In-Memory Neural Stochastic Differential Equations with Probabilistic Differential Pair Achieved by In-Situ P-bit using CMOS Integrated Voltage-Controlled Magnetic Tunnel Junctions

Zhihua Xiao[†1,2], Yaoru Hou[†1], Zihan Tong[1], Yicheng Jiang[1], Yiyang Zhang[3], Xuezhao Wu[1], Albert Lee[4], Di Wu[4], Hao Cai[5] and Qiming Shao[1,2,3#]

[1]Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, China;
[2]AI Chip Center for Emerging Smart Systems, InnoHK Centers, Hong Kong Science Park, Hong Kong, China;
[3]Department of Physics, The Hong Kong University of Science and Technology, Hong Kong, China;
[4]InstonTech, Suzhou, China;[5]School of Integrated Circuit, Southeast University, Nanjing, China;
[†]Equal Contribution; [#]email: eeqshao@ust.hk

***Abstract*—**The probabilistic bit (P-bit) is the core of probabilistic computing. We propose a novel in-situ P-bit compatible with compute-in-memory (CIM) schemes using voltage-controlled magnetic tunnel junctions (MTJs) to eliminate the generation-sample-transfer-compute paradigm of current P-bits. The conventional approach of sampling and transferring random sequences between separate P-bits and computing units reintroduces the memory bottleneck seen in von Neumann architectures, thereby limiting the efficiency of probabilistic computing. By pairing a data-bit and a P-bit as a probabilistic differential pair in a crossbar array, we enable random sequences to be directly utilized for computing. This generation-compute scheme eradicates the sampling and transfer costs associated with previous probabilistic computing methods. Full reuse of devices in the differential cells allows for probabilistic CIM with a large number of P-bits and high parallelism, suitable for real-world probabilistic computing tasks. We demonstrated in-memory neural stochastic differential equations for the reverse diffusion process in generative models. The results shows that without the bottlenecks, in-situ P-bit throughput is 6× faster and 2.19× more efficient than ex-situ P-bits using the same technology. Compared to other devices and schemes, the proposed scheme is 3× faster than state-of-the-art CMOS designs and 1.36× more energy efficient.

I. Introduction

Stochastic differential equations (SDEs) are critical mathematical tools for modeling, analyzing, and predicting stochastic systems across various scientific, engineering, and financial domains. As shown in **Fig. 1**, the differential of an SDE can be generalized by two terms: the probability flow ordinary differential equation (ODE) (drifting term) and the stochastic injection (Langevin diffusion term)[1]. Solving SDEs requires multiple numerical iterations, making the process computationally intensive. To mitigate this complexity, neural networks are being adapted to solve SDEs. An artificial neural network (ANN) can efficiently model the deterministic probability flow ODE[2], while the stochastic injection relies on true random number generation (TRNG). The quality and efficiency of the TRNG is crucial for the stochasticity of SDEs. Although many emerging devices based on P-bits are designed for probabilistic computing, current P-bit TRNGs follow a generation-sample-transfer-compute scheme (**Fig. 2**)[3], [4]. Despite the efficiency of the emerging devices in TRNGs, sampling the random numbers requires amplifiers or filters, which are energy-consuming and slow. Additionally, transferring random sequences to probabilistic computing units can cause significant delays, reintroducing the memory bottleneck seen in von Neumann architectures.

Our work aims to overcome the existing bottlenecks in probabilistic computing schemes by integrating P-bits directly into computing units as in-situ P-bits. Utilizing the voltage-controlled magnetic anisotropy (VCMA) effect (**Fig. 3a**), we control the energy barrier of the MTJ to function as either a data-bit or a P-bit. By applying different control signals to conventional differential pairs (**Fig. 3b**), we pair a data-bit and a P-bit in a cell, forming a probabilistic differential pair and enabling in-memory probabilistic computing (**Fig. 3c**). This probabilistic switching of P-bits facilitates in-situ Monte-Carlo (MC) drop-connect sampling for probabilistic neural network computing, streamlining the process into a generation-compute scheme and overcoming the sampling and transfer bottlenecks present in conventional ex-situ P-bit systems. Experimental evaluations demonstrate that the proposed in-situ P-bits achieve significant speedup and energy efficiency due to the elimination of sample-transfer bottlenecks. An additional advantage of in-situ P-bits is the 100% reuse of devices and circuits, enabling P-bits within computing units with zero overhead. This approach potentially allows for the integration of a massive number of P-bits in future probabilistic computers.

II. Standalone and CMOS-Integrated VCMA-MTJ Characterization

To evaluate the characteristics of the designed VCMA-MTJ devices, we fabricated two types of chips with standalone and CMOS-integrated devices. The standalone device chip (**Fig. 4a** and **4b**) allows precise test pulses to be applied for basic switching behavior tests using probe stations (**Fig. 4c**). Additionally, array-level characterization and system demonstration were performed on the CMOS-integrated chip (**Fig. 5**).

A. Data-bit Characterization

The field-induced switching hysteresis loop in **Fig. 6** shows that the magnetic tunnel ratio (TMR) measured on the standalone devices is above 100%. **Fig. 7** shows the TMR distribution from four arrays, with low variations and a mean TMR measured under a 200mV read voltage of 65%. The resistance

distribution across four arrays is plotted in **Fig. 8**. Despite some resistance variation between arrays in different locations, the read margin remains clear. When two devices are combined as one probabilistic differential pair, the conductance of three states is shown in **Fig. 9**, with each state having a read margin of around 6σ. Differential pairs were manually programmed to corresponding states to store a logo, and the output conductance was measured. Results in **Fig. 10** indicate the high readability of the differential pair output.

B. P-bit Characterization

To set a VCMA-MTJ as a P-bit, we first measured the switching probability concerning pulse amplitude and duration on the standalone devices. **Fig. 11** shows the duration-controlled switching probability, with the first peak occurring at 700ps. **Fig. 12** illustrates the voltage-controlled switching probability, with a 50% switching probability found at 1.5V and a stochastic switching window of approximately 1V. **Fig. 13** shows the switching probability within an array, where variations are observed due to device variations and circuit delay. The pulse duration and amplitude with a 50% mean switching probability were used for TRNG/P-bit configuration.

III. True Random Number Generation of the In-Situ P-bit with VCMA-MTJ

The quality of true random number generation from the P-bit is critical for SDEs. The VCMA-MTJ P-bit's stochasticity arises from thermal noise-induced random motion of the spins, enabling high-quality random number generation without further calibration. To evaluate generation quality experimentally, a write pulse was applied to the integrated MTJs, and the device state was read out. **Fig. 14** shows the stochastic bit streams (SBS) from a single device, and **Fig. 15** plots the conductance of each random bit. The low cycle-to-cycle variation of MTJ devices ensures a high read margin of two states and calibration-free TRNG. The high-quality TRNG of our devices is validated by several randomness tests, including the autocorrelation test (**Fig. 16**), inter-sample Hamming distance of different SBS (**Fig. 17**), Shannon entropy of SBS (**Fig. 18**), and the NIST SP800-22 randomness test (**Table I**). Additionally, **Fig. 19** demonstrates the high endurance of the P-bit devices for 10^{12} write cycles without significant conductance degradation. **Fig. 20** shows the schematic and layout of the VCMA-MTJ probabilistic differential pair. The topology of the proposed design is the same as conventional differential cells that can be easily integrated with other CIM schemes.

IV. Demonstration of In-Memory Neural SDE Using Probability Differential Pair for Image Generation

Diffusion models for generative AI are among the most promising applications of neural SDEs. While many works use probability flow ODEs as a simplified version of reverse diffusion SDE to reduce the generation cost, SDEs still outperform in terms of generation quality[1]. Our goal is to generate high-quality samples efficiently using the proposed probabilistic differential pair. **Fig. 21** shows the computing scheme of conventional ex-situ P-bit reverse diffusion SDE in the latent space[5]. The latent code $\boldsymbol{Z}_T$ is forwarded to the ANN to calculate the probability flow ODE. The ex-situ P-bit generates SBS, which is added to the ANN output to get the SDE differential. The proposed in-situ P-bit reverse diffusion SDE uses the inherent stochasticity of MC dropconnect Bayesian neural networks achieved by probabilistic differential pairs to directly calculate the SDE differential. After a few iterations, the latent code at the SDE edge condition is decoded by a variational autoencoder (VAE) to generate the sample. The fundamental difference between ex- and in-situ P-bit methods lies in the point of stochasticity injection. For a fair comparison, we used the same network for models with or without MC dropconnect in the following experiments. As shown in **Fig. 22a**, the lack of stochasticity in ANNs necessitates an ex-situ stochastic injection step. In contrast, MC dropconnect models are inherently stochastic and can inject stochasticity in situ[6]. **Fig. 22b** demonstrates a good regression result of MC dropconnect models on the same distribution. We compared performance in terms of generation quality, energy consumption, and delay using the same compute-in-memory (CIM) emulator for neural network computing for both ex- and in-situ methods (**Fig. 23**). The conductance of each differential pair was measured from the CMOS-integrated VCMA-MTJ arrays, with the in-situ method taking additional SBS from arrays for in-situ P-bits. As shown in **Fig. 24**, the proposed in-memory neural SDE has a distribution[7] similar to the conventional ex-situ method, indicating good generation quality. Benchmarking the energy and delay of systems with different P-bit generation schemes (**Table II**), the proposed in-situ P-bit generation achieves $3\sim1\times10^6$ higher throughput and $1.36\sim1.9\times10^5$ higher energy efficiency with the lowest additional area cost to make the CIM system support probabilistic computing.

V. Conclusion

We experimentally demonstrated the behavior of data-bits and P-bits in both standalone and integrated VCMA-MTJ devices, validating that the proposed probabilistic differential pair based on these devices exhibits a high read margin with MC dropconnect support for CIM. The TRNG quality of our devices was experimentally evaluated, achieving a 3× speedup compared to existing state-of-the-art CMOS TRNGs. We further adapted the probabilistic differential pairs to demonstrate their application in in-memory neural SDEs for image generation. The results indicate that our proposed probabilistic computing scheme, which overcomes the sampling and transfer bottleneck of conventional methods, significantly increases efficiency. The complete reuse of devices in the differential cells also enables the integration of a large number of P-bits and high parallelism in probabilistic computing.

Acknowledgment This research was supported by ACCESS – AI Chip Center for Emerging Smart Systems, sponsored by InnoHK funding, Hong Kong SAR, and ITF (ITS/153/22). The authors also acknowledge IMEC for the VCMA-MTJ samples and CMOS-integrated chips.

References [1] T. Karras *et al.*, *Advances in Neural Information Processing Systems*, 2022. [2] Y. Song *et al.*, *ICLR 2021*. [3] T. Gong *et al.*, in *IEDM*, 2023. [4] N. S. Singh *et al.*, *IEDM*, 2023. [5] R. Rombach *et al.*, *CVPR*, 2021. [6] A. Mobiny *et al.*, *Sci. Reports* 2021.[7] L. Mcinnes, *et al.*, 2020. [8] Kim J *et al.*, *JSSC*, 2024 [9] Singh N *et al.*, *Nature Communication*, 2024 [10]B. Lin *et al., IEDM*, 2019. [11] Frustaci F. *et al.*, *TCAS II*, 2023

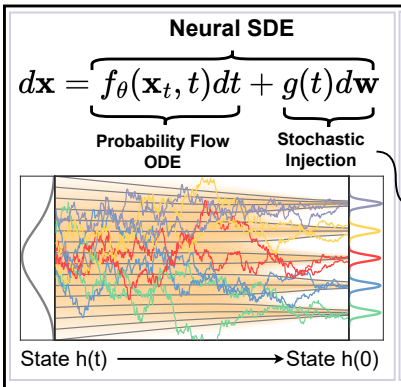


Fig. 1. A neural SDE uses a neural network to fit the transition of a probability flow ODE. Then a stochastic injection from TRNG is added for langevin diffusion. Finding the edge condition requires iterative steps of the SDE.

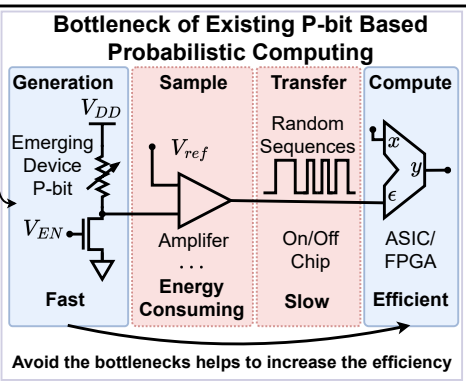


Fig. 2. The Existing P-bit based probabilistic computing follows a generation-sample-transfer-compute scheme. However, the sample and transfer are either energy-consuming or slow, which becomes the bottleneck of probabilistic computing.

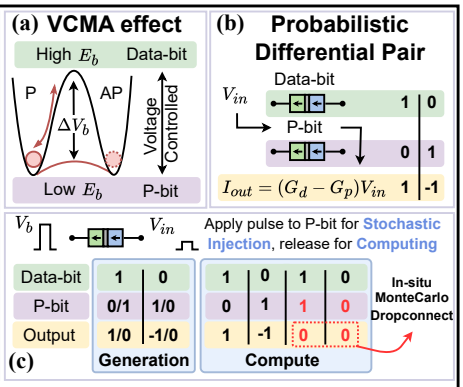


Fig. 3. **(a)** The VCMA effect can change the behavior of the device for data/probabilistic usage. **(b)** A probabilistic differential pair with VCMA-MTJ. **(c)** The generation-compute scheme of probabilistic differential pair can avoid bottlenecks.

Standalone and CMOS-Integrated Voltage Controlled Magnetic Anisotrophy Magnetic Tunnel Junction Characterization

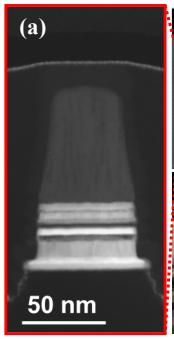


Fig. 4. **(a)** TEM of the VCMA-MTJ. **(b)** Standalone VCMA-MTJ test chip. **(c)** The standalone devices are tested with a probe station to apply accurate pulse for detailed analysis.

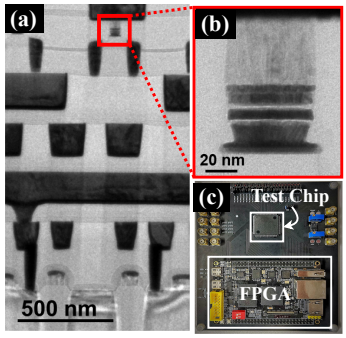


Fig. 5. **(a)** TEM of the CMOS-integrated VCMA-MTJ. **(b)** VCMA-MTJ fabricated with BEOL process showing clear edge. **(c)** The test platform of integrated devices.

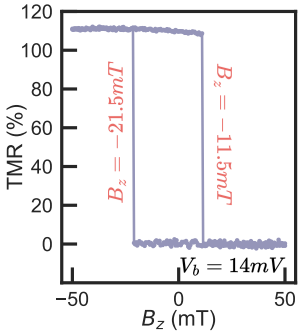


Fig. 6. Field induced switching hysteresis curves for standalone devices. Measured with 14mV and the TMR is above 100% with high yield.

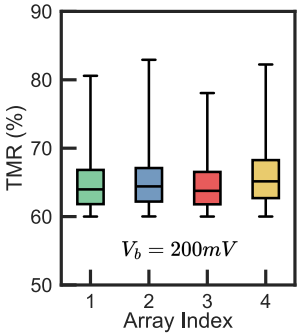


Fig. 7. TMR measured from 4 arrays with 200mV of integrated VCMA-MTJ devices. The MTR variation between different arrays is small at a mean value of 65%.

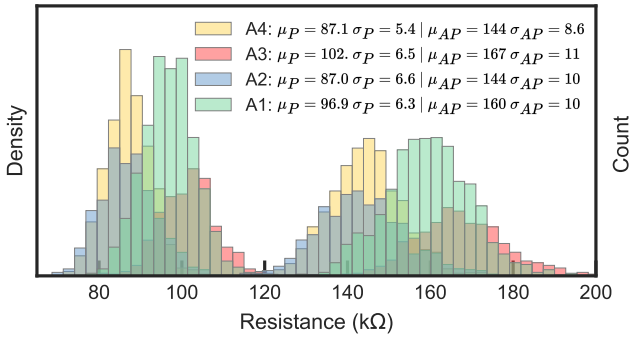


Fig. 8. Resistance distribution obtained from 4 arrays of the integrated VCMA-MTJ test chip. The mean and sigma values are annotated separately, showing good uniformity accross arrays.

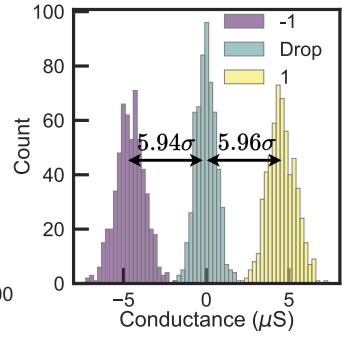


Fig. 9. Conductance distribution of the states of the probabilistic three differential pairs. The distance is 6σ.

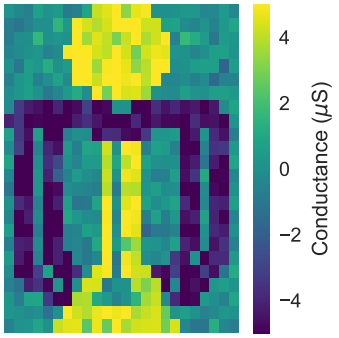


Fig. 10. Manually program the state of the probabilistic differential pairs into targeted states to faithfully store a logo.

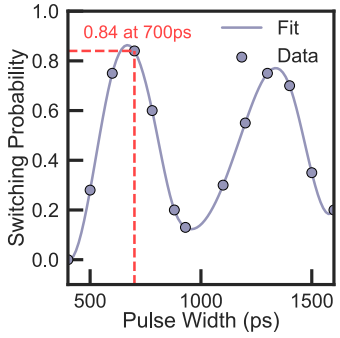


Fig. 11. The switching probability versus pulse width obtained from standalone VCMA-MTJ devices. The highest switching probability is obtained at 700ps write pulse.

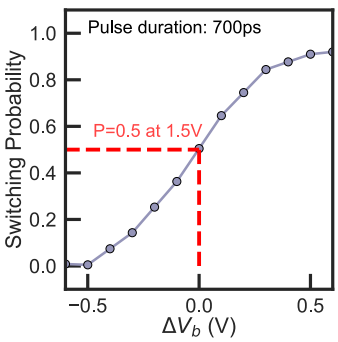


Fig. 12. The switching probability versus pulse amplitude obtained from standalone VCMA-MTJ devices. The 50% switching probability is obtained at 1.5V write pulse with a large stochastic window.

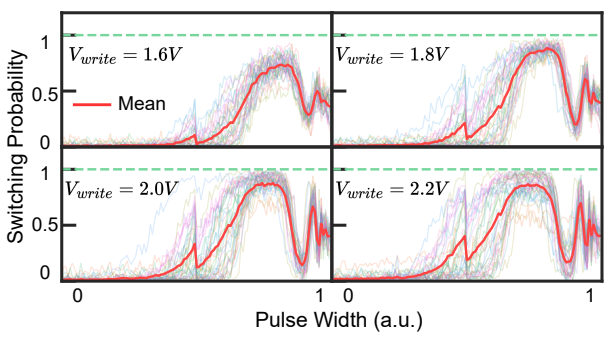


Fig. 13. The switching probability versus pulse amplitude obtained from integrated VCMA-MTJ devices under different write voltages. Due to the different locations of the devices and device variations the switching probability also varies. The mean value of the switching probability is highlighted in red.

True Random Number Generation of the In-Situ P-bit with VCMA-MTJ

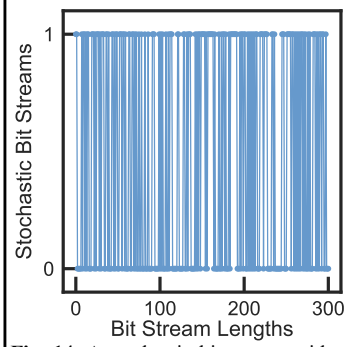


Fig. 14. A stochastic bit stream with a length equal to 300. Showing uniform distribution of the ones and zeros.

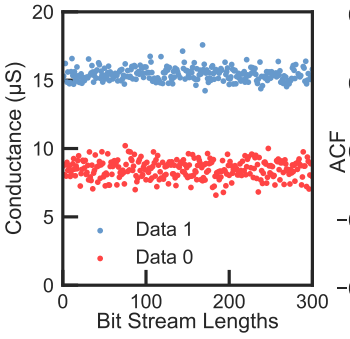


Fig. 15. The conductance measurement of the random sequence with length equal to 300. Conductances for 1/0 are stable.

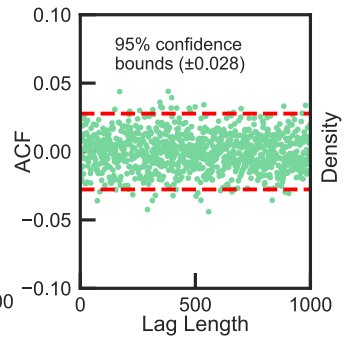


Fig. 16. The autocorrelation test result of the random sequences at 95% confidence level (±0.028).

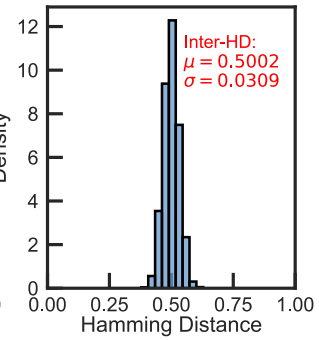


Fig. 17. The inter-sample hamming distance between multiple random sequences with a mean value of 0.5.

Test	P-Value	Results
Frequency	0.810	Pass
Block Frequency	0.921	Pass
Cumulative Sums	0.991	Pass
Runs	0.831	Pass
Longest Runs of Ones	0.202	Pass
Random Excursion	0.815	Pass
FFT	0.897	Pass
Non Overlapping Template	All Pass	
Serial	0.975	Pass
Approximate Entropy	0.970	Pass

Table I. The Nist SP800-22 Randomness Test of P-bit in Probabilistic Differential Pair

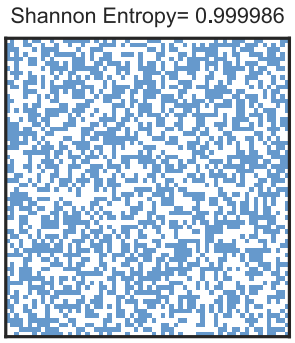


Fig. 18. The Shannon entropy calculated from a random sequence with length equal to 4096 obtained from the integrated chip.

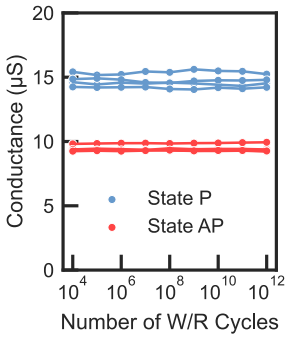


Fig. 19. The conductance of 4 devices after applied for 1e12 write-read pulses without significant conductance change.

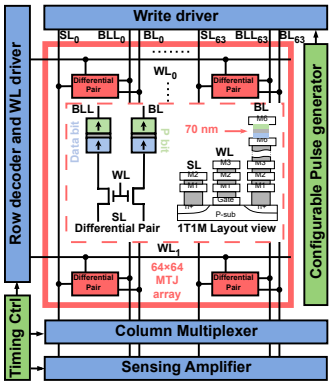


Fig. 20 Macro schematic of VCMA-MTJ probabilistic differential pair.

Demonstration of In-Memory Neural SDE Using Probability Differential Pair for Image Generation

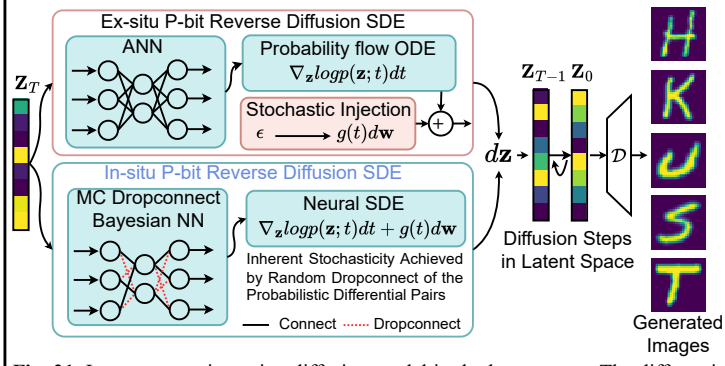


Fig. 21. Image generation using diffusion model in the latent space. The differential of each step is calculated by the neural SDE. Conventional ex-situ P-bit method injects the stochasticity explicitly while in-situ P-bit directly include the langevin motion implicitly in the inherent model stochasticity of MC dropconnect bayesian NN. The latent code at edge condition is decoded to pixel space by a VAE decoder.

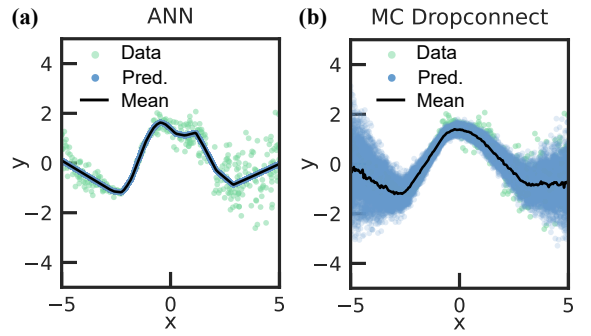
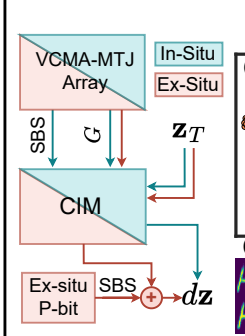


Fig. 22. A heteroscedastic regression task. **(a)** The ANN tends to predict the mean value of the data and completely misses the knowledge of stochasticity. **(b)** MC dropconnect models with probabilistic differential pairs can correctly capture the variation of the data with its inherent stochasticity.



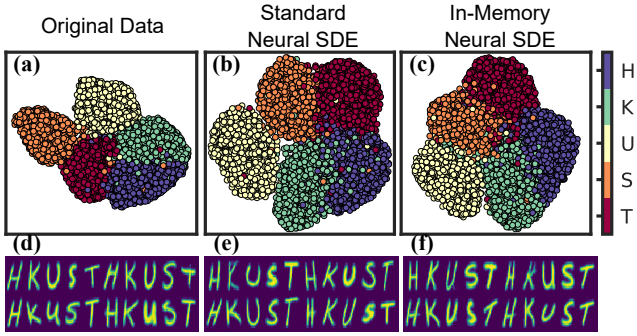


Fig. 23. Data flow of schemes are based on the same CIM system. **Fig. 24.** Data distribution clustered by UMAP[7] method. **(a)** Original In/Ex-situ P-bit Neural data distribution. **(b)** Generated data by the ex-situ P-bit generation SDE computing. Both method. **(c)** Generated data by in-situ P-bit method. **(d)-(f)** Sample data from each method. The generated samples show highly similar images to the original data, indicating a good generation quality.

	P-bit Device	Location	Additional Area	Throughput (bit/ns)↑	Efficiency (pJ/bit)↓
This Work	VCMA-MTJ	In-situ	0	29.8	0.089
		Ex-situ	609.8μm^2	4.97	0.196
JSSC '24[8]	CMOS		2254μm^2	10	0.121
IEDM '23[4]	IMA-sMTJ		PCB	5e-4	16800
NC'24 PMA[9]	PMA-sMTJ		PCB	3.2e-5	10000
NC'24 IMA[9]	IMA-sMTJ		PCB	10	10
IEDM '19[10]	RRAM		Array	1.1e-3	3.51
TCAS II '24[11]	CMOS		FPGA	0.3	400

↑ higher is better ↓ lower is better

Table II. P-bit/TRNG performance comparison of different works