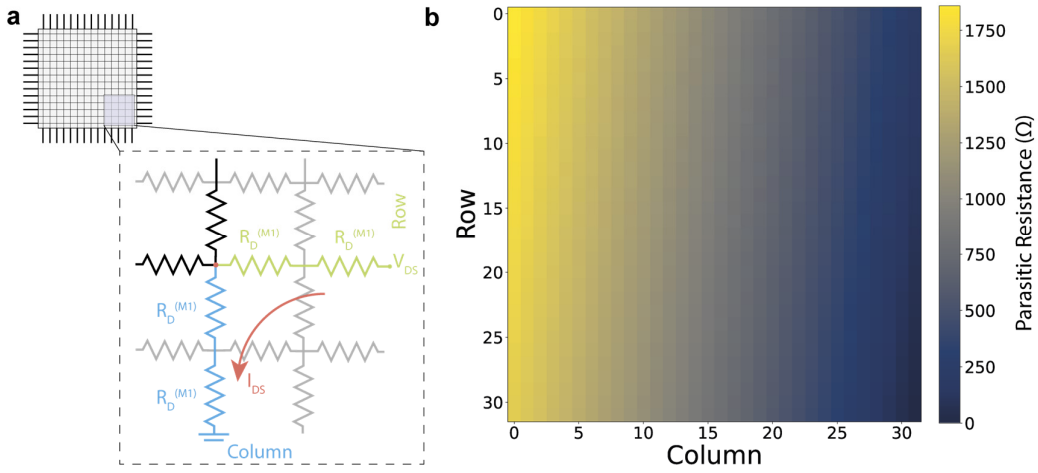

Figure S17 Parasitics modelling. a, Parasitics modeling of the wire resistances in the matrix. Shown in blue are source metal (M2) resistances and in green the drain line metal resistance (M1). **b,** Map of the parasitic resistance estimate.

7.2. Open-loop Compensation

Since we observe a linear increase of the resistance from right to left, there will also be a voltage drop from right to left. To compensate for this effect, we apply a linearly increasing gain ($G_{LINEAR}^{(j)}$) in each column output. The linear gain is now expressed in terms of a tensor matrix with gains in its diagonals.

$$G_{LINEAR} = 4 \cdot 10^5 \begin{bmatrix} 10 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 32 \end{bmatrix} \tag{7}$$

$$y_{OUT} = A_{TI} \cdot (G_{LINEAR} \cdot V_{OUT}^{t})^{t} \tag{8}$$

The memory state can be fixed at $w = y_{OUT}$ for a defined V_{DS}, in this case 1V. By applying the previously described compensation method we compare the weight distribution of the memory array for different programming voltages. Figure S18a-d shows the comparison of the histograms of the output weights for different programming voltages with both compensations (constant/linear) and on different scales (linear/$\log_{10}$). Initially, comparing the states for the linear scale, we observe that with constant compensation, the distribution is broader and has a gamma-like shape. After applying the linear compensation, the distribution becomes closer to a gaussian, and has a smaller spread. A similar behaviour can be seen on the $\log_{10}$ scale, where the linear compensation has a smaller spread and more defined states.

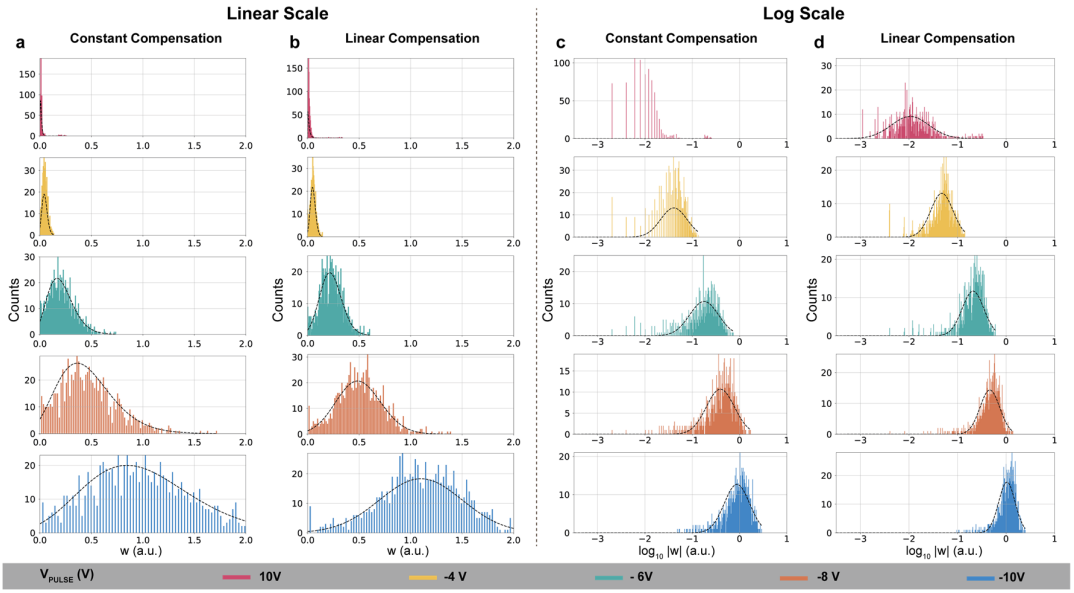

Figure S18. Output state histograms and fitting for different programming voltages. a, With constant compensation in linear scale. **b,** With linear compensation and linear scale. **a-b** where fitted using a gamma distribution. **c,** With constant compensation on $\log_{10}$ scale. **d,** With linear compensation on $\log_{10}$ scale. c-d where fitted using a Gaussian distribution.

Another way of analyzing this behavior is by comparing the empirical cumulative distribution function which gives the percentage of devices that have a certain value up to a certain level (w_i). For sharper transitions from 0 to 100%, the memory states are more closely found around the same w_i value. In Figure S19a-d shows the ECDF function of the different compensations at different scales. A similar conclusion can be taken, that linear compensation results in memory states (w) forming a narrower distribution, more concentrated around a similar value when compared to constant compensation. Another interesting piece of information that these plots show is that for the linear compensation on $\log_{10}$ scale, the states are practically independent which makes us conclude that a multi-state storage within a single pulse is possible. The multi-level storage has some imperfections, but with improvements in fabrication and material, these are expected to decrease. Furthermore, many applications of in-memory computing are robust to small errors in the programmed weights values of the matrix.

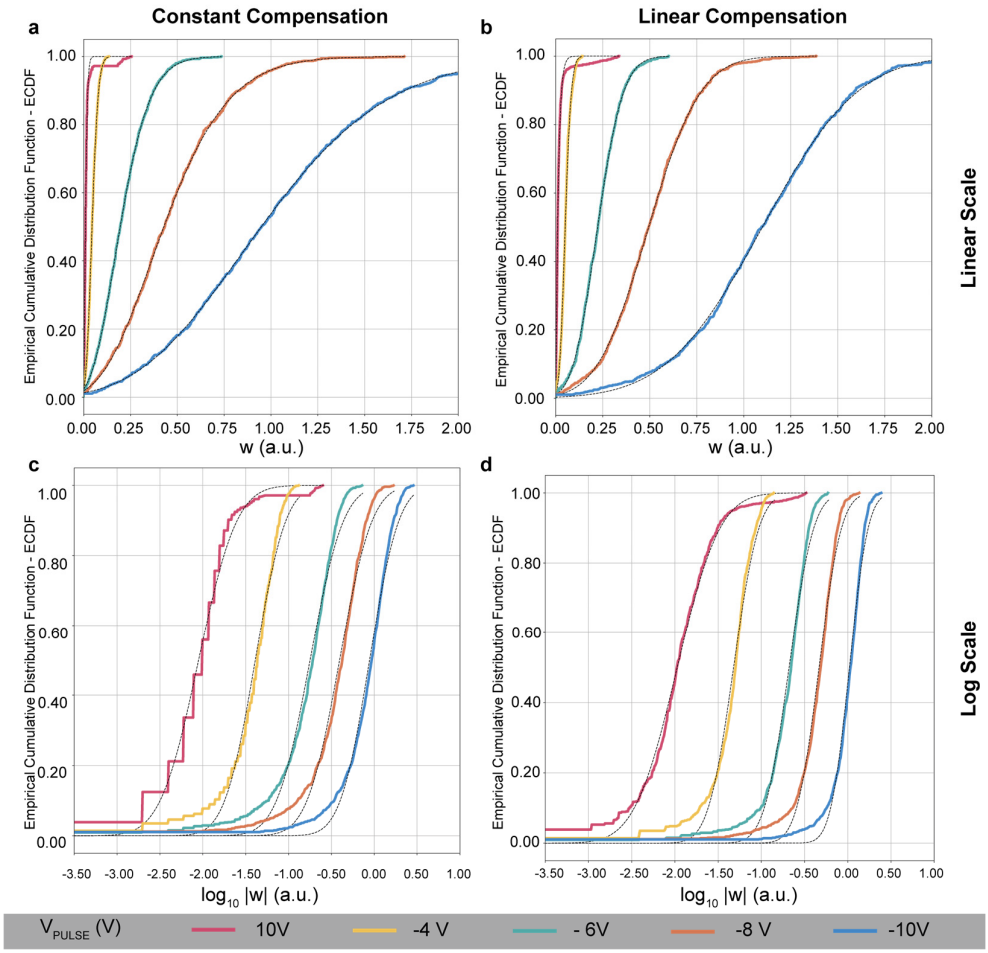

Figure S19. Empirical cumulative distribution function. a, With constant compensation in linear scale. **b,** With linear compensation and linear scale. **c,** With constant compensation and $\log_{10}$ scale. **d,** With linear compensation and $\log_{10}$ scale.

With the fitted values for the distributions with log and linear scale, we can define and plot the average state ($\langle w \rangle$) *versus* programming voltage (V_{PROG}) for both constant and linear compensation. We fitted the intermediate our experimental values with the following functions:

$$y = a \cdot 1.1^{-b \cdot x} + c \qquad \text{(linear scale)}$$

$$y = a \cdot \tanh\big(b(x - c)\big) + d \qquad \text{(}\log_{10}\text{ scale)}$$

To take into account the variability of the programming, we estimate the variability of the function for a standard deviation from the principal curve. Figure S20 shows the experimental values with fitting for the 4 cases with constant/linear compensation and linear/$\log_{10}$ scale. The shaded area gives the error of the fitting for a standard deviation of the fitted curve.

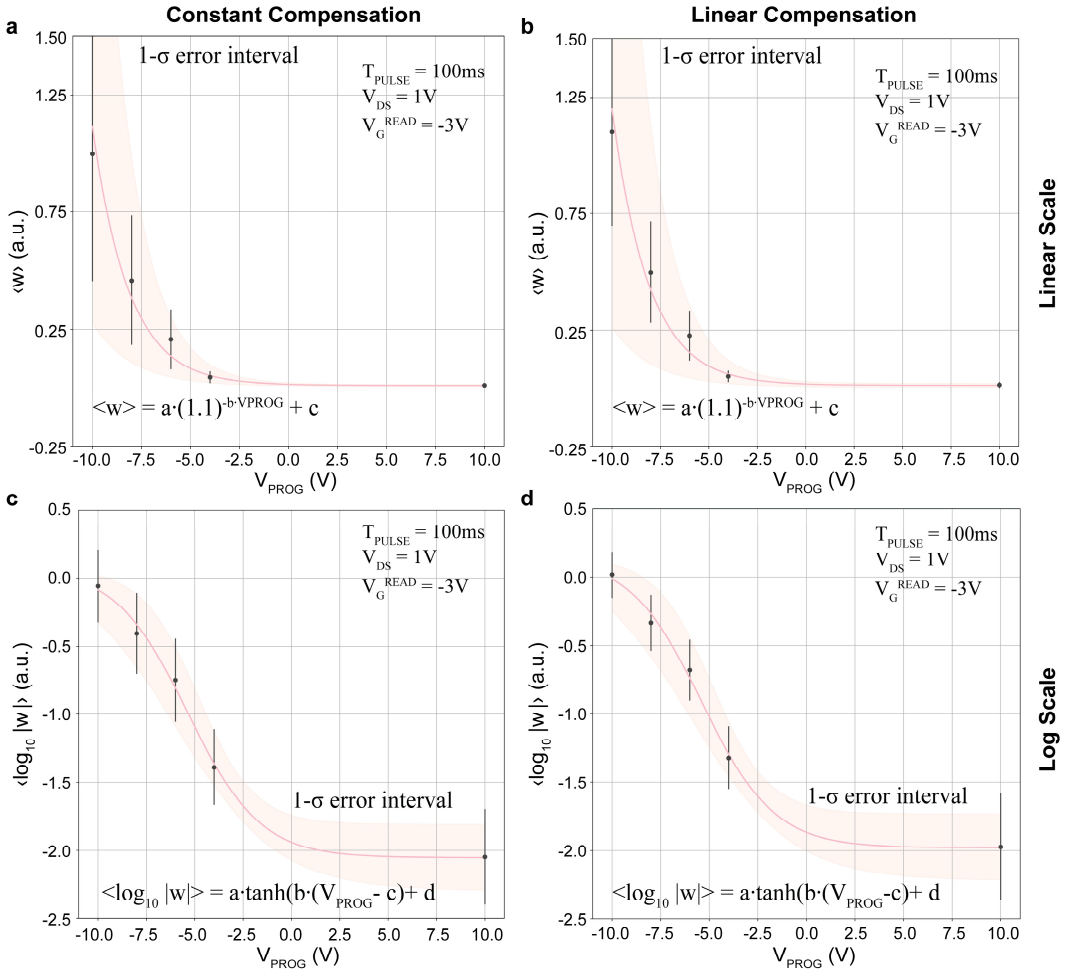

Figure S20. State value (<w>) *versus* programming voltage (V_{PROG}). a, With constant compensation in linear scale. **b,** With linear compensation and linear scale. **c,** With constant compensation and $\log_{10}$ scale. **d,** With linear compensation and $\log_{10}$ scale.

7.3. Reliability Analysis

To analyze the performance of the memories, we compared the memories with constant and linear compensation in regard to their bit programming error, initially for 1-bit programming. To perform this analysis, we take advantage of the Weibull reliability analysis. We construct the failure function (F) as follows:

$$\mathrm{F} = \frac{i - 0.5}{N} \tag{9}$$

This function will be used to assess how many devices will have a certain state in increasing order. Combined with the Weibull formula below we are able to predict and extrapolate its behavior:

$$ln\left(ln\left(\frac{1}{1-F}\right)\right) = A \cdot \ln(y_{OUT}) + B$$

Therefore, we first plot the Weibull formula as a function of the ln(y_{OUT}) output state for each programming state with constant and linear compensation. Figure S21 shows the plots of the mentioned data with the linear fitting for each plot.

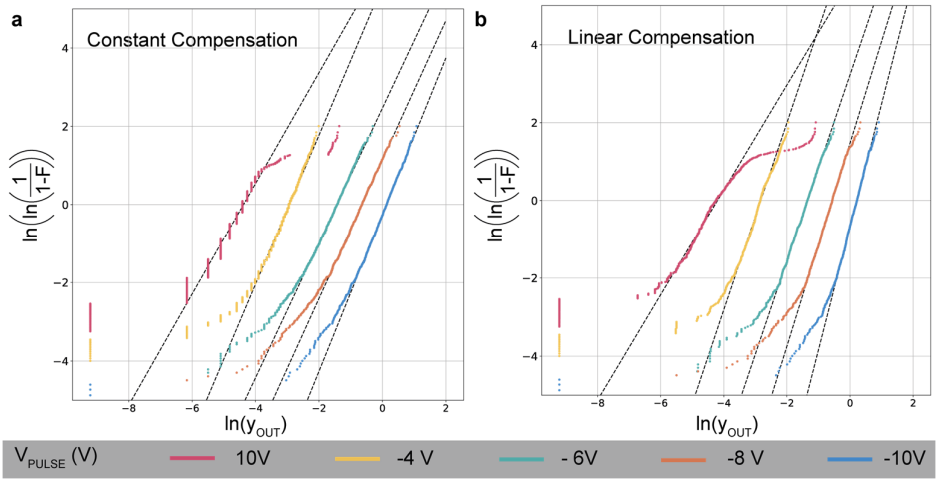

Figure S21. Weibull analysis of the programming error of the memories with each type of compensation. a, Ln(ln(1/(1-F))) *versus* ln(y_{OUT}) with output using a constant compensation. **b,** Ln(ln(1/(1-F))) *versus* ln(y_{OUT}) with output using a linear compensation.

Table 3 shows the coefficient values for the linear regressions plotted in Figure S21.

	Constant Compensation		**Linear Compensation**	
V_{PROG}	**A**	**B**	**A**	**B**
10V	1.416	6.203	1.349	5.663
-4V	1.908	5.590	2.405	6.805
-6V	1.722	2.465	2.392	3.228
-8V	1.777	1.147	2.611	1.475
-10V	2.001	-0.257	3.161	-0.677

Table 3, Weibull fitting parameters for different programming voltages (V_{PROG}) and different types of compensation.

The programming error of our floating-gate memories depends on the threshold (y_{OUT}^{TH}) chosen while the device operates with a single-bit precision. However, depending on how the threshold is chosen/calculated the programming error for bit 0 and bit 1 are different. Therefore, to determine the programming error, we have some degrees of freedom, such as: the programming error of bit 0 or 1 and threshold y_{OUT}^{TH} and we should fix some criteria to estimate the value of the parameters.

As a first calculation, we set a desired error of 1 per million while programming the bit 0 (erase state). This means that 999 999 of the programming values ($y_{OUT}^{(0)}$) fall below the threshold y_{OUT}^{TH}. We first consider the constant compensation method, before applying the same methodology for linear compensation. By design, we can estimate the failure rate of the Weibull distribution:

$$F = \frac{999999}{1000000}, for\ y_{OUT}^{(0)} \leq y_{OUT}^{TH} \tag{10}$$

Thus,

$$ln\left(ln\left(\frac{1}{1-F}\right)\right) = 2.626 \tag{11}$$

Since we have already experimentally extracted the parameters for the Weibull analysis, we can determine the y_{OUT}^{TH}:

$$1.376 \cdot ln(y_{OUT}^{TH}) + 6.020 = 2.626 \tag{12}$$

Hence,

$$\begin{gathered} R_{2D} = \rho \frac{l}{W} ln(y_{OUT}^{TH}) = -2.526 \\ y_{OUT}^{TH} = 0.080 \end{gathered} \tag{13}$$

Now that we know the threshold for the data, we can estimate how many devices fall below this calculated threshold when we are programming the values with bit 1.

$$ln\left(ln\left(\frac{1}{1-F}\right)\right) = 2.001 \cdot (-2.526) - 0.257 = -5.3112 \tag{14}$$

$$\frac{1}{1-F} = 1.0049 \rightarrow \mathrm{F} = 4921 \tag{15}$$

This means that with this threshold, we are going to find about 6000 devices in million that are intended to be programmed with a pulse for bit 1 but fall in the bit 0 zone. This causes 6000 programming errors per million. A visual representation of this analysis is shown in panel a. We applied the same procedure and constraints for estimating the programming error with the linear compensation.

$$ln\left(ln\left(\frac{1}{1-F}\right)\right) = 2.626$$

$$1.322 \cdot ln(y_{OUT}) + 5.540 = 2.626$$

$$ln(y_{OUT}) = -2.204$$

$$y_{OUT}^{TH} = 0.11$$

$$ln\left(ln\left(\frac{1}{1-F}\right)\right) = 3.156 \cdot (-2.204) - 0.674 = -7.630$$

$$\frac{1}{1-F} = 1.0004 \rightarrow \mathrm{F} = 486 \text{ errors per mllion}$$

For the linear compensation, we find 500 programming errors per million that fall in the 0-bit zone. This is a 10× improvement in programming reliability as a result of applying the linear compensation method instead of the constant compensation method. This estimation does not perfectly match reality, since experimentally, we observe a tail in the more negative part of the experimental values. However, this is not detrimental to our analysis as a first-order estimation. This analysis can be extrapolated to the error for a higher number of bits since what affects most directly the estimation of the programming error is the curve slope. In b we see that the slope of the linear compensation method is consistently higher than the one for the constant method.

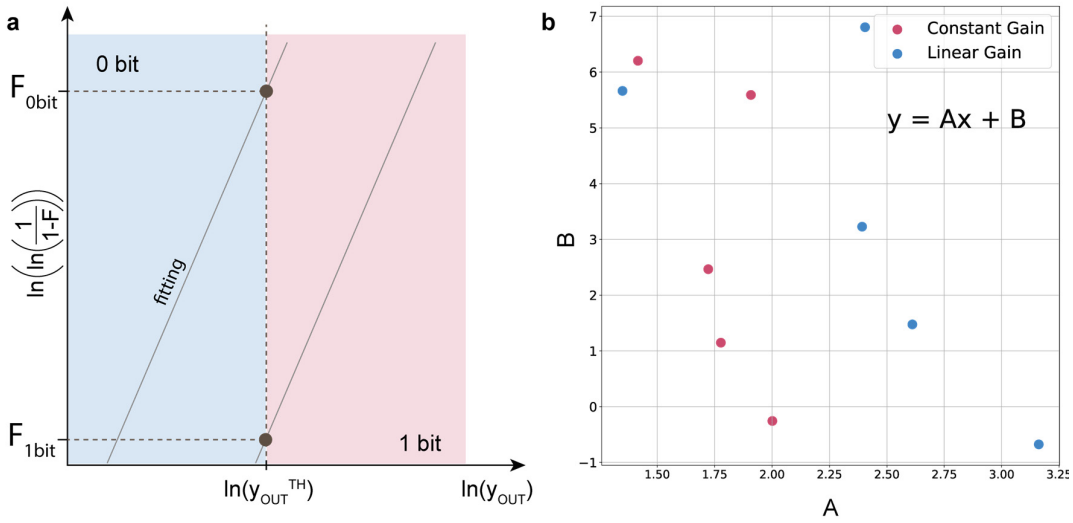

Figure S22. **Extraction of parameters for the Weibull analysis for the programming error of the memories. a,** B *versus* A parameters **b,** B *versus* A parameters.

References

(1) Migliato Marega, G.; Wang, Z.; Paliy, M.; Giusi, G.; Strangio, S.; Castiglione, F.; Callegari, C.; Tripathi, M.; Radenovic, A.; Iannaccone, G.; Kis, A. Low-Power Artificial Neural Network Perceptron Based on Monolayer MoS_2. *ACS Nano* **2022**, *16* (3), 3684–3694. https://doi.org/10.1021/acsnano.1c07065.

(2) Migliato Marega, G.; Zhao, Y.; Avsar, A.; Wang, Z.; Tripathi, M.; Radenovic, A.; Kis, A. Logic-in-Memory Based on an Atomically Thin Semiconductor. *Nature* **2020**, *587* (7832), 72–77. https://doi.org/10.1038/s41586-020-2861-0.

(3) Giusi, G.; Marega, G. M.; Kis, A.; Iannaccone, G. Impact of Interface Traps in Floating-Gate Memory Based on Monolayer MoS $_\text2$. *IEEE Trans. Electron Devices* **2022**, 1–6. https://doi.org/10.1109/TED.2022.3208804.

(4) Bertolazzi, S.; Krasnozhon, D.; Kis, A. Nonvolatile Memory Cells Based on MoS_2/Graphene Heterostructures. *ACS Nano* **2013**, *7* (4), 3246–3252. https://doi.org/10.1021/nn3059136.

(5) Sup Choi, M.; Lee, G.-H.; Yu, Y.-J.; Lee, D.-Y.; Hwan Lee, S.; Kim, P.; Hone, J.; Jong Yoo, W. Controlled Charge Trapping by Molybdenum Disulphide and Graphene in Ultrathin Heterostructured Memory Devices. *Nat. Commun.* **2013**, *4* (1), 1624. https://doi.org/10.1038/ncomms2652.

(6) Paul, T.; Ahmed, T.; Tiwari, K. K.; Thakur, C. S.; Ghosh, A. A High-Performance MoS_2 Synaptic Device with Floating Gate Engineering for Neuromorphic Computing. *2D Mater.* **2019**, *6* (4), 045008. https://doi.org/10.1088/2053-1583/ab23ba.
(7) Han, S.-T.; Zhou, Y.; Wang, C.; He, L.; Zhang, W.; Roy, V. a. L. Layer-by-Layer-Assembled Reduced Graphene Oxide/Gold Nanoparticle Hybrid Double-Floating-Gate Structure for Low-Voltage Flexible Flash Memory. *Adv. Mater.* **2013**, *25* (6), 872–877. https://doi.org/10.1002/adma.201203509.
(8) Han, S.-T.; Zhou, Y.; Chen, B.; Zhou, L.; Yan, Y.; Zhang, H.; Roy, V. a. L. Two-Dimensional Molybdenum Disulphide Nanosheet-Covered Metal Nanoparticle Array as a Floating Gate in Multi-Functional Flash Memories. *Nanoscale* **2015**, *7* (41), 17496–17503. https://doi.org/10.1039/C5NR05054E.
(9) Zhang, E.; Wang, W.; Zhang, C.; Jin, Y.; Zhu, G.; Sun, Q.; Zhang, D. W.; Zhou, P.; Xiu, F. Tunable Charge-Trap Memory Based on Few-Layer MoS_2. *ACS Nano* **2015**, *9* (1), 612–619. https://doi.org/10.1021/nn5059419.
(10) Li, D.; Wang, X.; Zhang, Q.; Zou, L.; Xu, X.; Zhang, Z. Nonvolatile Floating-Gate Memories Based on Stacked Black Phosphorus–Boron Nitride–MoS2 Heterostructures. *Adv. Funct. Mater.* **2015**, *25* (47), 7360–7365. https://doi.org/10.1002/adfm.201503645.
(11) Wang, T.-Y.; Meng, J.-L.; He, Z.-Y.; Chen, L.; Zhu, H.; Sun, Q.-Q.; Ding, S.-J.; Zhou, P.; Zhang, D. W. Ultralow Power Wearable Heterosynapse with Photoelectric Synergistic Modulation. *Adv. Sci.* **2020**, *7* (8), 1903480. https://doi.org/10.1002/advs.201903480.
(12) Tian, H.; Deng, B.; Chin, M. L.; Yan, X.; Jiang, H.; Han, S.-J.; Sun, V.; Xia, Q.; Dubey, M.; Xia, F.; Wang, H. A Dynamically Reconfigurable Ambipolar Black Phosphorus Memory Device. *ACS Nano* **2016**, *10* (11), 10428–10435. https://doi.org/10.1021/acsnano.6b06293.
(13) Migliato Marega, G.; Zhao, Y.; Avsar, A.; Wang, Z.; Tripathi, M.; Radenovic, A.; Kis, A. Logic-in-Memory Based on an Atomically Thin Semiconductor. *Nature* **2020**, *587* (7832), 72–77. https://doi.org/10.1038/s41586-020-2861-0.
(14) Migliato Marega, G.; Wang, Z.; Paliy, M.; Giusi, G.; Strangio, S.; Castiglione, F.; Callegari, C.; Tripathi, M.; Radenovic, A.; Iannaccone, G.; Kis, A. Low-Power Artificial Neural Network Perceptron Based on Monolayer MoS_2. *ACS Nano* **2022**, *16* (3), 3684–3694. https://doi.org/10.1021/acsnano.1c07065.
(15) Liu, L.; Liu, C.; Jiang, L.; Li, J.; Ding, Y.; Wang, S.; Jiang, Y.-G.; Sun, Y.-B.; Wang, J.; Chen, S.; Zhang, D. W.; Zhou, P. Ultrafast Non-Volatile Flash Memory Based on van Der Waals Heterostructures. *Nat. Nanotechnol.* **2021**, *16* (8), 874–881. https://doi.org/10.1038/s41565-021-00921-4.
(16) Sup Choi, M.; Lee, G.-H.; Yu, Y.-J.; Lee, D.-Y.; Hwan Lee, S.; Kim, P.; Hone, J.; Jong Yoo, W. Controlled Charge Trapping by Molybdenum Disulphide and Graphene in Ultrathin Heterostructured Memory Devices. *Nat. Commun.* **2013**, *4* (1), 1624. https://doi.org/10.1038/ncomms2652.