NANO LETTERS

pubs.acs.org/NanoLett

Letter

Restricted Boltzmann Machines Implemented by Spin−Orbit Torque Magnetic Tunnel Junctions

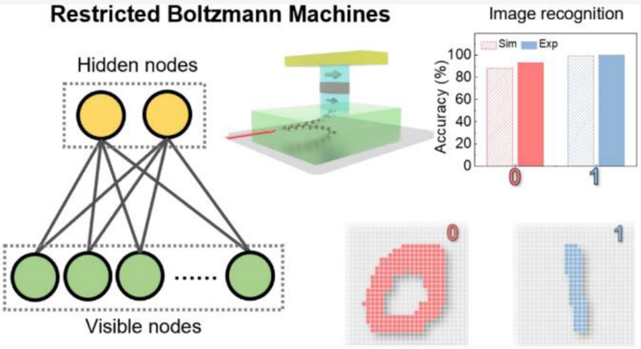
Xiaohan Li,[#] Caihua Wan,[*,#] Ran Zhang,[#] Mingkun Zhao, Shilong Xiong, Dehao Kong, Xuming Luo, Bin He, Shiqiang Liu, Jihao Xia, Guoqiang Yu, and Xiufeng Han[*]

Cite This: https://doi.org/10.1021/acs.nanolett.3c04820

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Artificial intelligence has surged forward with the advent of generative models, which rely heavily on stochastic computing architectures enhanced by true random number generators with adjustable sampling probabilities. In this study, we develop spin−orbit torque magnetic tunnel junctions (SOT-MTJs), investigating their sigmoid-style switching probability as a function of the driving voltage. This feature proves to be ideally suited for stochastic computing algorithms such as the restricted Boltzmann machines (RBM) prevalent in pretraining processes. We exploit SOT-MTJs as both stochastic samplers and network nodes for RBMs, enabling the implementation of RBM-based neural networks to achieve recognition tasks for both handwritten and spoken digits. Moreover, we further harness the weights derived from the preceding image and speech training processes to facilitate cross-modal learning from speech to image generation. Our results clearly demonstrate that these SOT-MTJs are promising candidates for the development of hardware accelerators tailored for Boltzmann neural networks and other stochastic computing architectures.



KEYWORDS: *restricted Boltzmann machines, spin−orbit torque, magnetic tunnel junctions, true random number generators*

Rapid advancements in artificial intelligence (AI) have led to the emergence of powerful generative models, essentially large-scale stochastic neural networks underpinned by probability and statistics principles.[1−12] Despite their remarkable capabilities, training these models is a computationally intensive task, relying heavily on an advanced complementary-metal-oxide semiconductor (CMOS) based central processing unit (CPU) or graphics processing unit (GPU) hardware.[13−15] As the evolution of conventional CMOS technology approaches the Moore's Law limit and the classical von Neumann computing architecture grapples with the significant memory wall issue, the progressive trajectory of generative AI is expected to be revolutionized by emerging nonvolatile memory technologies.[16−19]

Nonvolatile random-access memory (RAM) technologies such as resistive RAM,[20−22] phase change RAM,[23] magnetoresistive RAM (MRAM),[24] and ferroelectric RAM[25] have been proven exceptionally suitable and successful in hardware-accelerated matrix multiplication, a burdensome operation predominant in AI training and computing, owing to their nonvolatility and crossbar structure. Nevertheless, matrix multiplication is just one side of the coin in stochastic neural networks. Stochastic sampling forms the other vital operation aligning with the stochastic computing architecture inherent to generative neural networks.[1,2,26−31]

Boltzmann machines and restricted Boltzmann machines (RBMs) are a standout among stochastic computing algorithms to implement generative neural networks known for their effectiveness in unsupervised classification and recognition tasks and widely employed for pretraining unannotated data.[32−34] In the RBM algorithm, besides the matrix multiplication, a stochastic sampling process named as the Gibbs sampling plays a weighty role, which has to stochastically sample the state of network accurately according to a probability-distribution-function (PDF) predefined by its network parameters. Only if this stochastic sampling process is precisely executed can the energy of the network be assured to converge toward its global minimum. Conventionally, this sampling procedure involves generation of many pseudorandom numbers calculated by the CPU according to the PDFs, which is undeniably lengthy and inefficient. Alternatively, novel materials such as two-dimensional materials or memristor, and

Received: December 8, 2023
Revised: April 22, 2024
Accepted: April 24, 2024

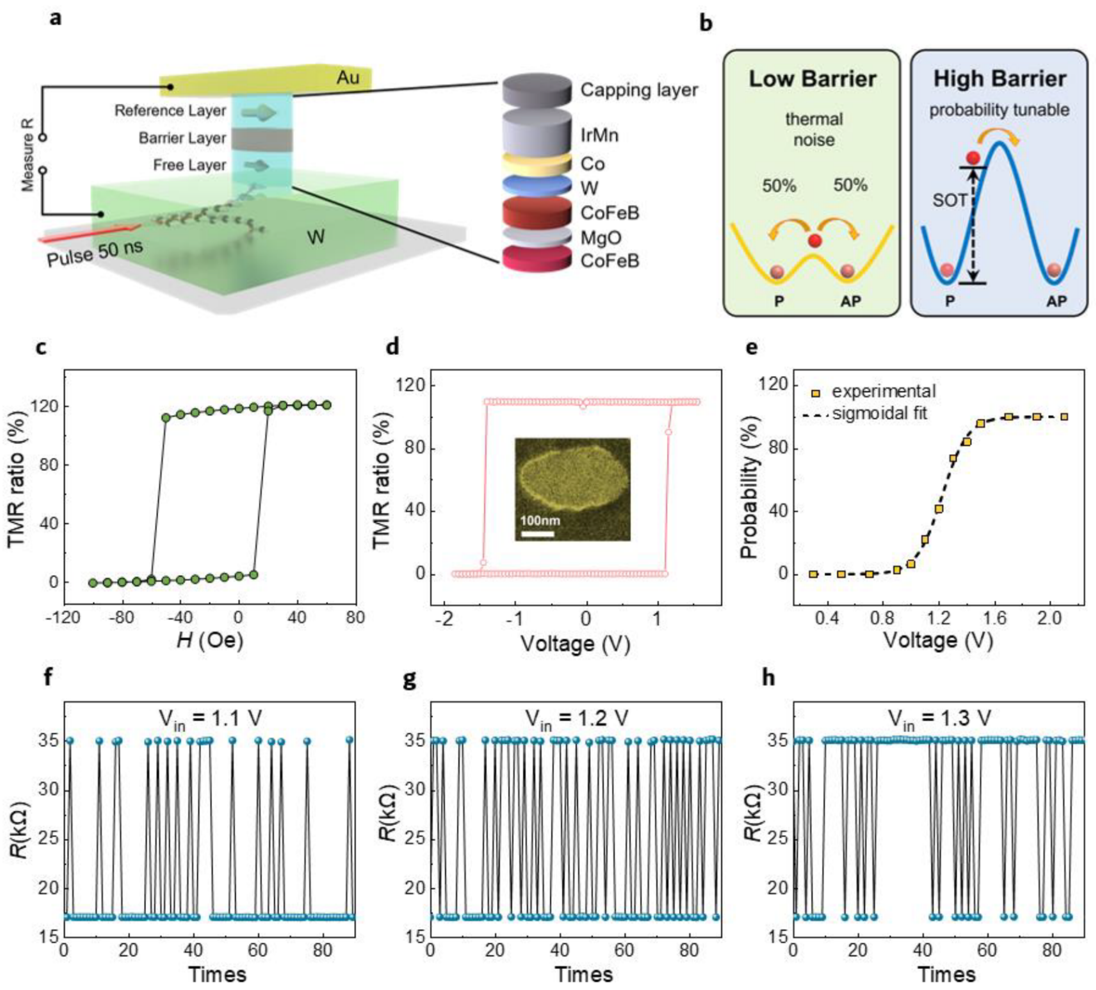


Figure 1. Stochastic switching performance of SOT-MTJs. (a) The stack structure of SOT-MTJs. (b) The schematic energy profile of low barrier and high barrier MTJs. (c) $R-H$ loop obtained by scanning an in-plane magnetic field along the easy axis of the MTJ. (d) Field-free magnetization switching driven by 50 ns pulse currents. The inset is the top view of the scanning electron microscope (SEM) image of an MTJ. (e) Switching probability (P) as a function of V_{in} and its fitting result with a well-matched sigmoid curve. (f–h) Switching results obtained from continuous testing at voltages of 1.1 V (f), 1.2 V (g), and 1.3 V (h).

their stochastic dynamics, are adopted for the sampling procedure in hardware.[35,36]

Especially, nonvolatile spintronic devices like magnetic tunnel junctions (MTJs) with high speed and endurance could potentially be a game-changer[37−43] despite their rareness in this field before. The intrinsically stochastic spin dynamics of MTJs,[44−46] especially those employing SOT,[47−52] make them competitive true random number generators (TRNGs),[53−56] thereby extending their utility to generative AI applications. Here, we aim to address the stochastic sampling challenge in spintronic hardware, enhance the efficiency of the Gibbs sampling operation in RBM by leveraging high-performance SOT-MTJs, and experimentally demonstrate the high compatibility of SOT-MTJs with the needs of RBM. To begin, we characterized the stochastic switching performance of SOT-MTJs, achieving a probability–voltage ($P-V$) tunability that aligns well with the desired sigmoid function. The SOT-MTJs were then deployed as both network nodes and stochastic samplers in image and speech recognition and even cross-modal learning tasks.[57] In addition, we explored the SOT-MTJ based RBMs in various fields including AI generation, image recovery, data encryption, integer factorization, and reversible logic. Our study clearly manifests that the implementation of Gibbs sampling using SOT-MTJs can significantly expedite the development of generative neural networks via spintronic hardware.

The MTJ device, as shown in Figure 1a, encompasses a W(3)/$Co_{20}Fe_{60}B_{20}$(1.4)/ MgO(1.5)/$Co_{20}Fe_{60}B_{20}$(3)/W(0.4)/Co(2.7)/$Ir_{25}Mn_{75}$(10)/Ru(4 nm) nanopillar and Au electrodes. It has a well-defined cross-section, measuring approximately 330 × 170 nm, akin to an ellipse as shown in the top view (Figure 1d inset), thus enhancing the in-plane uniaxial magnetic anisotropy along the ellipse's long axis. We patterned a 600 nm wide writing channel along the short axis of the MTJ, facilitating the Y-type SOT switching mode[58,59] in our in-plane MTJs.

When a pulse current (50 ns) is channeled through the writing channel (W), the magnetization of the bottom $Co_{20}Fe_{60}B_{20}$ (1.4) free layer adjacent to W is subjected to a spin–orbit torque (SOT) due to the spin Hall effect and the induced spin current. The magnetization can subsequently be manipulated or even switched if the pulse current is of sufficient strength. The magnetic configurations of the MTJ device, that is, the parallel (P) or antiparallel (AP) alignment

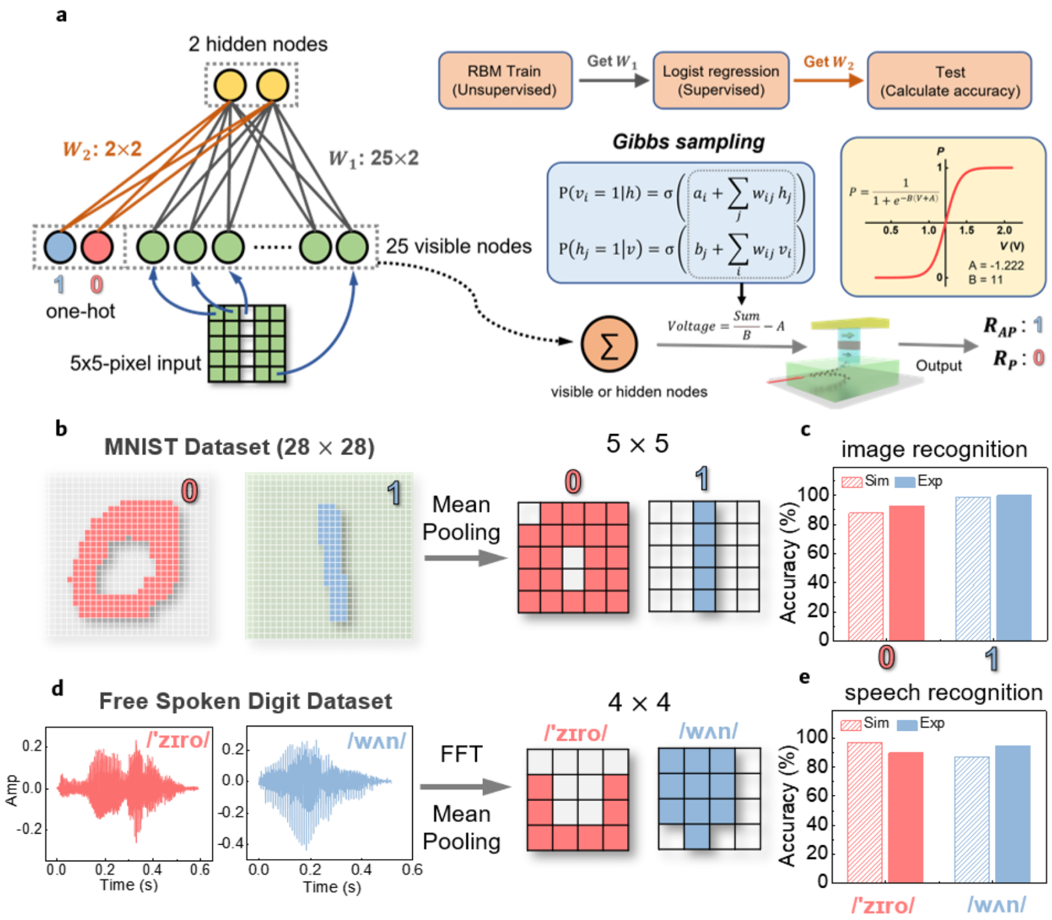


Figure 2. Image and speech recognition based on SOT-MTJ-RBMs in experiments. (a) A typical RBM structure as an example with 25 visible nodes (v_i) and 2 hidden nodes (h_j). The training process is divided into three parts, RBM training (unsupervised), logistic regression (supervised), and a final test (calculating accuracy). The Gibbs sampling of each node can be achieved with SOT-MTJs. Specifically, the probabilistic sampling $P(v_i = 1|h)$ and $P(h_j = 1|v)$ according to the sigmoid function $\sigma(\cdot)$ can be directly replaced by simple switching operations of SOT-MTJs. The conditional probability P of the visible (hidden) nodes depends on the overall weighted influence from their connected hidden (visible) nodes. The dependence complies with the sigmoid function. The parameters A and B here are used for renormalization. (b) "0" and "1" examples of MNIST data sets and final training data after mean pooling. (d) "0" and "1" examples of Free-Spoken-Digit-Data sets and final training data after fast Fourier transform (FFT) and mean pooling. (c and e) The experiment and simulation accuracy of image recognition (c) and speech recognition (e).

of the free layer relative to the top pinned $Co_{20}Fe_{60}B_{20}$ layer, can be probed by the low or high junction resistance of the MTJ device, respectively. For instance, the MTJ resistance can be switched by an external magnetic field or the pulse current, as illustrated in Figure 1c,d. The tunneling magnetoresistance (TMR) ratio is over 100%, indicating the high MgO barrier quality. The critical voltage for the AP to P transition is −1.2 V, and for the P to AP transition, it is 1.3 V. This voltage range can be managed by advanced CMOS technology.[60] The corresponding current density is approximately 8.3×10^6 A/cm^2 @ 50 ns, consistent with the levels reported for W/$Co_{20}Fe_{60}B_{20}$ systems.[58] These high-performance SOT-MTJ devices thus present an exceptional platform for stochastic computing applications, as made evident by our experimental demonstrations below.

In our exploration, the SOT-MTJ devices serve the elementary function of a Bernoulli true random number generator (TRNG) for stochastic sampling, acting as a binary TRNG with a tunable probability P or $(1 - P)$ to sample 1 or 0, respectively. Typically, MTJ-based TRNGs can be categorized into two types: low-barrier[44,61−69] and high-barrier systems (Figure 1b). The former features a barrier comparable to k_BT (low thermal stability factor), and hence, its magnetic configuration or junction resistance can stochastically flip between the P and AP states due to thermal fluctuations. Without external controls, both states have an occurrence probability of 50%.

On the other hand, the high-barrier TRNG possesses a much higher thermal stability factor. Consequently, only when an external stimulus activates a high-barrier MTJ to a specific critical state can thermal fluctuations significantly influence the stochastic switching dynamics. In this case, we can tune the switching probability (P) continuously by modulating the amplitude of the external stimuli (in this instance, spin−orbit torque), as illustrated in Figure 1e. Figure 1f−h demonstrates the occurrence of the P and AP states at three typical voltages (1.1, 1.2, and 1.3 V) corresponding to P = 20%, 50%, and 80%, respectively, directly reflecting the P-tunability. Furthermore, the $P - V$ dependency of this SOT-MTJ aligns perfectly with the sigmoid function (Figure 1e), endorsing its application as a Bernoulli TRNG in the Boltzmann machines, an efficient stochastic computing architecture. Moreover, a high barrier

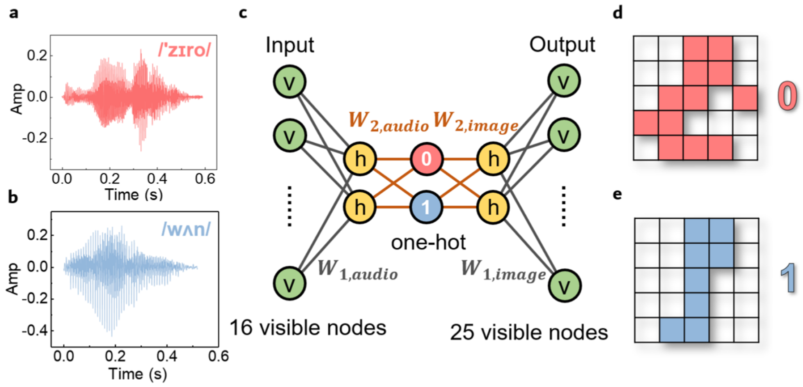


Figure 3. Experimental implementation of image and speech crossmodal learning on a merged RBM. (a and b) Presentation of audios for zero (a) and one (b) in the Free-Spoken-Digit-Data set. (c) Two RBMs are seamlessly connected through a labeled one-hot layer. The left input layer contains audio information, and the rightmost layer outputs images, and vice versa. (d and e) Using the network in (c) to generate images of 0 (d) and 1 (e) after sourcing the network with speech data in (a) and (b), respectively.

renders our SOT-MTJs nonvolatile, a feature that is also essential for implementing Boltzmann machines.[70]

Boltzmann machines (BMs) offer a potent stochastic computing architecture that is applicable to optimization, classification, and recognition tasks. They possess a network structure typically comprising numerous binary neurons (x_i) interconnected mutually by a symmetric weight ($w_{ij} = w_{ji}$), which parametrizes their interaction. The node x_i can be only 0 or 1. A BM is assigned a global energy $E = -\sum w_{ij}x_ix_j$ to quantify its overall interactions, which is physically similar to that of the Ising model. Here, bias terms b_ix_i are realized by adding a constant node ($x_0 = 1$) to the network. The occurrence probability (P) of a state depends on its nominal energy via $P \propto \exp(-E/T)$, mimicking a physical system adhering to the Boltzmann distribution. The BM can ideally settle down to its global minimum when it follows certain rules to update. One such rule is the Gibbs sampling, wherein the probability of a neuron (ith, for example) being 0 or 1 depends on the combined influence of other connected neurons. This influence is expressed in terms of probability: the probability of sampling the ith node into 1 should precisely align with the sigmoid-style PDF $P(x_i = 1) = \sigma(-\Delta E_i/T) = 1/[1 + \exp(-\Delta E_i/T)]$ with $\Delta E_i \equiv E_i(x_i = 0) - E(x_i = 1)$ the energy difference as $x_i = 0$ or 1 and T a renormalized temperature for tuning noise level. This Gibbs sampling rule ensures a lower energy state more probabilistically favorable and improves the chance for a network to converge to its global optimal. As demonstrated below, SOT-MTJs are well qualified to implement the Gibbs sampling for RBMs.

The restricted Boltzmann machine (RBM) is a simplified version of the BM, offering enhanced parallelism. Figure 2a provides a schematic illustration of the RBM network structure and its implementation. The RBM includes a visible layer with 25 nodes and a hidden layer with 2 nodes. The visible (hidden) layer is used for inputting/outputting image or speech data (storing latent recognition data). These two layers are interconnected with a weight matrix $\mathbf{W}_1$, containing 25 × 2 variables.

For image recognition, we initially input 5 × 5 pixels data ($\mathbf{V}$) of an image (after mean pooling from a 28 × 28 image in the MNIST Data set) into the visible layer. We then employ the Gibbs rule to sample the hidden layer. Before sampling, we execute matrix multiplication $\mathbf{V} \times \mathbf{W}_1$. Its result, postrescaling, is translated into the updating voltages of the SOT-MTJs in the hidden layer. These voltages are subsequently applied to SOT-MTJs in the hidden layer, helping them stochastically sample their states directly without the need of generating random numbers calculated by CPU as usual. The updated states of the hidden layer can be read out through the resistances of the SOT-MTJs, thanks to their nonvolatility, with high and low resistances denoting output 1 and 0, respectively.

After sampling the hidden layer, reconstruction from the hidden nodes to the visible nodes can be carried out using the same Gibbs sampling approach. In principle, iterative compressions from visible to hidden and reconstructions in the opposite direction can enable an RBM to reach its thermodynamic equilibrium. After training, an RBM can not only grasp the probabilistic distribution of the training data but also generate new data according to the learned distribution, which accounts for the powerful generation capability of RBMs. Here, we employ the contrastive divergence method[34] to facilitate rapid training of RBMs. SOT-MTJs serve as hardware accelerators to perform Gibbs sampling during the learning and computation procedures. To conduct the Gibbs sampling in an RBM, the hidden (visible) nodes should be kept tightly invariant in their current states when updating nodes in the visible (hidden) layer. In this case, it is impractical to ensure indispensable synchronism with volatile MTJs. In contrast, it is easy for nonvolatile MTJs to do so by a clock signal. This is the reason why nonvolatile MTJs are indispensable for the following learning and training tasks. Supplementary Notes S1 and S11 provide more algorithm and implementation details.

Figure 2b−e depicts the image and speech recognition results achieved by the RBM utilizing SOT-MTJs. Figure 2b presents two typical "0" and "1" images from the MNIST data set, alongside their mean pooling results. The image-recognition accuracy for one (zero) is as high as 100% (93%) (Figure 2c). Figure 2d displays the raw "0" and "1" audio data from the Free-Spoken-Digit data set and their FFT results after mean pooling (specific details can be found in Supplementary Notes S1−4). The RBM attains 95% (90%) accuracy for recognizing one (zero) (Figure 2e). Furthermore, we constructed a deep belief network (DBN) based on the SOT-MTJs in a simulation to accomplish the recognition of all handwritten digits and speech from 0 to 9. Supplementary Note S7 illustrates the results and implementation details.

https://doi.org/10.1021/acs.nanolett.3c04820

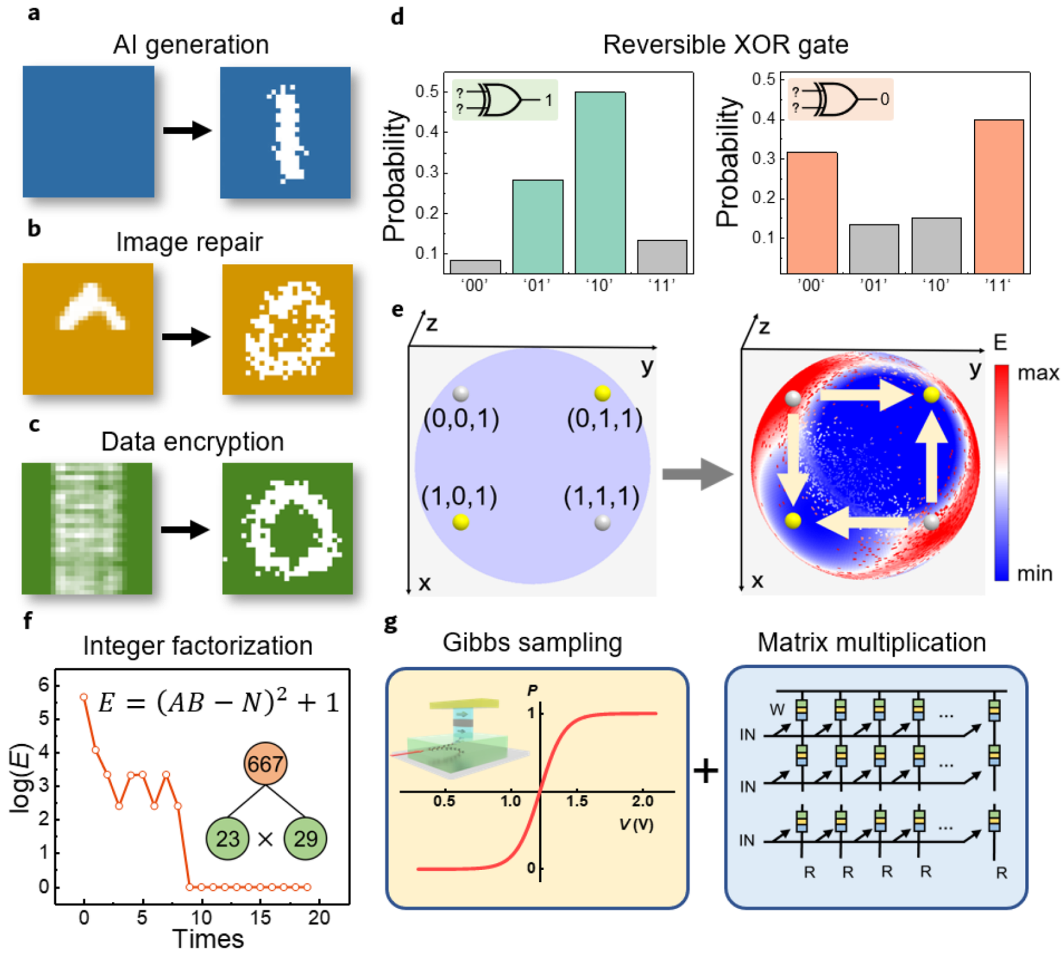


Figure 4. Envision of RBM based on SOT-MTJs. (a−c) The simulation results of RBM in generation (a), image repair (b), and data encryption (c). (d) The experimental results of the inverted XOR gate, inferring possible [AB] combinations from a given C. (e) The energy (E) defined by network weights as a function of the space coordinates of input (x, y) and output (z) for the inverted XOR gate. The left figure is the phase space with all zero initialization weights; the right is the phase space after training. The yellow dots highlight the correct XOR patterns corresponding to the output of 1 while the silver dots indicate the wrong patterns that a XOR gate should avoid. The selection of the correct patterns is achieved by properly shaping the energy landscape here. (f) The integer factorization process using an RBM based on SOT-MTJs in the experiment. (g) The Gibbs sampling based on SOT-MTJs can be combined with matrix multiplication also based on MTJ arrays to build a full spintronic hardware RBM accelerator.

In addition to the capabilities in image and speech recognitions, we further exemplify the multimodal and crossmodal abilities of our SOT-MTJ-based RBM. The capability to transform audio into an image using pretrained weights, as demonstrated in the experiment, is depicted in Figure 3. Similarly, the conversion of an image into audio is illustrated in Supplementary Note S8. The network architecture employed in our experiment is outlined in Figure 3c. The hidden layers of the two RBMs, which were separately responsible for image and speech recognitions in Figure 2, are interconnected via a one-hot layer. The pretrained weights ($\mathbf{W}_{1,\mathrm{audio}}$ and $\mathbf{W}_{1,\mathrm{image}}$) from Figure 2 were utilized in this configuration. The weights $\mathbf{W}_{2,\mathrm{audio}}$ and $\mathbf{W}_{2.\mathrm{image}}$ were easily computed to facilitate the seamless integration of the two RBM networks. The input audio data for zero and one fed into the combined RBM are depicted in Figure 3a,b, while Figure 3d,e displays the resulting output images generated by the RBM, correspondingly. It is important to note that the generated images are not replicas of any specific images from the MNIST data set; they are synthesized by the trained RBM owing to its stochastic nature. This combined RBM architecture can be feasibly extended to implement crossmodal perception and enhance associative power in neural networks.

In addition to image and speech recognition, we have also explored potential applications for the SOT-MTJ-based RBM in Figure 4. Once sufficiently trained, the RBM can generate an output image representing digit 1 even when the input is an entirely black image, as seen in Figure 4a. Furthermore, the RBM can assist in reducing noise from an incomplete input image to produce a significantly clearer output (illustrated in Figure 4b). Even in situations in which the noise is so intense that the original pattern becomes visually indistinguishable, the RBM still possesses the capability to extract the hidden information from the image. As demonstrated in Figure 4c, these noise-reducing capabilities could be valuable for data encryption, with the structure and weights of the RBM network serving as its decryption key. This feature implies that the RBM networks store information in their well-trained weights in a distributed manner, contributing to their resilience against intense noise.

To gain a deeper understanding of the operational principle of the RBM and to expand its applications, we further conducted an experimental demonstration of a reversible logic XOR gate, a milestone challenge in the history of AI because of the absence of its linear separability. The implementation details are explained in Supplementary Note S5. This XOR gate consists of two input nodes, A and B, and an output node, C. This reversible gate should have the ability not only to perform forward derivation (A⊕B → C) but also to enable reverse inference, allowing [AB] combinations inferred from C (C → [AB]). Our SOT-MTJ-RBM demonstrated 100% accuracy in derivation, as detailed in Supplementary Note S9. For inference, as shown in Figure 4d, when input C = 1 (0), the gate indeed outputs [AB] = [01] or [10] ([00] or [11]) with at least double the probabilities compared to the remaining two.

Figure 4e visualizes the space expanded by the RBM energy as a function of the coordinates A, B, and C. This gate has only three visible degrees of freedom, allowing us to easily visualize the phase space's landscapes. Before training, the weights of the RBM are all set to zero, resulting in a uniform space (Figure 4e). However, after adequate training, four attractors (local energy minima) appear in the energy landscape. Figure 4e also shows two attractors marked by two yellow dots located precisely around the positions (1, 0, 1) and (0, 1, 1). The other two attractors (0, 0, 0) and (1, 1, 0) are located on the back of Figure 4e, as shown in Figure S8 (Supplementary Note S5).

By analyzing this landscape, we can comprehend the working principle and robustness of the RBM networks. For recognition and classification, the fundamental process of training an RBM is to build a phase space by adjusting its weights, enabling attractors to emerge at points that correspond to the observed patterns. For optimization and denoising applications in a trained RBM, even if the input is somewhat distant away from the corresponding attractor, the Gibbs sampling rule, which favors a lower energy state with a higher probability, can still guide the system toward convergence at the attractor. Thus, it becomes understandable, as seen in Figure 4e, that the [AB] combinations of [10] and [01] around the attractors are more likely when C = 1.

Moreover, we extended our exploration of the RBM into optimization applications, focusing on the problem of integer factorization, as depicted in Figure 4f. Assuming an integer N can be factorized into $A \times B$, and the two factors can be further binary-encoded into $\sum a_i 2^i$ and $\sum b_j 2^j$, we can reframe the integer factorization problem into an energy minimization problem using the equation $E = (N - AB)^2 + 1 = [N - (\sum a_i 2^i)\cdot(\sum b_j 2^j)]^2 + 1$. In this equation, a_i and b_j are [0, 1] binary random variables, and they can be represented by nodes in different layers of an RBM. The weights between nodes a_i and b_j can be directly determined by the energy equation. More implementation specifics can be found in Supplementary Note S6.

By constructing the corresponding RBM based on the energy equation and adhering to the Gibbs sampling rule, we are guided to the minimal energy, which also reveals the factorization result. As Figure 4f demonstrates, log(E) rapidly declines and ultimately reaches 0 through iterations, enabling the decomposition of 667 into 23 × 29. This experiment attests to the capabilities of the RBM in solving such NP-hard (nondeterministic polynomial-time-hard) combinational problems.

Furthermore, high-barrier SOT-MTJs, when employed as nodes of RBMs, can store the results of sampling, which facilitates their recall without the need for unnecessary data transfers. Thus, this MTJ-based hardware acceleration approach, which can speed both sampling and matrix multiplication (as illustrated in Figure 4g), seems promising for the future development of RBMs and other stochastic computing architectures. We evaluated the energy consumption of each random number in Supplementary Note S10.

In conclusion, we have developed high-performance SOT-MTJ nanodevices that demonstrate a sigmoid-style switching probability feature and utilized these SOT-MTJs in the Gibbs sampling process for the RBM algorithm. The RBM powered by SOT-MTJs exhibits versatility and effectiveness in various tasks, including image and speech recognitions, crossmodal learning, noise reduction, reversible logic, and integer factorization. These demonstrations prove that high barrier SOT-MTJs are an excellent fit for the Gibbs sampling task and can well serve as nodes of RBMs. This study sheds light on the potential use of high-barrier SOT-MTJs for spintronic generative AI hardware accelerators in the future, which might speed up the pace of the field by hardware implementation and provide more efficient and effective ways of processing and interpreting data in stochastic computing architecture, thereby also broadening the application scope of SOT-MTJs.

■ ASSOCIATED CONTENT

Data Availability Statement

All data needed to evaluate the conclusions in the paper are presented in the paper. Additional data and codes are available from authors upon reasonable request.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.nanolett.3c04820.

MNIST data set image recognition; free spoken speech recognition; schematic diagram of pooling; simulation and experiment results about recognition rate and convergence speed; XOR gate inverted logic; integer factorization; image and speech recognition based on deep belief network; crossmodal training from images to speeches; the verification of the Boltzmann distribution and forward derivation results of the XOR gate; and energy consumption evaluation and details of the experiment (PDF)

Crossmodal training from image "0" to speech "zero" (Audio) (AVI)

Crossmodal training from image "1" to speech "one" (Audio) (AVI)

Crossmodal source code (TXT)

Image and speech recognition source code (TXT)

■ AUTHOR INFORMATION

Corresponding Authors

Caihua Wan − *Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Beijing 100190, China; Songshan Lake Materials Laboratory, Dongguan, Guangdong 523808, China*; Email: wancaihua@iphy.ac.cn

Xiufeng Han − *Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Beijing 100190, China; Songshan Lake Materials Laboratory,* *Dongguan, Guangdong 523808, China; Center of Materials Science and Optoelectronics Engineering, University of Chinese Academy of Sciences, Beijing 100049, China;* orcid.org/0000-0001-8053-793X; Email: xfhan@iphy.ac.cn

Authors

Xiaohan Li − *Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Beijing 100190, China; Center of Materials Science and Optoelectronics Engineering, University of Chinese Academy of Sciences, Beijing 100049, China*

Ran Zhang − *Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Beijing 100190, China*

Mingkun Zhao − *Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Beijing 100190, China*

Shilong Xiong − *Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Beijing 100190, China*

Dehao Kong − *Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Beijing 100190, China*

Xuming Luo − *Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Beijing 100190, China*

Bin He − *Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Beijing 100190, China*

Shiqiang Liu − *Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Beijing 100190, China*

Jihao Xia − *Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Beijing 100190, China*

Guoqiang Yu − *Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Beijing 100190, China; Songshan Lake Materials Laboratory, Dongguan, Guangdong 523808, China;* orcid.org/0000-0002-7439-6920

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.nanolett.3c04820

Author Contributions

[#](X.L., C.W., and R.Z.) These authors contributed equally to this work. X.H., C.W., and G.Y. planned the study. X.L., M.Z., R.Z., X.L., S.L., and J.X. prepared and characterized the MTJ devices. X.L., R.Z., D.K., S.X., and B.H. developed the algorithm and experiment. All authors contributed to the writing of the manuscript. All authors discussed the results.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China (MOST) (Grant No. 2022YFA1402800 and 2017YFA0206200), the National Natural Science Foundation of China (NSFC) (Grant Nos. 51831012, 51620105004, 11974398, and 12134017), the Strategic Priority Research Program (B) of Chinese Academy of Sciences (CAS) (Grant Nos. XDB33000000), and Beijing Natural Science Foundation (Grant No. Z201100004220006). C. H. Wan appreciates financial support from the Youth Innovation Promotion Association, CAS (Grant No. 2020008). Guangdong Special Support Project (2019BT02X030) and Shenzhen Peacock Group Plan (KQTD20180413181702403).

REFERENCES

(1) Misra, S.; Bland, L. C.; Cardwell, S. G.; Incorvia, J. A. C.; James, C. D.; Kent, A. D.; Schuman, C. D.; Smith, J. D.; Aimone, J. B. Probabilistic Neural Computing with Stochastic Devices. *Adv. Mater.* **2023**, *35*, No. 2204569.

(2) Kaiser, J.; Borders, W. A.; Camsari, K. Y.; Fukami, S.; Ohno, H.; Datta, S. Hardware-Aware In Situ Learning Based on Stochastic Magnetic Tunnel Junctions. *Physical Review Applied* **2022**, *17* (1), No. 014016.

(3) Bohm, F.; Alonso-Urquijo, D.; Verschaffelt, G.; Van der Sande, G. Noise-injected analog Ising machines enable ultrafast statistical sampling and machine learning. *Nat. Commun.* **2022**, *13* (1), 5847.

(4) Aadit, N. A.; Grimaldi, A.; Carpentieri, M.; Theogarajan, L.; Martinis, J. M.; Finocchio, G.; Camsari, K. Y. Massively parallel probabilistic computing with sparse Ising machines. *Nature Electronics* **2022**, *5* (7), 460−468.

(5) Shao, Y.; Sinaga, S. L.; Sunmola, I. O.; Borland, A. S.; Carey, M. J.; Katine, J. A.; Lopez-Dominguez, V.; Amiri, P. K. Implementation of artificial neural networks using magnetoresistive random-access memory-based stochastic computing units. *IEEE Magnetics Letters* **2021**, *12*, 1−5.

(6) Niazi, S.; Aadit, N. A.; Mohseni, M.; Chowdhury, S.; Qin, Y.; Camsari, K. Y. Training Deep Boltzmann Networks with Sparse Ising Machines. *arXiv*, 2023, 2303.10728. https://arxiv.org/abs/2303.10728 (accessed 2024-01-23).

(7) Yang, Q.; Mishra, R.; Cen, Y.; Shi, G.; Sharma, R.; Fong, X.; Yang, H. Spintronic Integrate-Fire-Reset Neuron with Stochasticity for Neuromorphic Computing. *Nano Lett.* **2022**, *22* (21), 8437−8444.

(8) Safranski, C.; Kaiser, J.; Trouilloud, P.; Hashemi, P.; Hu, G.; Sun, J. Z. Demonstration of Nanosecond Operation in Stochastic Magnetic Tunnel Junctions. *Nano Lett.* **2021**, *21* (5), 2040−2045.

(9) Singh, N. S.; Kobayashi, K.; Cao, Q.; Selcuk, K.; Hu, T.; Niazi, S.; Aadit, N. A.; Kanai, S.; Ohno, H.; Fukami, S.; et al. CMOS plus stochastic nanomagnets enabling heterogeneous computers for probabilistic inference and learning. *Nat. Commun.* **2024**, *15* (1), 2685.

(10) Grollier, J.; Querlioz, D.; Camsari, K. Y.; Everschor-Sitte, K.; Fukami, S.; Stiles, M. D. Neuromorphic Spintronics. *Nat. Electron* **2020**, *3* (7), 360−370.

(11) Mizrahi, A.; Hirtzlin, T.; Fukushima, A.; Kubota, H.; Yuasa, S.; Grollier, J.; Querlioz, D. Neural-like computing with populations of superparamagnetic basis functions. *Nat. Commun.* **2018**, *9* (1), 1533.

(12) Kaiser, J.; Faria, R.; Camsari, K. Y.; Datta, S. Probabilistic Circuits for Autonomous Learning: A Simulation Study. *Front Comput. Neurosci* **2020**, *14*, 14.

(13) Horowitz, M. Computing's energy problem (and what we can do about it). In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*; IEEE: 2014; pp 10−14.

(14) Keckler, S. W.; Dally, W. J.; Khailany, B.; Garland, M.; Glasco, D. GPUs and the future of parallel computing. *IEEE micro* **2011**, *31* (5), 7−17.

(15) Song, J.; Cho, Y.; Park, J.-S.; Jang, J.-W.; Lee, S.; Song, J.-H.; Lee, J.-G.; Kang, I. An 11.5 TOPS/W 1024-MAC butterfly structure dual-core sparsity-aware neural processing unit in 8nm flagship mobile SoC. In *2019 IEEE International Solid-State Circuits Conference-(ISSCC)*; IEEE: 2019; pp 130−132.

(16) Sebastian, A.; Le Gallo, M.; Khaddam-Aljameh, R.; Eleftheriou, E. Memory devices and applications for in-memory computing. *Nat. Nanotechnol* **2020**, *15* (7), 529−544.

(17) Wang, Z.; Wu, H.; Burr, G. W.; Hwang, C. S.; Wang, K. L.; Xia, Q.; Yang, J. J. Resistive switching materials for information processing. *Nature Reviews Materials* **2020**, *5* (3), 173−195.

(18) Wang, W.; Danial, L.; Li, Y.; Herbelin, E.; Pikhay, E.; Roizin, Y.; Hoffer, B.; Wang, Z.; Kvatinsky, S. A memristive deep belief neural network based on silicon synapses. *Nature Electronics* **2022**, *5* (12), 870−880.

(19) Huang, P.; Gu, Y.; Fu, C.; Lu, J.; Zhu, Y.; Chen, R.; Hu, Y.; Ding, Y.; Zhang, H.; Lu, S. SOT-MRAM-Enabled Probabilistic Binary Neural Networks for Noise-Tolerant and Fast Training. *arXiv*, 2023, 2309.07789, https://arxiv.org/abs/2309.07789 (accessed 2023-09-20).

(20) Wan, W.; Kubendran, R.; Schaefer, C.; Eryilmaz, S. B.; Zhang, W.; Wu, D.; Deiss, S.; Raina, P.; Qian, H.; Gao, B.; et al. A compute-in-memory chip based on resistive random-access memory. *Nature* **2022**, *608* (7923), 504−512.

(21) Yao, P.; Wu, H.; Gao, B.; Tang, J.; Zhang, Q.; Zhang, W.; Yang, J. J.; Qian, H. Fully hardware-implemented memristor convolutional neural network. *Nature* **2020**, *577* (7792), 641−646.

(22) Xiao, Z.; Naik, V. B.; Cheung, S. K.; Lim, J. H.; Kwon, J.-H.; Ren, Z.; Wang, Z.; Shao, Q. Device Variation-Aware Adaptive Quantization for MRAM-based Accurate In-Memory Computing Without On-chip Training. In *2022 International Electron Devices Meeting (IEDM)*; IEEE: 2022; pp 10.15.11−10.15.14.

(23) Ambrogio, S.; Narayanan, P.; Tsai, H.; Shelby, R. M.; Boybat, I.; di Nolfo, C.; Sidler, S.; Giordano, M.; Bodini, M.; Farinha, N. C. P.; et al. Equivalent-accuracy accelerated neural-network training using analogue memory. *Nature* **2018**, *558* (7708), 60−67.

(24) Jung, S.; Lee, H.; Myung, S.; Kim, H.; Yoon, S. K.; Kwon, S. W.; Ju, Y.; Kim, M.; Yi, W.; Han, S.; et al. A crossbar array of magnetoresistive memory devices for in-memory computing. *Nature* **2022**, *601* (7892), 211−216.

(25) Kim, I. J.; Kim, M. K.; Lee, J. S. Highly-scaled and fully-integrated 3-dimensional ferroelectric transistor array for hardware implementation of neural networks. *Nat. Commun.* **2023**, *14* (1), 504.

(26) Camsari, K. Y.; Faria, R.; Sutton, B. M.; Datta, S. Stochastic p-Bits for Invertible Logic. *Physical Review X* **2017**, *7* (3), No. 031014.

(27) Dalgaty, T.; Castellani, N.; Turck, C.; Harabi, K.-E.; Querlioz, D.; Vianello, E. In situ learning using intrinsic memristor variability via Markov chain Monte Carlo sampling. *Nature Electronics* **2021**, *4* (2), 151−161.

(28) Hassan, O.; Faria, R.; Camsari, K. Y.; Sun, J. Z.; Datta, S. Low-barrier magnet design for efficient hardware binary stochastic neurons. *IEEE Magnetics Letters* **2019**, *10*, 1−5.

(29) Zand, R.; DeMara, R. F. Snra: A spintronic neuromorphic reconfigurable array for in-circuit training and evaluation of deep belief networks. In *2018 IEEE International Conference on Rebooting Computing (ICRC)*; IEEE: 2018; pp 1−9.

(30) Zand, R.; Camsari, K. Y.; Pyle, S. D.; Ahmed, I.; Kim, C. H.; DeMara, R. F. Low-Energy Deep Belief Networks Using Intrinsic Sigmoidal Spintronic-based Probabilistic Neurons. In *Proceedings of the 2018 on Great Lakes Symposium on VLSI*; Association for Computing Machinery: 2018.

(31) Deng, J.; Miriyala, V. P. K.; Zhu, Z.; Fong, X.; Liang, G. Voltage-controlled spintronic stochastic neuron for restricted Boltzmann machine with weight sparsity. *IEEE Electron Device Lett.* **2020**, *41* (7), 1102−1105.

(32) Fernandez-de-Cossio-Diaz, J.; Cocco, S.; Monasson, R. Disentangling Representations in Restricted Boltzmann Machines without Adversaries. *Physical Review X* **2023**, *13* (2), No. 021003.

(33) Patel, S.; Canoza, P.; Salahuddin, S. Logically synthesized and hardware-accelerated restricted Boltzmann machines for combinatorial optimization and integer factorization. *Nature Electronics* **2022**, *5* (2), 92−101.

(34) Hinton, G. E. Training products of experts by minimizing contrastive divergence. *Neural computation* **2002**, *14* (8), 1771−1800.

(35) Sebastian, A.; Pendurthi, R.; Kozhakhmetov, A.; Trainor, N.; Robinson, J. A.; Redwing, J. M.; Das, S. Two-dimensional materials-based probabilistic synapses and reconfigurable neurons for measuring inference uncertainty using Bayesian neural networks. *Nat. Commun.* **2022**, *13* (1), 6139.

(36) Woo, K. S.; Kim, J.; Han, J.; Kim, W.; Jang, Y. H.; Hwang, C. S. Probabilistic computing using Cu0.1Te0.9/HfO2/Pt diffusive memristors. *Nat. Commun.* **2022**, *13* (1), 5762.

(37) Yang, J.; Wu, D.; Lee, A.; Razavi, S. A.; Gupta, P.; Wang, K. L.; Pamarti, S. A calibration-free in-memory true random number generator using voltage-controlled MRAM. In *ESSDERC 2021-IEEE 51st European Solid-State Device Research Conference (ESSDERC)*; IEEE: 2021; pp 115−118.

(38) Zink, B. R.; Lv, Y.; Wang, J.-P. Review of magnetic tunnel junctions for stochastic computing. *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits* **2022**, *8* (2), 173−184.

(39) Choi, W. H.; Lv, Y.; Kim, J.; Deshpande, A.; Kang, G.; Wang, J.-P.; Kim, C. H. A magnetic tunnel junction based true random number generator with conditional perturb and real-time output probability tracking. In *2014 IEEE International Electron Devices Meeting*; IEEE: 2014; pp 12.15.11−12.15.14.

(40) Zink, B. R.; Lv, Y.; Wang, J.-P. Telegraphic switching signals by magnet tunnel junctions for neural spiking signals with high information capacity. *J. Appl. Phys.* **2018**, *124* (15), No. 152121.

(41) Lv, Y.; Bloom, R. P.; Wang, J.-P. Experimental demonstration of probabilistic spin logic by magnetic tunnel junctions. *IEEE Magnetics Letters* **2019**, *10*, 1−5.

(42) Lv, Y.; Wang, J.-P. A single magnetic-tunnel-junction stochastic computing unit. In *2017 IEEE International Electron Devices Meeting (IEDM)*; IEEE: 2017; pp 36.32.31−36.32.34.

(43) Hong, J.; Li, X.; Xu, N.; Chen, H.; Cabrini, S.; Khizroev, S.; Bokor, J.; You, L. A Dual Magnetic Tunnel Junction-Based Neuromorphic Device. *Advanced Intelligent Systems* **2020**, *2* (12), No. 2000143.

(44) Borders, W. A.; Pervaiz, A. Z.; Fukami, S.; Camsari, K. Y.; Ohno, H.; Datta, S. Integer factorization using stochastic magnetic tunnel junctions. *Nature* **2019**, *573* (7774), 390−393.

(45) Zand, R.; Camsari, K. Y.; Datta, S.; Demara, R. F. Composable Probabilistic Inference Networks Using MRAM-based Stochastic Neurons. *ACM Journal on Emerging Technologies in Computing Systems* **2019**, *15* (2), 1−22.

(46) Shao, Y.; Duffee, C.; Raimondo, E.; Davila, N.; Lopez-Dominguez, V.; Katine, J. A.; Finocchio, G.; Khalili Amiri, P. Probabilistic computing with voltage-controlled dynamics in magnetic tunnel junctions. *Nanotechnology* **2023**, *34* (49), No. 495203.

(47) Honjo, H.; Nguyen, T.; Watanabe, T.; Nasuno, T.; Zhang, C.; Tanigawa, T.; Miura, S.; Inoue, H.; Niwa, M.; Yoshiduka, T. First demonstration of field-free SOT-MRAM with 0.35 ns write speed and 70 thermal stability under 400° C thermal tolerance by canted SOT structure and its advanced patterning/SOT channel technology. In *2019 IEEE International Electron Devices Meeting (IEDM)*; IEEE: 2019; pp 28.25.21−28.25.24.

(48) Song, M. Y.; Lee, C. M.; Yang, S. Y.; Chen, G. L.; Chen, K. M.; Wang, I. J.; Hsin, Y. C.; Chang, K. T.; Hsu, C. F.; Li, S. H.; et al. High speed (1ns) and low voltage (1.5V) demonstration of 8Kb SOT-MRAM array. In *2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*; IEEE: 2022.

(49) Liu, L.; Pai, C.-F.; Li, Y.; Tseng, H.; Ralph, D.; Buhrman, R. Spin-torque switching with the giant spin Hall effect of tantalum. *Science* **2012**, *336* (6081), 555−558.

(50) Cai, K.; Talmelli, G.; Fan, K.; Van Beek, S.; Kateel, V.; Gupta, M.; Monteiro, M. G.; Chroud, M. B.; Jayakumar, G.; Trovato, A.; et al. First demonstration of field-free perpendicular SOT-MRAM for

ultrafast and high-density embedded memories. In *2022 International Electron Devices Meeting (IEDM)*; IEEE: 2022.

(51) Song, M.; Duan, W.; Zhang, S.; Chen, Z.; You, L. Power and area efficient stochastic artificial neural networks using spin−orbit torque-based true random number generator. *Appl. Phys. Lett.* **2021**, *118* (5), 052401.

(52) Li, R.; Song, M.; Guo, Z.; Li, S.; Duan, W.; Zhang, S.; Tian, Y.; Chen, Z.; Bao, Y.; Cui, J.; et al. In-Memory Mathematical Operations with Spin-Orbit Torque Devices. *Adv. Sci. (Weinh)* **2022**, *9* (25), No. 2202478.

(53) Ostwal, V.; Appenzeller, J. Spin−orbit torque-controlled magnetic tunnel junction with low thermal stability for tunable random number generation. *IEEE Magnetics Letters* **2019**, *10*, 1−5.

(54) Liu, Y.; Wang, Z.; Li, Z.; Wang, X.; Zhao, W. A spin orbit torque based true random number generator with real-time optimization. In *2018 IEEE 18th International Conference on Nanotechnology (IEEE-NANO)*; IEEE: 2018; pp 1−4.

(55) Li, X. H.; Zhao, M. K.; Zhang, R.; Wan, C. H.; Wang, Y. Z.; Luo, X. M.; Liu, S. Q.; Xia, J. H.; Yu, G. Q.; Han, X. F. True random number generator based on spin−orbit torque magnetic tunnel junctions. *Appl. Phys. Lett.* **2023**, *123* (14), No. 142403.

(56) Chen, H.; Zhang, S.; Xu, N.; Song, M.; Li, X.; Li, R.; Zeng, Y.; Hong, J.; You, L. Binary and ternary true random number generators based on spin orbit torque. In *2018 IEEE International Electron Devices Meeting (IEDM)*; IEEE: 2018; pp 36.35.31−36.35.34.

(57) Srivastava, N.; Salakhutdinov, R. R. Multimodal learning with deep boltzmann machines. *Advances in neural information processing systems* **2012**, *25*, 2222−2230.

(58) Zhao, M. K.; Zhang, R.; Wan, C. H.; Luo, X. M.; Zhang, Y.; He, W. Q.; Wang, Y. Z.; Yang, W. L.; Yu, G. Q.; Han, X. F. Type-Y magnetic tunnel junctions with CoFeB doped tungsten as spin current source. *Appl. Phys. Lett.* **2022**, *120* (18), No. 182405.

(59) Fukami, S.; Anekawa, T.; Zhang, C.; Ohno, H. A spin-orbit torque switching scheme with collinear magnetic easy axis and current configuration. *Nat. Nanotechnol* **2016**, *11* (7), 621−625.

(60) Kumar, M. Effective Control of Threshold Voltage of MOS Transistors. *Journal of Active and Passive Electronic Devices* **2015**, *10* (2), 121−127.

(61) Grimaldi, A.; Selcuk, K.; Aadit, N. A.; Kobayashi, K.; Cao, Q.; Chowdhury, S.; Finocchio, G.; Kanai, S.; Ohno, H.; Fukami, S. Experimental evaluation of simulated quantum annealing with MTJ-augmented p-bits. In *2022 International Electron Devices Meeting (IEDM)*; IEEE: 2022; pp 22.24.21−22.24.24.

(62) Hayakawa, K.; Kanai, S.; Funatsu, T.; Igarashi, J.; Jinnai, B.; Borders, W. A.; Ohno, H.; Fukami, S. Nanosecond Random Telegraph Noise in In-Plane Magnetic Tunnel Junctions. *Phys. Rev. Lett.* **2021**, *126* (11), No. 117202.

(63) Vodenicarevic, D.; Locatelli, N.; Mizrahi, A.; Friedman, J. S.; Vincent, A. F.; Romera, M.; Fukushima, A.; Yakushiji, K.; Kubota, H.; Yuasa, S.; et al. Low-energy truly random number generation with superparamagnetic tunnel junctions for unconventional computing. *Physical Review Applied* **2017**, *8* (5), No. 054045.

(64) Kaiser, J.; Rustagi, A.; Camsari, K. Y.; Sun, J. Z.; Datta, S.; Upadhyaya, P. Subnanosecond Fluctuations in Low-Barrier Nanomagnets. *Physical Review Applied* **2019**, *12* (5), No. 054056.

(65) Schnitzspan, L.; Kläui, M.; Jakob, G. Nanosecond True-Random-Number Generation with Superparamagnetic Tunnel Junctions: Identification of Joule Heating and Spin-Transfer-Torque Effects. *Physical Review Applied* **2023**, *20* (2), No. 024002.

(66) Yin, J.; Liu, Y.; Zhang, B.; Du, A.; Gao, T.; Ma, X.; Dong, Y.; Bai, Y.; Lu, S.; Zhuo, Y. *Scalable ising computer based on ultra-fast field-free spin orbit torque stochastic device with extreme 1-bit quantization*; IEEE: 2022; pp 36.31.31−36.31.

(67) Camsari, K. Y.; Torunbalci, M. M.; Borders, W. A.; Ohno, H.; Fukami, S. Double-Free-Layer Magnetic Tunnel Junctions for Probabilistic Bits. *Physical Review Applied* **2021**, *15* (4), No. 044049.

(68) Kanai, S.; Hayakawa, K.; Ohno, H.; Fukami, S. Theory of relaxation time of stochastic nanomagnets. *Phys. Rev. B* **2021**, *103* (9), No. 094423.

(69) Kobayashi, K.; Borders, W. A.; Kanai, S.; Hayakawa, K.; Ohno, H.; Fukami, S. Sigmoidal curves of stochastic magnetic tunnel junctions with perpendicular easy axis. *Appl. Phys. Lett.* **2021**, *119* (13), No. 132409.

(70) Ackley, D. H.; Hinton, G. E.; Sejnowski, T. J. A Learning Algorithm for Boltzmann Machines*. *Cognitive Science* **1985**, *9* (1), 147−169.

https://doi.org/10.1021/acs.nanolett.3c04820