

A multi-timescale synaptic weight based on ferroelectric hafnium zirconium oxide

Mattia Halter ^{1,2,3}✉, Laura Bégon-Lours ¹, Marilyne Sousa¹, Youri Popoff^{1,2}, Ute Drechsler¹, Valeria Bragaglia ¹ & Bert Jan Offrein ¹

Brain-inspired computing emerged as a forefront technology to harness the growing amount of data generated in an increasingly connected society. The complex dynamics involving short- and long-term memory are key to the undisputed performance of biological neural networks. Here, we report on sub- μm -sized artificial synaptic weights exploiting a combination of a ferroelectric space charge effect and oxidation state modulation in the oxide channel of a ferroelectric field effect transistor. They lead to a quasi-continuous resistance tuning of the synapse by a factor of 60 and a fine-grained weight update of more than 200 resistance values. We leverage a fast, saturating ferroelectric effect and a slow, ionic drift and diffusion process to engineer a multi-timescale artificial synapse. Our device demonstrates an endurance of more than 10^{10} cycles, a ferroelectric retention of more than 10 years, and various types of volatility behavior on distinct timescales, making it well suited for neuromorphic and cognitive computing.

¹IBM Research Europe - Zurich Research Laboratory, CH-8803 Rüschlikon, Switzerland. ²ETH Zurich - Integrated Systems Laboratory, CH-8092 Zurich, Switzerland. ³Present address: Lumiphase AG, CH-8712 Stäfa, Switzerland. ✉email: mattia@lumiphase.com

The amount of data created during the last thirty years (~320 ZB) compares with what will be created during the next three years (~364 ZB projected for 2022–2024), a trend also accelerated by the covid-19 pandemic^{1–3}. A considerable amount will be created by the rapidly growing Internet of Things (IoT)⁴, which connects the physical world and computing entities. The development of sensors and actuators that connect to the internet, comes in pair with the emergence of Artificial Neural Networks (ANNs) for data processing. The conventional von-Neumann architecture cannot sustain such evolution, because of the energy and performance bottleneck coming from the massive data movement between the physically separated memory and processing units. Novel processing architectures, device technologies, and computational paradigms have therefore recently emerged. In-memory computing⁵ co-locates memory and processing, eliminating inefficient data movement, which is especially beneficial for ANN training by data-intense machine-learning algorithms. In addition to ANNs, computing systems like Spiking Neural Networks (SNNs), which mimic the type of information processing in the human brain, promise low-power computation and dynamic learning in the context of complex data⁶. A major challenge for SNNs remains in supporting operations under a wide range of effective timescales⁷. This requirement arises from the need to adapt the computation to the input timescale in real-time online applications (e.g., real-world sensory signals) and because multi-timescales are inherent in spiking neurodynamics (e.g., neural activation decay, combination of short- and long-term plasticity mechanisms^{7–9}). To emulate the complex and multi-timescale plasticity processes of biological synapses⁷, physical effects acting at different timescales must be orchestrated in an artificial synapse⁹. Being able to tune the timescale (tunable volatility) in SNNs would bring the ability to mimic many of the basic processing and storage operations of the mammalian brain¹⁰ and facilitate reservoir computing or unsupervised learning¹¹.

In ANNs and SNNs, analog non-volatile resistive memory elements, so-called memristors, are used to emulate synaptic functionality, to locally store network parameters^{12,13}, and to serve as analog computing element¹⁴. Lately, multi-terminal memristive devices have gained a lot of attention due to the ability of using extra terminals to tune their switching dynamics (volatility, plasticity timescales), opening opportunities to implement advanced bio-inspired learning rules. For example, reward-modulated spike-timing-dependent plasticity was demonstrated in a 4-terminal cell based on ferroelectric P(VDFTrFE)¹⁵. The concept of multi-terminal can also be extended to combining electrical and light spikes to control the synapse plasticity¹⁶. Hafnia-based technologies^{17–19} are CMOS-compatible; the low-power and multi-state nature of hafnia-based Ferroelectric Field-Effect Transistors (FeFETs) makes them a viable candidate for the development of a multi-timescale synaptic element in neuromorphic circuits. FeFETs^{20–22} have the advantage over two-terminal devices of separating the read and write path²³. On one hand, this permits to write to a high-impedance gate with low-power. On the other hand, the reading current flows through the channel that can be engineered to be ohmic²⁴.

Scaled FeFETs nevertheless suffer from a finite number of conductance levels²¹ due to the discrete number of ferroelectric domains available in each device. Integrating the FeFETs in the Back-End-Of-Line (BEOL) leaves more area for the control electronics while at the same time it enables analog weight updates through relaxed size constraints^{24–26}.

In this work we present a BEOL-compatible, back-gated FeFET utilizing a 10 nm thick HfZrO₄ (HZO) ferroelectric gate dielectric and a 4 nm tungsten oxide (WO_x) thin-film channel. The screening of the ferroelectric polarization charges of HZO results

in an accumulation or depletion of electrons in the WO_x channel, effectively changing its resistivity. Beyond the absence of interfacial dielectric layers (which limit the performance of FeFETs with doped-Si channels²⁷), oxide channels are of interest for the intrinsic presence of slow ionic drift and diffusion mechanisms, and hence are viable candidates for a slow plasticity timescale effect. WO_x was chosen due to its relatively mobile oxygen-ions and thereby tunable conductivity as a function of oxygen content^{28,29}. Furthermore, the junction less contacts to the WO_x permit a ohmic conduction between source and drain²⁴.

We engineered our device by using a channel with the appropriate stoichiometry and by having it in direct contact with HZO. In the first section, we capture how these measures facilitate a resistance modulation of the channel when the device is programmed with different timescales: on one hand the resistance modulation is driven by fast electronic screening of the HZO polarization charges, and on the other hand by the slow ionic drift and diffusion process of oxygen in the WO_x. In the second section, we characterize the synaptic weight and provide an evaluation of the proposed technology's performance when implemented in an ANN.

Results and discussion

Artificial synaptic weight structure. We developed a back-gated FeFET with two metal lines (M1, M2) for the Gate (G) and two (M3, M4) for the Source (S) and Drain (D) to establish a three-terminal device configuration. The FeFET gate stack is made of a TiN contact and a 10 nm HZO ferroelectric gate dielectric. The thin-film channel consists of a 4 nm WO_x layer on top of the HZO. The S and D contacts are made of Pt/W and the passivation layer above the WO_x consists of an Al₂O₃ and a SiO₂ layer. In our design, the gate fully overlaps with S and D to avoid large topographic steps beneath the channel. The device schematic is depicted in Fig. 1a. In this work, single, stand-alone devices are studied. In view of their integration in passive pseudo-crossbar circuits, the geometry of the channel is designed to optimize the channel resistance (we refer to Supplementary Note 1 where the pseudo-crossbar operation is further discussed). For the integration in active pseudo-crossbar circuits (as proposed for example by Jerry et al.²⁰), the devices would be embedded in at least four metal levels. Therefore, within our stand-alone device processing scheme, we include these challenges inherent to the complexity of the active crossbar realization.

The structural analysis of our device material stack is described in Supplementary Note 2. Grazing-Incidence X-Ray Diffraction GIXRD scans confirm the absence of the monoclinic phase, the peak at 30.7° corresponds to the ferroelectric orthorhombic phase of HZO³⁰. Bright-field Scanning Transmission Electron Microscopy (BF-STEM) analysis confirms the expected layer thickness. Energy-dispersive X-ray spectroscopy reveals the diffusion of Al into the WO_x layer, a metal expected to promote the reduction of WO_x³¹.

In addition to the ferroelectric field-effect, this work exploits the high mobility of oxygen in WO_x, as observed during reduction anneals³² and in WO_x nanowires³³. More generally, WO_x was chosen as a channel for its electron-doping ability through the removal of oxygen^{34,35}. In this work, the as-deposited WO_x was close to a stoichiometry of $x = 3$ (insulating). We then leverage the mobility of oxygen: during the deposition of the passivation layers (Al₂O₃ and SiO₂), the WO_x was slightly reduced (consistently with the Al diffusion found earlier). The resulting mobility and carrier concentration are suitable for a large channel conductance modulation by the polarization charges via Coulomb coupling²⁹: the screening of the HZO polarization charges in the WO_x occurs over a distance

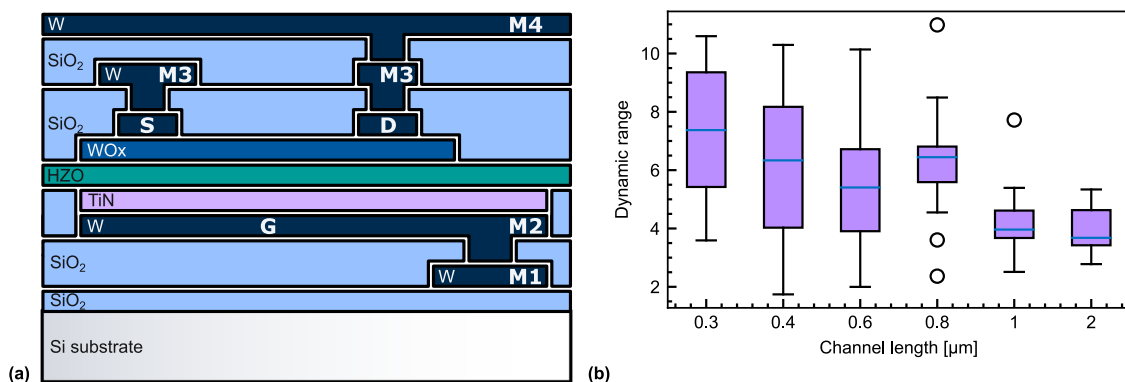


Fig. 1 Structural data of the FeFET. **a** Schematic illustration of the FeFET, indicating a source (S), a drain (D), a gate (G), a WO_x channel, a ferroelectric HZO gate dielectric, G access metal line (M1, M2), S access metal line (M3), and D access metal line (M4). **b** Channel length effect: The dynamic range of 120 devices with a channel width of 600 nm and a varying channel length from 300 nm to 2 μm . Write pulses with an amplitude (V_w) from -6 V to 6 V and width (t_w) of 500 μs were applied. The same data for ($t_w = 500$ ms are available in Supplementary Fig. 6a). The boxes extend from the lower to upper quartile values of the data, with a line at the median. The whiskers extend from the box to show the range of the data. Flier points are those past the end of the whiskers.

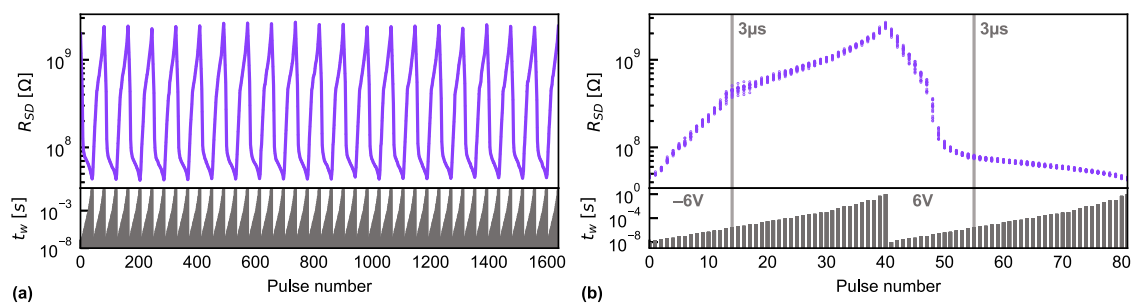


Fig. 2 Potentiation and depression cycles with a constant pulse amplitude scheme. The write pulse width t_w was modulated from 10 ns to 1 s while keeping the write pulse amplitude V_w constant at 6 V (-6 V) for the potentiation (depression) on a FeFET with $L = 400$ nm and $W = 1$ μm . **a** Source-drain resistance as a function of the pulse number for 20 consecutive cycles. On the bottom, the corresponding duration t_w of each pulse is provided. **b** Superposition of the data points shown in **a**) and represented as a function of the pulse number in each cycle, to visualize the small cycle-to-cycle variations and potentiation and depression shape.

comparable to the channel thickness²⁴. A 2D time-dependent Ginzburg-Landau model of the HZO/ WO_x FeFET³⁶ provided guidelines for the optimization of the channel thickness and stoichiometry. Moreover, the resulting channel resistances in the M Ω to G Ω range are well suited for energy-efficient and scalable crossbar array operation³⁷.

The capacitance, resistance, and polarization characterization of the gate stack are found in the Supplementary Notes 3 and 4.

Switching components and timescales. Potentiation (depression) cycles on 120 devices with constant channel width ($W_{\text{ch}} = 600$ nm) and a channel length varying from $L_{\text{ch}} = 300$ nm to 2 μm were measured by applying write pulses with an amplitude (V_w) up to 6 V (-6 V) and a width (t_w) of 500 μs . In this work, the device is programmed symmetrically, i.e., by applying V_w on the gate while grounding both the S and D contacts. The dynamic range (DR) is defined as the ratio of the source-to-drain resistance (R_{SD}) in the high resistive state (HRS) to the low resistive state (LRS). First, the influence of the channel length on the DR is examined by performing a statistical analysis on 120 devices to average out variations due to local fabrication inhomogeneities (see further discussion in section “Characteristics for ANNs”). The average DR and the distribution are represented as a function of L_{ch} in Fig. 1b. A detailed analysis of the DR dependence on L_{ch} is found in the Supplementary Note 5. We found that the two components of the source-to-drain resistance

$R_{\text{SD}} = R_{\text{ch}} + 2R_{\text{c}}$, the channel resistance R_{ch} and the contact resistance R_{c} , are modulated to a different extend by the write pulses. While R_{c} displays a modulation of $DR_{R_{\text{c}}} = 8.4$, the dynamic range of the channel $DR_{R_{\text{ch}}} = 2.8$ is three times less. Consequently, the total DR in Fig. 1b, that includes both contributions, decreases for longer channels as the relative contribution of $DR_{R_{\text{ch}}}$ increases. Moreover, an overall increase of the DR by a factor of ~ 2.8 was observed by increasing t_w to 500 ms, a first indication of another R_{SD} modulation than caused by the polarization switching at short t_w (500 μs).

The influence of t_w on the dynamic range was further analyzed by performing a sequence of 20 constant amplitude potentiation ($V_w = 6$ V) and depression ($V_w = -6$ V) cycles with changing t_w from 10 ns to 1 s (Fig. 2a). In Fig. 2b, the same 20 potentiation/depression cycles are superposed, to highlight the low cycle-to-cycle variability and better visualize the potentiation and depression shape. A dynamic range of almost 60 was reached (average of 20 cycles), a considerable increase as compared to the state-of-the-art oxide-channel FeFETs^{24,25}. The R_{SD} shows a strong dependence on t_w : there are two different regimes for both the potentiation and depression. For short pulses ($t_w > 3$ μs , $V_w = -6$ V) a steep depression is observed, followed by a less steep change ($t_w > 3$ μs , $V_w = -6$ V) that does not saturate up to the maximum t_w of 1 s. In the long pulse regime, cumulative switching is observed: repeating the same pulse (e.g., $t_w = 100$ ms, $V_w = -6$ V) 100 times increased R_{SD} after the first 10 cycles still

by a factor 2 (Supplementary Note 6). The absence of a saturation and a change of slope indicates that the resistance modulation for $t_w > 3 \mu\text{s}$ is based on an additional and slower physical process with respect to the previous regime ($t_w > 3 \mu\text{s}$): the first regime ($t_w > 3 \mu\text{s}$) is attributed to the ferroelectric switching as it is known to occur at very fast timescales³⁸. In the second regime ($t_w > 3 \mu\text{s}$), the energy of each pulse was potentially large enough to additionally enable oxygen migration^{33,39} between HZO and WO_x . The oxidation or reduction of the WO_x channel in turn adds to the modulation of R_{SD} .

The underlying conduction mechanisms of the two regimes are discussed next. Several electrode-limited and bulk-limited conduction mechanisms depend on temperature in different ways⁴⁰. Temperature-dependent $I_{\text{DS}} - V_{\text{DS}}$ measurements of the channel were performed ($20^\circ\text{C} - 60^\circ\text{C}$) after programming the device in the LRS ($V_w = 6 \text{ V}$) and in the HRS ($V_w = -6 \text{ V}$). From Fig. 2b, we observe that fast, ferroelectric effects occur already from pulse widths of 100 ns. The onset of the change of regime is observed for pulses of $3 \mu\text{s}$. To reveal that there are indeed two different mechanisms to modulate the resistance, the effect of the temperature on the DR was investigated at different timescales. Write pulses of $t_w = 500 \mu\text{s}$ (where the ferroelectric effect is fully saturated and oxygen migration just starts) and longer write pulse trains of $90 \cdot t_w = 100 \text{ ms}$ (to enhance the oxygen migration) were applied.

No difference in the conduction mechanism between the two timescales was observed, as discussed in Supplementary Note 7. Both the LRS and HRS show a linear $I_{\text{D}} - V_{\text{DS}}$ characteristic and are best fitted with the ohmic conduction model⁴¹.

In contrast, comparing the DR as a function of temperature for the two timescales (Fig. 3a, b) displays a clear difference. For the pulses where the ferroelectric effect dominates ($t_w = 500 \mu\text{s}$), the DR decreases with increasing temperature. This moderate effect can be attributed to a phase transition from the orthorhombic ferroelectric phase of HZO to its tetragonal, anti-ferroelectric phase at elevated temperatures, as observed in ZrO_2 ⁴² and Si:HfO_2 ⁴³. From $50^\circ\text{C} - 60^\circ\text{C}$, an increase in the dynamic range is observed, indicating an additional mechanism that modulates the resistance becoming dominant. This is explained by the chosen pulse width of $t_w = 500 \mu\text{s}$ that ensures saturation of the polarization but is also large enough for the second effect to become noticeable (predicted by Fig. 2b for pulses above $3 \mu\text{s}$). In contrast, for the slower timescale, an increase in the dynamic range with temperature is observed in the whole temperature range and is one to two orders of magnitude larger. From $50^\circ\text{C} - 60^\circ\text{C}$, the relative increase in the dynamic range is the highest. The temperature range was limited to $20^\circ\text{C} - 60^\circ\text{C}$ since above, irreversible changes of the gate were observed for some devices when applying a bias of 6 V. These observations are consistent with increased oxygen-ion mobility at elevated temperatures, which facilitates the channels oxidation or

reduction. In this scenario, the observed increase of the DR for $500 \mu\text{s}$ pulses above 50°C suggests a dual dependency of the oxygen-ion mobility on the temperature and on the pulse duration. The different types of behavior with temperature confirm the presence of two effects modulating our channel resistance. A schematic diagram illustrating the two effects can be found in Supplementary Fig. 10.

Retention measurements at elevated temperatures (85°C) and with long write pulses ($t_w = 500 \text{ ms}$) display a DR of almost three orders of magnitude, a direct consequence of the aforementioned increased oxygen-ion mobility (Fig. 3c). The larger the resistance change obtained within this retention study, the faster (hours to minutes) it drifted back. The improved retention of the intermediate states compared to the extreme states rules out the effect of a depolarization field, as observed by Muller et al. in $\text{TiN/SiHfO}_2/\text{TiN}$ capacitors⁴⁴. Furthermore, the observation that both the LRS and HRS relax towards intermediate resistances confirm that the dominant mechanism is not a depolarization field due to a limited number of carriers in the WO_x . Therefore, a resistance modulation by displacing oxygen-ions will eventually relax back to equilibrium by ion-diffusion processes, introducing the possibility to tune the plasticity timescale by controlling the oxygen-ion mobility. The more consecutive pulses are applied, the larger the resistance change and the longer the relaxation back to a stable long-term state. This characteristic can be used to implement short-term plasticity⁴⁵. In dedicated neuromorphic hardware, the above-mentioned relaxation occurs on long timescales (seconds, minutes, hours): this should then also hold for the stimuli or events processed by the network, as for example in the application field of health monitoring or autonomous driving. At shorter timescales, our ferroelectric synaptic functionalities ascribed to the existence of a threshold voltage for domain switching (coercive field), makes our devices good candidates for implementation in spiking neural networks. For example, spike-timing-dependent plasticity (STDP) was reported in perovskite⁴⁶ and hafnia⁴⁷ based thin films. Finally, the gradual and field-driven switching of the ferroelectric domains is well suited for implementing voltage-based STDP models⁴⁸⁻⁵⁰. In recent work, Garg et al.⁵¹ showed that such unsupervised, voltage-dependent-plasticity learning rules outperforms STDP in classification tasks.

Building on this finding of multiple timescales and physical locations where the R_{SD} modulation occurs, we envision adding a 4th electrode with a non-ferroelectric gate on top of the WO_x dielectric to further oxidize or reduce the channel without affecting the ferroelectric state. This additional terminal provides a volatile resistance component to tune the switching dynamics, similarly to⁵². Having such an additional volatile component permits the implementation of a wide range of neuromorphic engineering paradigms⁵³.

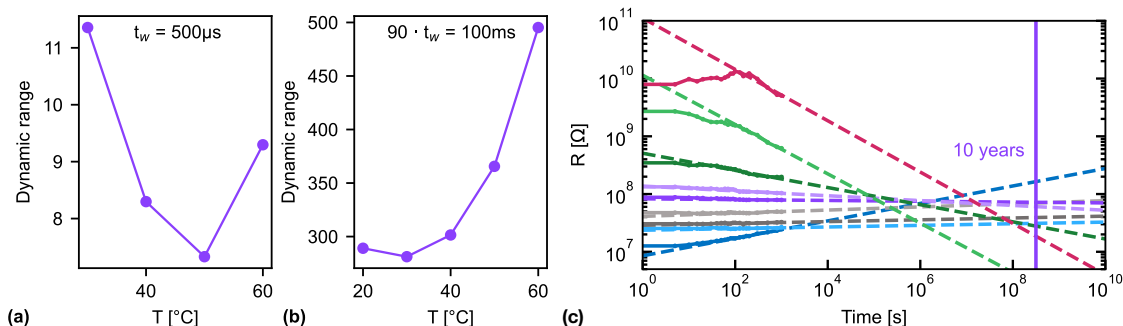


Fig. 3 Temperature-dependent measurements. Dynamic range modulation with temperature for **a** $t_w = 500 \mu\text{s}$ and **b** longer write pulse trains of $90 \cdot t_w = 100 \text{ ms}$. **c** Retention measurements at elevated temperature (85°C) for long set pulses $t_w = 500 \text{ ms}$.

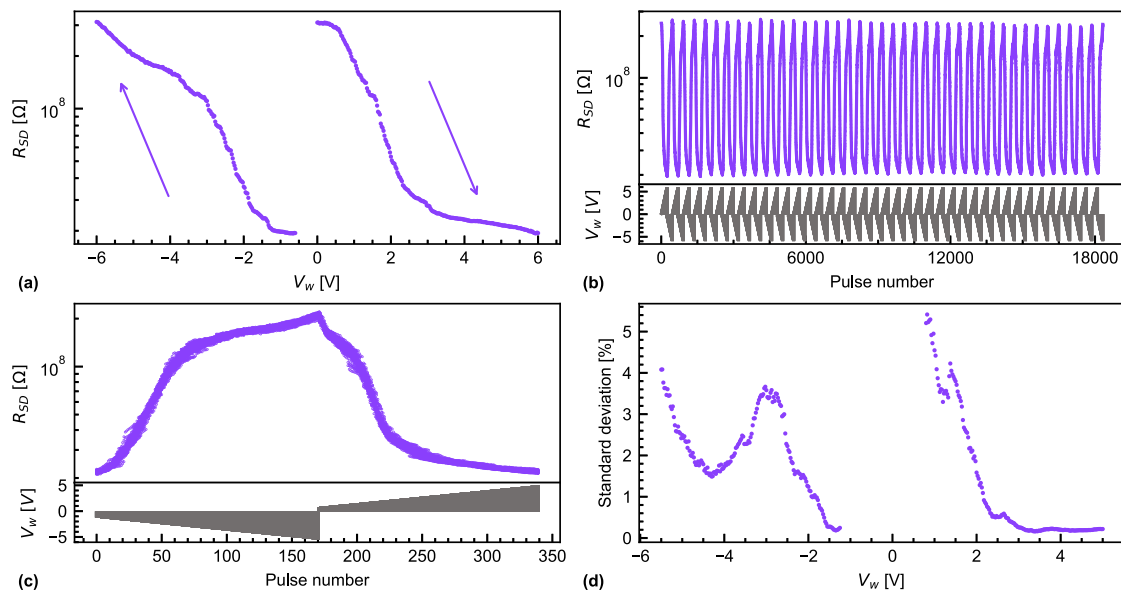


Fig. 4 Potentiation and depression cycles with a constant pulse width scheme. The FeFET has a channel length of 300 nm and a channel width of 2 μm : **a** one potentiation (0 V to 6 V) and depression (-0.6 V to -6 V) cycle with a constant $t_w = 500\ \mu\text{s}$ showing quasi-continuous channel resistance states (dep: 241 states, pot: 217). **b** 40 sub-range potentiation (0.6 V to 5 V, 171 levels) and depression (1.25 V to -5.5 V , 169 levels) cycles. Channel resistance R_{SD} as a function of pulse number (top) and the corresponding V_w (bottom) are reported. **c** Superposition of all 40 cycles from **(b)** to visualize the cycle-to-cycle variability. **d** Standard deviation of a channel resistance state normalized by the resistance window (HRS-LRS) for each V_w .

Characteristics for ANNs. In contrast to SNNs, where a tunable plasticity is desired, the efficient operation of ANNs on analog memristive crossbar arrays requires artificial synapses that behave as long-term memory with long data retention and low variability. Figure 4a shows a single potentiation and (depression) cycle of R_{SD} by increasing (decreasing) V_w from 0 V to 6 V (-0.6 V to -6 V) in 25 mV steps. The programming and R_{SD} measurement protocols are described in the section “Methods, Electrical characterization”; for a sampling bias of 200 mV, the conduction in the channel is Ohmic. The pulse duration t_w was kept constant at 500 μs . The 241 (217) steps for the potentiation (depression) exhibit a quasi-continuous resistance range with a monotonic change of the resistance and a DR of 16. The Cycle-to-Cycle Variability (CtCV) was analyzed by performing 40 sub-range cycles of depression (-0.6 V to -5 V) and potentiation (-1.25 V to -5.5 V) as shown in Fig. 4b. The corresponding V_w to each pulse is given at the bottom. Figure 4c is a superposition of all 40 cycles to help visualize the CtCV. By reducing the V_w range, the DR decreases to 10.4 on average. Figure 4d is a visualization of the CtCV (standard deviation) as a percentage of the channel resistance range (HRS-LRS). The CtCV does not exceed 6% and is as low as 1.9% on average.

To study the device-to-device variability, 20 identical FeFETs were measured and the normalized standard deviation of the HRS (39%), LRS (39%), and DR (28%) was extracted. The device-to-device variability can be explained by process variations across the sample. The polycrystalline nature of HZO results in different ferroelectric properties from device to device¹⁹. Also, the WO_x conductivity is rather sensitive to its oxygen content and can be reduced at elevated temperatures ($>250\text{ }^\circ\text{C}$) by other oxides or nitrides interfacing it, such as SiO_2 , Al_2O_3 , or SiN . Hence, local temperature differences during processing can cause different local reduction states of the WO_x .

The performance of our FeFETs as artificial synapses in a crossbar array was investigated by using the MLP+ NeuroSimV3.0⁵⁴ framework. The on-line learning accuracy of a pseudo-crossbar array of 400 input, 250 hidden, and 10 output neurons trained on the MNIST database was simulated by using

the aforementioned values. The non-linearity (NL) parameters were extracted by fitting the potentiation and depression curves according to ref. ⁵⁴ (Fig. 5a). They are 2.32 and -4.63 for the potentiation and depression, respectively. Moreover, Fig. 4a shows that operating the FeFET in a subloop regime ($\pm 3\text{ V}$) would improve the linearity even further, but at the cost of a smaller dynamic range. Here, the MLP+ NeuroSimV3.0 code was slightly adapted (Supplementary information Note 8) to apply a random Conductance Range Variation (CRV) to every device in the network, which takes into account our device-to-device variability. The spread of the HRS, LRS, and DR around the average (No var) are depicted in Fig. 5b, c. Figure 5d reports the learning accuracy on the MNIST database. By only considering the NL parameters and the Finite Number-of-States (FNoS) as a non-ideality, an excellent performance of 92% recognition accuracy is achieved. This metrics show that the proposed technology compares to state-of-the-art FeFET evaluated with the same platform (Supplementary Table 3). By further introducing the CtCV to the simulation, the performance remains as high as 89% and with the HRS, LRS, and DR spread included, still 88% is reached.

We then perform endurance and retention tests to evaluate the benefit of using a metallic oxide channel, compared to state-of-the-art FeFET on Si, for which is typically reported an endurance of 10^6 ¹⁹ cycles, although lately up to 10^8 ⁵⁵ cycles were reported. The endurance of a FeFET with $L = 800\text{ nm}$ and $W = 600\text{ nm}$ is shown in Fig. 6a. Cycling pulses of $\pm 3\text{ V}$ and $\pm 4\text{ V}$ with a frequency of 100 KHz were applied to the gate, while the source and drain were grounded. When switching at $\pm 3\text{ V}$ we observed a small continuous decay of the DR to $\sim 70\%$ of its initial value after 10^{10} cycles, but no failure could be identified. To the best of our knowledge this is the best endurance reported on hafnia-based FeFETs. Especially when compared with Si-based channel FeFETs this is a major improvement. Increasing the cycling voltage to $\pm 4\text{ V}$ (on another device with the same dimensions) accelerated the fatigue and the device failed after $8 \bullet 10^9$ cycles. Cycling at even higher fields was not tested as online learning happens in small changes and not by constantly switching between the extreme states.

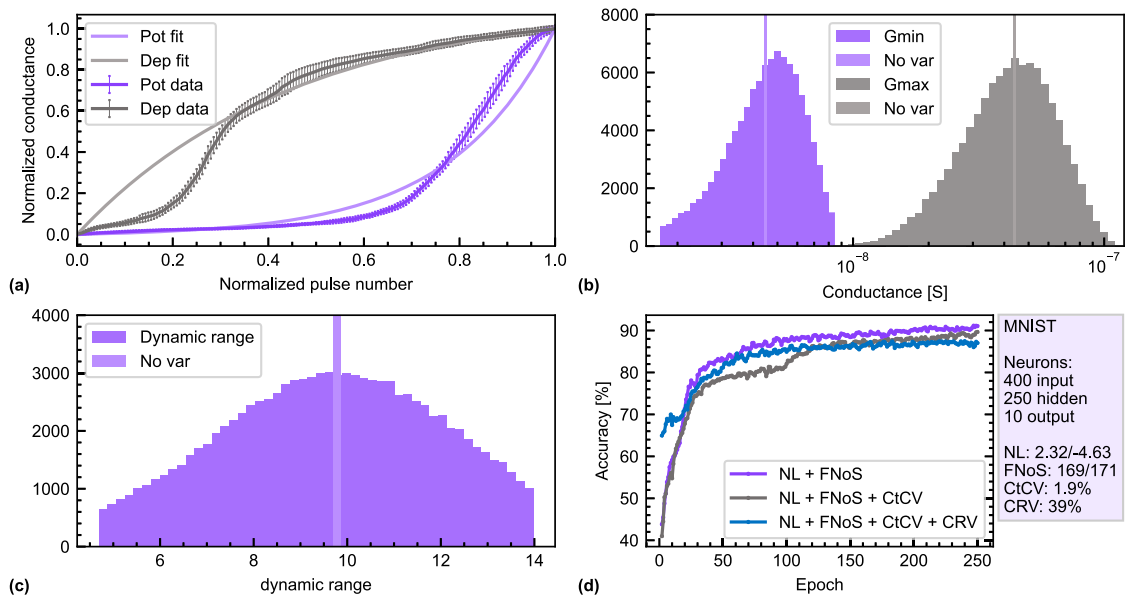


Fig. 5 Online learning performed with MLP+ NeuroSimV3.0. **a** Exponential fit of the potentiation and depression curves to extract the corresponding non-linearity parameter. **b, c** Histogram of the HRS, LRS, and DR after taking the device-to-device variability into account. **d** MNIST classification performance of our FeFETs with different degrees of non-idealities included: non-linearity factors and finite number of steps (purple), + cycle to cycle variation (gray), and + conductance range variation (blue).

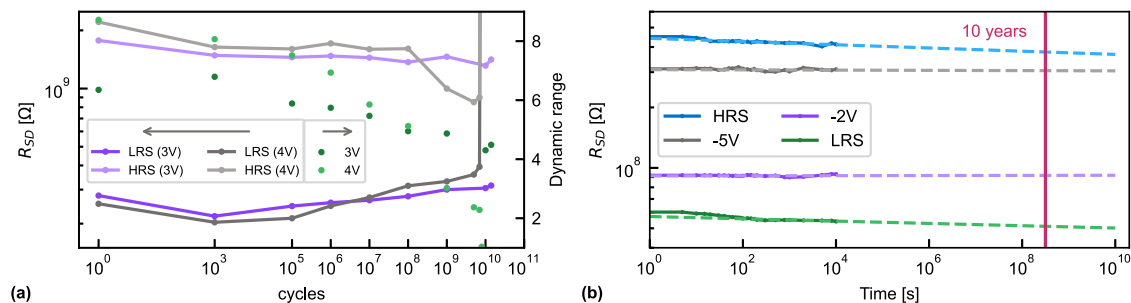


Fig. 6 Endurance and retention. **a** Endurance of a FeFET with $L = 800$ nm and $W = 600$ nm. Triangular pulses with a frequency of 100 kHz were applied up to 10^{10} cycles. The amplitude of the pulses was ± 3 V and ± 4 V. The evolution of the HRS and LRS (left axis) and the corresponding dynamic range (right axis) are shown. **b** Retention measurements at room temperature for a FeFET with $L = 300$ nm and $W = 2$ μ m showing a good retention of >10 years for the four programmed states. Only the HRS has a small drift. The solid lines are the experimental data and the dashed lines are linear extrapolations in the log-log scale.

With a channel resistance between 20 M Ω and 2 G Ω (depending on the geometry) and a read voltage of 100 mV, between 5 pW and 500 pW are dissipated during a read operation ($P_{\text{read}} = V_{\text{SD}}^2/R_{\text{SD}}$). The write operation of a single device ($t_{\text{read}} = 500$ μ s) to the high gate impedance (Supplementary Fig. 4b) has a lower energy consumption, between 1.7 fJ and 1.2 pJ for V_w of 1 V and 6 V, respectively ($E_{\text{write}} = t_w \cdot V_w^2/R_G$). The low-power operation, long endurance, and promising MNIST classification performance make this a viable candidate for large crossbar implementations for ANN training.

Retention measurements are depicted in Fig. 6b. By fitting a linear regression in the log-log scale, no drift is observed for the intermediate states. Both, the LRS and HRS display a small drift towards lower values. Extrapolating the fit to 10 years yields a change of about 11% and 15%, respectively. This excellent retention time confirms the advantage of using metal-oxide thin films over Si as channels as there is no back-switching of the ferroelectric domains due to charge trapping at the oxide interlayer formed between Si and the ferroelectric⁵⁶. This low drift opens the path to inference and memory applications for the FeFET devices presented in this study.

Conclusion

We demonstrated a scaled (sub- μ m) BEOL-compatible FeFET artificial synapse with an amorphous $\text{WO}_{x < 3}$ channel. The device concept was engineered to leverage two controllable resistance modulation mechanisms activated on two different write pulse timescales: a fast ferroelectric field effect ($t_w < 3$ μ s) and an oxidation/reduction of the channel by oxygen movement at a slower timescale ($t_w > 3$ μ s). Key enablers were the control of the channel oxidation state and of its thickness down to 4 nm, as well as having the channel in direct contact with the ferroelectric HZO gate without the formation of a spurious interlayer.

The dual nature of the resistance modulation mechanisms was derived from the write time-dependent dynamics and from the two different potentiation and depression slopes. The temperature-dependent current measurements showed opposite dynamic range trends for the two timescales, confirming that the resistance changes originate from two different mechanisms. Moreover, the temperature-dependent retention measurements highlighted the role of the oxygen drift across the layers in the slow regime ($t_w > 3$ μ s).

With this extra option to extend the dynamic range, our scaled FeFETs have a dynamic HRS/LRS ratio at room temperature that

is 30 times larger than the state-of-the-art²⁴. The plasticity of our synapse shows a different response over multiple timescales, making our FeFETs interesting candidates for neuromorphic engineering. The ohmic nature and a resistance of our WO_x channel in the MΩ regime are excellent features for precise and low-power readout operations. The extremely fine-grained, quasi-continuous monotonic resistance changes with more than 200 steps between the LRS and HRS, together with an excellent cycle-to-cycle variability led to a good MNIST classification accuracy of 88% with the NeuroSim framework. The endurance was extended to >10¹⁰ cycles and an excellent retention of >10 years was obtained with only little dynamic range loss. Therefore, our FeFET technology is not only promising for online learning but also for in-memory computing and neural network inference applications. A 4th electrode acting as top gate could introduce a volatile component, a desired property for bio-inspired neuromorphic engineering paradigms.

Methods

Sample preparation. Our FeFET is a back-gated device to allow the crystallization of the ferroelectric HZO gate dielectric before the deposition of the thin-film WO_x channel. First, 500 nm SiO₂ were formed by thermally oxidizing the Si substrate. A 100 nm thick W layer was deposited by sputtering and consecutively etched in a Reactive Ion Etcher (RIE) with an SF₆ plasma to form the first gate metal level (M1). Then, a 100 nm thick SiO₂ passivation was deposited by Plasma-Enhanced Chemical Vapor Deposition (PECVD) at 300 °C and vias to M1 were etched using an RIE with a CHF₃/O₂ plasma. M2 was deposited as M1 and continued by a 10 nm TiN layer deposited using a tetrakis-(dimethylamino)titanium (TDMAT) precursor and N₂/H₂ plasma in an Oxford Instruments plasma-enhanced atomic layer deposition (PEALD) system. To avoid large topographies, 600 nm of SiO₂ were deposited on top of M2 and then removed by chemical mechanical polishing (CMP) until only a thin layer of SiO₂ was left above M2. This effectively removed the topography introduced by M1 and M2 (Supplementary Fig. 3d). The last few nm of SiO₂ were then etched by RIE to expose the TiN on top of M2. An ~10 nm thick layer of HZO was grown in a process using alternating cycles of tetrakis-(ethylmethylamino)-hafnium (TEMAH) and bis-(methyl-η5-cyclopentadienyl)-methoxymethyl-zirconium (ZrCMMM) at 300 °C and then capped by 30 nm of W deposited by sputtering. The crystallization of the HZO was then performed by a millisecond flash lamp anneal (ms-FLA)³⁰ where the sample was heated to 375 °C. After the crystallization the W was removed by a wet etch in H₂O₂ at 50 °C. 4 nm of WO_x was then deposited using a bis-(tert-butylimino)-bis-(dimethylamino)-tungsten precursor and an oxygen plasma at 375 °C in a PEALD system. To fully oxidize the WO_x, it was annealed in a rapid thermal annealer at 350 °C with 50 sccm O₂ for 6 min. After the structuring of the WO_x by an SF₆ plasma in the RIE, source (S) and drain (D) contacts (Pt/W) were deposited on top of the WO_x channel by lift-off. Finally, the structures were passivated by 100 nm of SiO₂ and two metal levels (M3, M4) were fabricated as M1 and M2.

Structural characterization. Grazing-Incidence X-Ray Diffraction (GIXRD) measurements were executed on a Bruker D8 Discover diffractometer equipped with a copper rotating anode generator. Cross-sectional cuts and lamellas for Scanning Transmission Electron Microscope (STEM) analyses were prepared by Focused Ion Beam (FIB) using an FEI Helios NanoLab 450 S. STEM analysis was carried out on a double spherical aberration-corrected JEOL JEM-ARM200F microscope. Bright-Field STEM (BF-STEM) images were acquired at 200 kV, and energy-dispersive X-ray spectroscopy (EDS) line profiles were performed using a liquid-nitrogen-free silicon drift detector.

Electrical characterization. The electrical characterization was performed using a probe station in atmosphere. For the temperature-dependent measurements, the stage was heated and the temperature controlled by a thermocouple. Otherwise, the measurements were performed at room temperature. Potentiation and depression, resistance, retention, endurance, and capacitance measurements were performed on an Agilent B1500A with a B1530A waveform generator/fast measurement unit and a B1520A multi-frequency capacitance measurement unit. Write pulses were generated by a remote sense unit module close to the probe and applied to the gate while source and drain were grounded. To read the channel resistance, four sampling points at V_{DS} = 200 mV and V_{DS} = -200 mV were taken with the source measurement unit at the source while the drain was grounded and the gate was floating. PUND measurements on MFSM capacitors were performed on a TF analyzer 2000 from aixACCT. The PUND signal of ±5 V at 1 kHz was applied to the W/TiN contact (equivalent of the gate in a FeFET) of the W/TiN/HZO/WO_x/W capacitor, while the other side was grounded. Before the PUND

measurement, devices were woken up by 10⁵ fully switching cycles with an amplitude of ±4 V.

Data availability

The data that support the findings of this study have been included in the manuscript and supplementary information. Any additional data are available from the corresponding author upon reasonable request.

Received: 25 April 2022; Accepted: 6 February 2023;

Published online: 17 February 2023

References

- Völske, M. et al. Web Archive Analytics. In *INFORMATIK 2020* (eds. Reussner, R. H., Koziol, A. & Heinrich, R.) 61–72 https://doi.org/10.18420/inf2020_05 (Gesellschaft für Informatik, Bonn, 2021).
- Umair, M., Cheema, M. A., Cheema, O., Li, H. & Lu, H. Impact of COVID-19 on IoT adoption in healthcare, smart homes, smart buildings, smart cities, transportation and industrial IoT. *Sensors* **21**, 3838 (2021).
- Reinsel, D., Rydning, J. & Gantz, J. Worldwide Global DataSphere Forecast, 2021–2025: the world keeps creating more data — now, what do we do with it all? *Doc # US46410421*. Available at: <https://www.idc.com/getdoc.jsp?containerId=US46410421> (2021).
- Abdo, G. & Mahale, V. Worldwide managed edge services forecast, 2021–2025. *Doc # US47308121* Available at: <https://www.idc.com/getdoc.jsp?containerId=US47308121> (2021).
- Yang, R. In-memory computing with ferroelectrics. *Nat. Electron.* **3**, 237–238 (2020).
- Christensen, D. V. et al. 2022 roadmap on neuromorphic computing and engineering. **2**, 022501 (2021).
- Feldman, D. E. Synaptic mechanisms for plasticity in neocortex. *Annu. Rev. Neurosci.* **32**, 33–55 (2009).
- Kasabov, N., Dhoble, K., Nuntalid, N. & Indiveri, G. Dynamic evolving spiking neural networks for on-line spatio- and spectro-temporal pattern recognition. *Neural Networks* **41**, 188–201 (2013).
- Beilliard, Y. & Alibart, F. Multi-terminal memristive devices enabling tunable synaptic plasticity in neuromorphic hardware: a mini-review. *Front. Nanotechnol.* **3**, 1–9 (2021).
- Youngblood, N. et al. Tunable volatility of Ge 2 Sb 2 Te 5 in integrated photonics. *Adv. Funct. Mater.* **29**, 1807571 (2019).
- Diehl, P. U. & Cook, M. Unsupervised learning of digit recognition using spike-timing-dependent plasticity. *Front. Comput. Neurosci.* **9**, 1–9 (2015).
- Payvand, M., Nair, M. V., Müller, L. K. & Indiveri, G. A neuromorphic systems approach to in-memory computing with non-ideal memristive devices: from mitigation to exploitation. *Faraday Discuss.* **213**, 487–510 (2019).
- Indiveri, G., Linares-Barranco, B., Legenstein, R., Deligeorgis, G. & Prodromakis, T. Integration of nanoscale memristor synapses in neuromorphic computing architectures. *Nanotechnology* **24**, 384010 (2013).
- Kub, F. J., Moon, K. K., Mack, I. A. & Long, F. M. Programmable analog vector-matrix multipliers. *IEEE J. Solid-State Circuits* **25**, 207–214 (1990).
- Zhou, Y. et al. A reconfigurable Two-WSe 2 -transistor synaptic cell for reinforcement learning. *Adv. Mater.* **34**, e2107754 (2022).
- Xue, F. et al. Optoelectronic ferroelectric domain-wall memories made from a single van der waals ferroelectric. *Adv. Funct. Mater.* **30**, 2004206 (2020).
- Han, H., Yu, H., Wei, H., Gong, J. & Xu, W. Recent progress in three-terminal artificial synapses: from device to system. *Small* **15**, 1900695 (2019).
- Yu, S. et al. Ferroelectric HfO 2 -based synaptic devices: recent trends and prospects. *Semicond. Sci. Technol.* **36**, 104001 (2021).
- Mulaosmanovic, H. et al. Ferroelectric field-effect transistors based on HfO 2: a review. *Nanotechnology* **32**, 502002 (2021).
- Jerry, M. et al. Ferroelectric FET analog synapse for acceleration of deep neural network training. in *2017 IEEE International Electron Devices Meeting (IEDM)* **6**, 6.2.1-6.2.4 (IEEE, 2017).
- Mulaosmanovic, H. et al. Switching kinetics in nanoscale hafnium oxide based ferroelectric field-effect transistors. *ACS Appl. Mater. Interfaces* **9**, 3792–3798 (2017).
- Dunkel, S. et al. A FeFET-based super-low-power ultra-fast embedded NVM technology for 22nm FDSOI and beyond. in *2017 IEEE International Electron Devices Meeting (IEDM)* **1**, 19.7.1-19.7.4 (IEEE, 2017).
- Breyer, E. T., Mulaosmanovic, H., Mikolajick, T. & Slesazek, S. Perspective on ferroelectric, hafnium oxide based transistors for digital beyond von-Neumann computing. *Appl. Phys. Lett.* **118**, 050501 (2021).

24. Halter, M. et al. Back-end, CMOS-compatible ferroelectric field-effect transistor for synaptic weights. *ACS Appl. Mater. Interfaces* **12**, 17725–17732 (2020).
25. Mo, F. et al. Experimental demonstration of ferroelectric HfO₂ FET with ultrathin-body IGZO for high-density and low-power memory application. In: 2019 Symposium on VLSI Technology T42–T43 <https://doi.org/10.23919/VLSIT.2019.8776553> (IEEE, 2019).
26. Kim, M.-K. & Lee, J.-S. Ferroelectric analog synaptic transistors. *Nano Lett.* **19**, 2044–2050 (2019).
27. Mulaosmanovic, H., Chicca, E., Bertele, M., Mikolajick, T. & Slesazek, S. Mimicking biological neurons with a nanoscale ferroelectric transistor. *Nanoscale* **10**, 21755–21763 (2018).
28. CHARLTON, M. G. Hydrogen reduction of tungsten trioxide. *Nature* **169**, 109–110 (1952).
29. Ingham, B., Hendy, S. C., Chong, S. V. & Tallon, J. L. Density-functional studies of tungsten trioxide, tungsten bronzes, and related systems. *Phys. Rev. B* **72**, 075109 (2005).
30. O'Connor, É. et al. Stabilization of ferroelectric Hf_xZr_{1-x}O₂ films using a millisecond flash lamp annealing technique. *APL Mater.* **6**, 121103 (2018).
31. Haubner, R., Schubert, W. D., Lassner, E. & Lux, B. Influence of aluminum on the reduction of tungsten oxide to tungsten powder. *Int. J. Refract. Hard Met.* **6**, 161–167 (1987).
32. Aird, A., Domeneghetti, M. C., Mazzi, F., Tazzoli, V. & Salje, E. K. H. Sheet superconductivity in: crystal structure of the tetragonal matrix. *J. Phys. Condens. Matter* **10**, L569–L574 (1998).
33. He, X. et al. Memristive properties of hexagonal WO₃ nanowires induced by oxygen vacancy migration. *Nanoscale Res. Lett.* **8**, 50 (2013).
34. Georg, A., Graf, W. & Wittwer, V. Comparison of electrical conductivity and optical properties of substoichiometrically and electrochemically coloured WO_x films of different crystallinity. *Sol. Energy Mater. Sol. Cells* **51**, 353–370 (1998).
35. Chien, W. C. et al. A study of the switching mechanism and electrode material of fully CMOS compatible tungsten oxide ReRAM. *Appl. Phys. A Mater. Sci. Process.* **102**, 901–907 (2011).
36. Wen, X., Halter, M., Bégon-Lours, L. & Luisier, M. Physical modeling of HZO-based ferroelectric field-effect transistors with a WO_x channel. *Front. Nanotechnol.* **4**, 1–11 (2022).
37. Gokmen, T. & Vlasov, Y. Acceleration of deep neural network training with resistive cross-point devices: design considerations. *Front. Neurosci.* **10**, 1–13 (2016).
38. Li, J. et al. Ultrafast polarization switching in thin-film ferroelectrics. *Appl. Phys. Lett.* **84**, 1174–1176 (2004).
39. Xiao, Z. et al. Controlling resistance switching performances of Hf_{0.5}Zr_{0.5}O₂ films by substrate stress and potential in neuromorphic computing. *Adv. Intell. Syst.* **4**, 2100244 (2022).
40. Chiu, F.-C. A review on conduction mechanisms in dielectric films. *Adv. Mater. Sci. Eng.* **2014**, 1–18 (2014).
41. Sze, S. M. & Ng, K. K. Physics of semiconductor devices. *America* **10**, (John Wiley & Sons, Inc., 2006).
42. Müller, J. et al. Ferroelectricity in simple binary ZrO₂ and HfO₂. *Nano Lett.* **12**, 4318–4323 (2012).
43. Böscke, T. S. et al. Phase transitions in ferroelectric silicon doped hafnium oxide. *Appl. Phys. Lett.* **99**, 112904 (2011).
44. Mueller, S., Muller, J., Schroeder, U. & Mikolajick, T. Reliability characteristics of ferroelectric Si:HfO₂ thin films for memory applications. *IEEE Trans. Device Mater. Reliab.* **13**, 93–97 (2013).
45. Ilyas, N. et al. Analog switching and artificial synaptic behavior of Ag/SiO_x:Ag/TiO_x/p++-Si memristor device. *Nanoscale Res. Lett.* **15**, 30 (2020).
46. Boyn, S. et al. Learning through ferroelectric domain dynamics in solid-state synapses. *Nat. Commun.* **8**, 14736 (2017).
47. Max, B., Hoffmann, M., Mulaosmanovic, H., Slesazek, S. & Mikolajick, T. Hafnia-based double-layer ferroelectric tunnel junctions as artificial synapses for neuromorphic computing. *ACS Appl. Electron. Mater.* **2**, 4023–4033 (2020).
48. Clopath, C., Büsing, L., Vasilaki, E. & Gerstner, W. Connectivity reflects coding: a model of voltage-based STDP with homeostasis. *Nat. Neurosci.* **13**, 344–352 (2010).
49. Ziegler, M., Riggert, C., Hansen, M., Bartsch, T. & Kohlstedt, H. Memristive hebbian plasticity model: device requirements for the emulation of hebbian plasticity based on memristive devices. *IEEE Trans. Biomed. Circuits Syst.* **9**, 197–206 (2015).
50. Diederich, N., Bartsch, T., Kohlstedt, H. & Ziegler, M. A memristive plasticity model of voltage-based STDP suitable for recurrent bidirectional neural networks in the hippocampus. *Sci. Rep.* **8**, 9367 (2018).
51. Garg, N. et al. Voltage-dependent synaptic plasticity: unsupervised probabilistic Hebbian plasticity rule based on neurons membrane potential. *Front. Neurosci.* **16**, 1–12 (2022).
52. He, C. et al. Artificial synapse based on van der waals heterostructures with tunable synaptic functions for neuromorphic computing. *ACS Appl. Mater. Interfaces* **12**, 11945–11954 (2020).
53. Xi, F. et al. Artificial synapses based on ferroelectric schottky barrier field-effect transistors for neuromorphic applications. *ACS Appl. Mater. Interfaces* **13**, 32005–32012 (2021).
54. Chen, P.-Y., Peng, X. & Yu, S. NeuroSim+: an integrated device-to-algorithm framework for benchmarking synaptic devices and array architectures. In: 2017 IEEE International Electron Devices Meeting (IEDM) 6.1.1–6.1.4 <https://doi.org/10.1109/IEDM.2017.8268337> (IEEE, 2017).
55. Bae, J. H. et al. Highly scaled, high endurance, Ω -gate, nanowire ferroelectric FET memory transistors. *IEEE Electron Device Lett.* **41**, 1637–1640 (2020).
56. Yurchuk, E. et al. Origin of the endurance degradation in the novel HfO₂-based 1T ferroelectric non-volatile memories. *IEEE Int. Reliab. Phys. Symp. Proc.* **2**, 1–5 (2014).

Acknowledgements

We thank the Cleanroom Operations Team of the Binnig and Rohrer Nanotechnology Center (BRNC) for their help and support. This work is funded by H2020: ULPEC (no. 732642) and BeFerroSynaptic (no. 871737).

Author contributions

L.B.L. contributed to the temperature-dependent conduction mechanism measurements and interpretation, to manuscript revision, and to technical guidance; M.S. contributed to the STEM analysis; Y.P. and U.D. contributed by doing CMP; V.B. contributed with result interpretation and manuscript revision; B.J.O. contributed with manuscript revision and technical guidance; M.H. contributed with the design and fabrication of the devices, with the acquisition and interpretation of all not yet mentioned data, adaption of software code and writing 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/s43246-023-00342-x>.

Correspondence and requests for materials should be addressed to Mattia Halter.

Peer review information Communications Materials thanks Kai Ni and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Aldo Isidori.

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

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



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

© The Author(s) 2023