Single-Step Extraction of Transformer Attention with Dual-Gated Memtransistor Crossbars

Nethmi Jayasinghe, Maeesha Binte Hashem, Dinithi Jayasuriya, Leila Rahimifard, Min-A Kang, Vinod K. Sangwan, *Senior Member, IEEE* Mark C. Hersam, *Fellow, IEEE* and Amit Ranjan Trivedi, *Senior Member, IEEE*

***Abstract*—We discuss how a dual-gated *memtransistor* crossbar can accelerate the extraction of the Transformer's attention scores. A memtransistor is a novel two-dimensional material-based device that offers non-volatile programmability and gate tunability. Leveraging these attributes, we demonstrate the extraction of quadratic-order products on a single memtransistor and the single-step extraction of attention scores without inferring intermediate query/key vectors. The query/key-free processing of memtransistor-based attention scoring results in 2.37× lower energy with less than half crossbar cells.**

***Index Terms*—Memtransistor, Transformers, Higher-Order Neural Processing, Time-series Prediction**

I. Introduction

Transformers have revolutionized machine learning, particularly for processing long-range sequence data. The model functions by mapping inputs into three distinct representations: keys, queries, and values, and utilizes self-attention to focus on the most relevant segments of the input dynamically. Initially designed for natural language processing, the unique attention mechanism of the model has now expanded into numerous other domains, such as image processing [1], video processing [2], event-driven computing [3], and cybersecurity [4].

Despite its predictive advantages, Transformers are also significantly more computationally expensive than predecessor architectures such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). This increased computational cost primarily arises from the quadratic scaling of their attention mechanism's complexity with the length of the input sequence, making it challenging to implement them on edge computing and resource-constrained devices.

In this work, we discuss a unique opportunity for accelerating Transformer's attention mechanisms by leveraging dual-gated *memtransistor* crossbars and in-memory computing [5]–[10]. Memtransistors, as shown in Fig. 1(a), are novel memory devices whose resistance can be programmed in a non-volatile manner, similar to a memristor, while also staying accessible for tunability by the top and bottom gates, similar to a transistor. Leveraging these unique attributes, we discuss how a single memtransistor can perform quadratic order products and how a crossbar of memtransistors can perform *Vector*×*Matrix*×*Vector* products in a single step.

This, in turn, enables the computation of attention scores directly from the Transformer's input tokens without having to extract query and key vectors as in traditional processing. Resultantly, the scheme significantly reduces the necessary multiplication-accumulation (MAC) operations and storage.

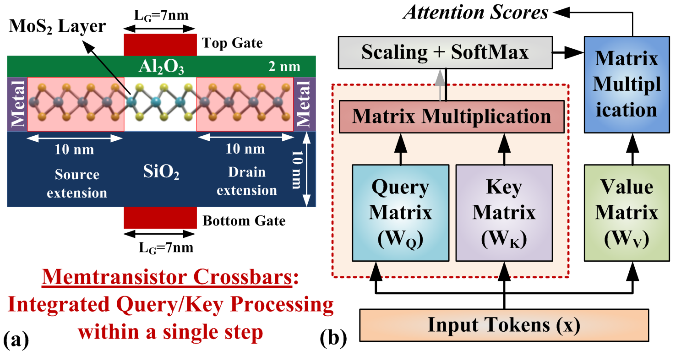

Figure 1: **(a)** Schematic of a dual-gated MoS$_2$ memtransistor. **(b)** The processing flow of the attention module in a Transformer. We leverage the non-volatile programmability and dual gate tunability of memtransistors for single-step attention extraction.

II. Transformer's Self-Attention Mechanism

The seminal study by [11] introduced Transformers, neural architectures built entirely on attention mechanisms. At the heart of the Transformer is the *multi-headed self-attention* (MHA) module, enabling the model to focus on different parts of the input sequence and weigh them based on their relevance. The MHA module processes an input tensor $\mathbf{x}$ of dimensions $[T, C]$, where T is the sequence length and C is the hidden size of input tokens. From $\mathbf{x}$, three linear projections for the query, key, and value are created as $\mathbf{Q}(\mathbf{x}) = W_Q\mathbf{x}$, $\mathbf{K}(\mathbf{x}) = W_K\mathbf{x}$, and $\mathbf{V}(\mathbf{x}) = W_V\mathbf{x}$, using the weight matrices W_Q, W_K, and W_V. The attention scores are then computed by $\mathbf{QK}^T$ and normalized using softmax$\left(\frac{\mathbf{QK}^T}{\sqrt{d_k}}\right)$. These scores are used to weigh the value vectors, producing the final output of the MHA module as softmax$\left(\frac{\mathbf{QK}^T}{\sqrt{d_k}}\right)\mathbf{V}$. The multiple attention heads capture various relationships within the sequence, enhancing the model's ability to learn complex patterns. Normalizing attention scores ensures stability and convergence, with the scaling factor $\sqrt{d_k}$ maintaining well-behaved gradients. By combining these elements, the Transformer architecture can efficiently process long sequences, capturing intricate dependencies and enabling state-of-the-art performance on a wide range of natural language processing tasks.

[1] L. Rahimifard, N. Jayasinghe, D. Jayasuriya and M. Binte Hashem are with the Department of Electrical and Computer Engineering, University of Illinois at Chicago, IL 60607 USA (e-mail: amitrt2@uic.edu). Min-A Kang, Vinod K. Sangwan, and Mark C. Hersam are with the Department of Materials Science and Engineering, Northwestern University, Evanston, IL 60208 USA

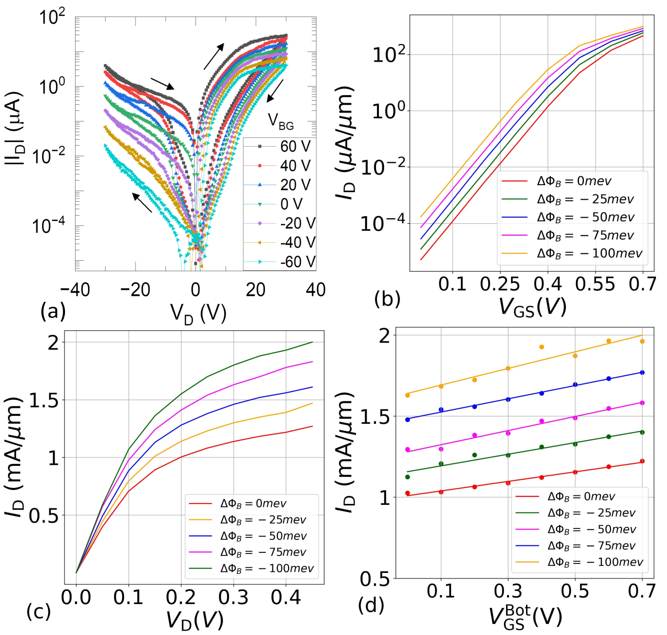

Figure 2: **(a)** Measured characteristics of memtransistor. For the scaled adaptation of memtransistor simulated using NEGF: (b) I_D-V_{GS} at varying Schottky Barrier (SB) height (V_{DS} = 0.4 V), (c) I_D-V_D at varying SB height while keeping the potential at the top and the bottom gates are 0.7 V, and (d) I_D-V_{GS}^{Bot} at varying SB height at V_{DS} = 0.4 V, sweeping the potential at the bottom gate, and the top gate set to 0.7 V.

III. Dual-Gated Memtransistor Crossbar for Single-Step Attention Scoring

A. Memtransistor: Physics and Characteristics

Sangwan and Hersam introduced a dual-gated memtransistor [12]–[16], using polycrystalline monolayer MoS_2 channel and Al_2O_3 and SiO_2 as top and bottom gate dielectrics, respectively. Device simulations, utilizes a scaled version of the memtransistor shown in Fig. 1(a). Four terminal MoS_2 memtransistors are programmable by drain voltage pulses that modulate Schottky barrier height ($\Delta\Phi_B$) at the source and drain contacts either by charge trapping or the migration of lattice defects like sulfur vacancies [17].

Fig. 2(a) shows the measured I_D-V_{GS} characteristics of fabricated memtransistor. Since the fabricated memtransistors have a larger dimension (gate length is ∼900 nm and oxide thickness is ∼30 nm), they require a larger voltage to operate. Therefore, to investigate the potential of nanoscale adaptation of the device in Fig. 1(a), we simulate them using a non-equilibrium Green's function (NEGF)-based model for current conduction and Schottky Barrier (SB) height modulation.

Fig. 2(b) shows exponential current conduction in the device while sweeping both gates, i.e., I_D-V_{GS} due to the thermionic emission-based current conduction, similar to measurements in Fig. 2(a). Programming $\Delta\Phi_B$ controls the device resistance in a non-volatile manner. Fig. 2(c) shows I_D-V_{DS} at increasing V_{DS} where the level of saturating current can be controlled by $\Delta\Phi_B$. Fig. 2(d) plots the current conduction at varying bottom gate voltage i.e., I_D-V_{GS}^{Bot}, while keeping the top gate voltage at 0.4 V and varying $\Delta\Phi_B$. Since the bottom gate has a much larger oxide thickness, it only weakly controls the channel electrostatics, resulting in almost linear control of channel conductance at varying V_{GS}^{Bot}.

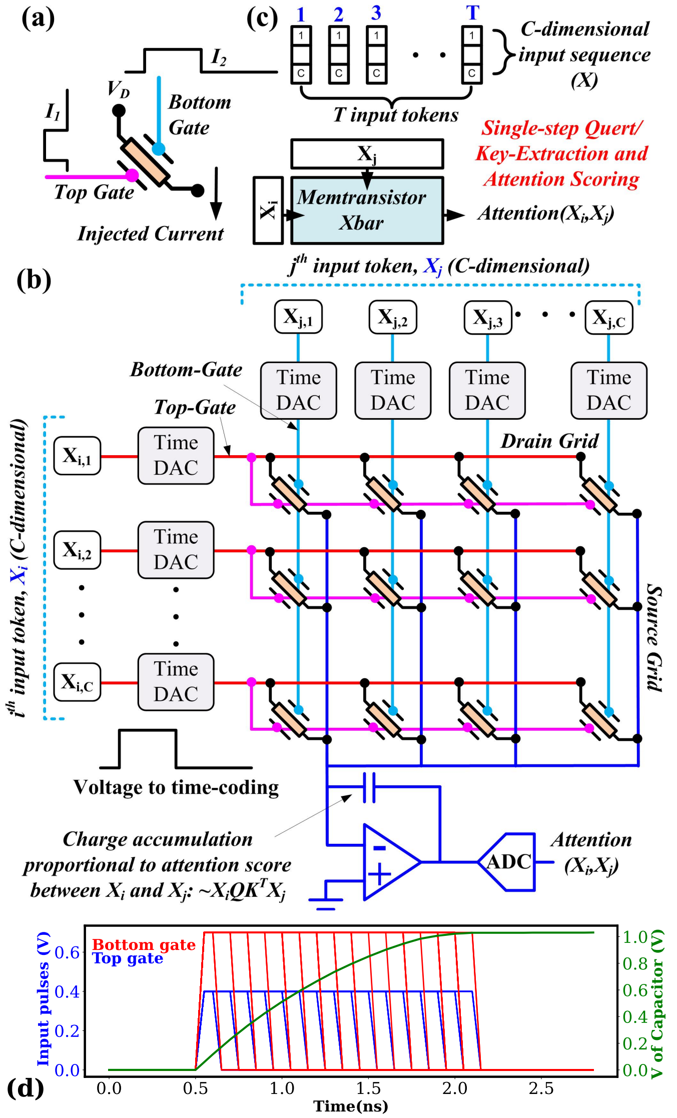

Figure 3: **(a)** Dual gate control of memtransistors for quadratic order interactions among inputs, I_1 and I_2, and programmed weight. **(b)** Memtransistor crossbar architecture for single-step key/query-free attention scoring in **(c)** and exemplary transient evolution of capacitor voltage in **(d)**.

B. Adapting Attention Loss to Memtransistor Electrostatics

For an input $\mathbf{x}$, consider the attention score A_{ij} computed between i^{th} and j^{th} tokens, $\mathbf{x}_i$ and $\mathbf{x}_j$, in a Transformer as

$$\alpha_{ij} = W_Q\mathbf{x}_i(W_K\mathbf{x}_j)^T = \sum_{l=1}^{d}\sum_{m=1}^{C}\sum_{n=1}^{C} x_{im}W_Q^{ml}W_K^{nl}x_{jn}$$
$$= \sum_{m=1}^{C}\sum_{n=1}^{C} x_{im}\left(\sum_{l=1}^{d} W_Q^{ml}W_K^{nl}\right)x_{jn} = \sum_{m=1}^{C}\sum_{n=1}^{C} x_{im}W_P^{mn}x_{jn} \tag{1}$$

Here, x_{mn} is the n^{th} element of the m^{th} token of input $\mathbf{x}$. W^{mn} is the m^{th} row and n^{th} column element of matrix W.

In Fig. 3, consider the memtransistor configuration to map the above attention coefficient (α_{ij}) computations. In Fig. 3(a), the conductance of the device is programmed in a non-volatile manner by programming $\Delta\Phi_B$ to match the desired weight value. Two inputs, I_1 and I_2, are processed on the device via time pulses at the top and bottom gates, respectively. The voltage generated follows a quadratic interaction of inputs and

This article has been accepted for publication in IEEE Electron Device Letters. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/LED.2024.3435540

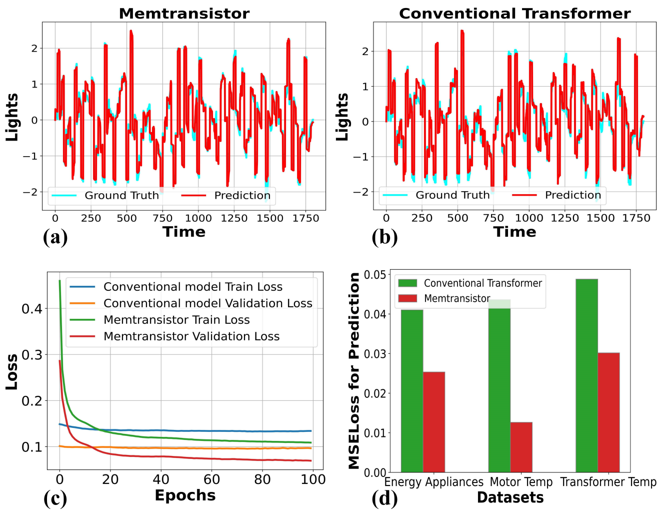

Figure 4: **(a)** and **(b)** Ground truth and prediction results for our Memtransistor and Conventional transformer for Energy Appliances dataset. **(c)** Comparison of train loss and validation loss for both transformers with the Energy Appliances dataset. **(d)** Comparison of MSE loss for prediction among three datasets.

weight, by integrating the current of the memtransistor on a downstream capacitor.

Fig. 3(b) shows the crossbar architecture that parallelizes this quadratic order interaction among inputs and weights to operate on all elements of tokens $\mathbf{x}_i$ and $\mathbf{x}_j$ in parallel, i.e., all indices m and n in Eq. (1). $\mathbf{x}_i$ is applied along the row electrodes using a digital-to-pulse converter (T-DAC). Likewise, $\mathbf{x}_j$ is applied along the column electrodes. The conductance of memtransistor at m^{th} row and n^{th} column is programmed proportional to $W_P^{mn} = \sum_{l=1}^{d} W_Q^{ml} W_K^{nl}$ as in (1) while keeping both top and bottom gates ON. With the above scheme, the capacitor voltage V_C follows

$$\alpha'_{ij} = \sum_{m=1}^{C} \sum_{n=1}^{C} \Big(\min(x_{im}, x_{jn}) W_P^{mn} + (x_{im} - x_{jn})^{+} W_{P1}^{mn} + (x_{jn} - x_{im})^{+} W_{P2}^{mn} \Big) \tag{2}$$

Above equation, W_{P1}^{mn} represents the translation of programmed weights at W_P^{mn} to the respective value when only the top gate is ON and the bottom gate is OFF. Likewise, W_{P2}^{mn} represents the translation of the programmed weight to the value when the top gate is OFF. Upon the application of row and column pulses in Fig. 3(b), the net charge from the crossbar is accumulated at a capacitor which follows the attention coefficients between the tokens in Fig. 3(d). The capacitor is coupled with an amplifier which enforces a virtual ground at its input port. The voltage output of the capacitor is digitized for storage and routing to other modules. For the signed implementation of W_P^{mn}, two memtransistors are used. One device retains data for positive weights, while the other retains data for negative weights; the intervening device, which is not in use, is set to a significantly high resistance. Similarly, inputs with signs are handled over two phases.

IV. Benchmarking of Memtransistor-based Transformers on Timeseries Datasets

Proposed single-step attention scoring using memtransistor crossbars is assessed on three time-series datasets: energy consumption data from household appliances [18], electric motor temperature variations [19], and electricity transformer temperature variations [20]. Notably, in Eq. (2), since the modified attention score (α'_{ij}) has an additional residual term, the loss function of Transformer processing was modified to account for this. Figs. 4(a,b) show the comparison between the ground truth and prediction on the energy appliances dataset for both the traditional transformer and memtransistor-based transformer. Fig. 4(c) shows the evolution of training and validation losses over the epochs. Fig. 4(d) compares the Mean Squared Error (MSE) of the prediction results across these datasets. Notably, across all three benchmark tests, our design consistently outperformed the original transformer in terms of predictive accuracy.

Table I: Memtransistor vs. Memristor for Transformer Inference

	Memtransistor	Memristor
# of crossbar cells	$T \times T$	$2 \times T \times d$
# of ADC conversions	1	$2 \times d$
Energy/Inference ($T = d = 64$)	104 pJ	246.8 pJ

Comments: T is the token length and d is the projection dimension of key/query matrices. Typically $d > T$ (in BERT and GPT-3).

Table 1 compares the efficiency of memtransistor and memristor-based processing of Transformer attention scores. For an application query of T tokens, single-step extraction of attention scores with memtransistor crossbar only requires $T \times T$ cells where the conventional query/key-based processing with memristors incurs $2T \times d$ cells. Note that in most Transformer models (such as BERT and GPT3), $d > T$; therefore, memtransistor processing results in significant area efficiency. Moreover, memtransistor crossbar requires only one digitization step per token sequence [Fig. 3(b)] whereas memristor-based conventional processing requires as many as in the projected dimension from query/key matrices, i.e., $2d$. Due to these efficiencies, even with a conservative assumption of $T = d$, the memtransistor-based design achieves a 2.37× lower energy than an equivalent memristor technology. The energy comparison was obtained by HSPICE simulation of memtransistor with 16nm CMOS-based peripherals and utilizing ADC and OP-AMP figures of merit from [21], [22].

Although our demonstration primarily focuses on memtransistor designs in [23], similar advantages are expected for other memtransistor technologies such as [24], which also offer non-volatile programming and gate tunability. The proposed framework can be adapted to other memtransistor technologies by fitting the $I_D - V_{GS}^{Bot}$ characteristics as shown in Fig. 2(c) and incorporating them into the training process.

V. Conclusions

This work introduced a framework to accelerate attention scoring in Transformer models by leveraging dual-gate tunability and non-volatile programming of conductance states in memtransistor crossbars. The proposed method reduces operations and storage needs by directly processing input tokens without separate query and key vector extraction.

Acknowledgement: This work was supported by the National Science Foundation (NSF) award CCF-2106964 and NSF EFRI BRAID Program under award EFMA-2317974.

References

[1] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 299–12 310, DOI:10.1109/CVPR46437.2021.01212.

[2] D. Neimark, O. Bar, M. Zohar, and D. Asselmann, "Video transformer network," in *Proceedings of the IEEE/CVF international conference on computer vision*, Feb 2021, pp. 3163–3172, DOI:10.1109/ICCVW54120.2021.00355.

[3] A. Sabater, L. Montesano, and A. C. Murillo, "Event transformer. a sparse-aware solution for efficient event data processing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, April 2022, pp. 2677–2686, DOI:10.1109/TPAMI.2023.3311336.

[4] Z. Wu, H. Zhang, P. Wang, and Z. Sun, "Rtids: A robust transformer-based approach for intrusion detection system," *IEEE Access*, vol. 10, pp. 64 375–64 387, Jan 2022, DOI:10.1109/ACCESS.2022.3182333.

[5] S. Yu, H. Jiang, S. Huang, X. Peng, and A. Lu, "Compute-in-memory chips for deep learning: Recent trends and prospects," *IEEE Circuits and Systems Magazine*, Nov 2022, DOI:10.1109/MCAS.2021.3092533.

[6] S. Yu, X. Sun, X. Peng, and S. Huang, "Compute-in-memory with emerging nonvolatile-memories: Challenges and prospects," in *2020 ieee custom integrated circuits conference (cicc)*. IEEE, March 2020, pp. 1–4, DOI:10.1109/CICC48029.2020.9075887.

[7] G. Finocchio, J. A. C. Incorvia, J. S. Friedman, Q. Yang, A. Giordano, J. Grollier, H. Yang, F. Ciubotaru, A. Chumak, A. Naeemi *et al.*, "Roadmap for unconventional computing with nanotechnology," *Nano Futures*, March 2024, DOI:10.1088/2399-1984/ad299a.

[8] P. Shukla, S. Nasrin, N. Darabi, W. Gomes, and A. R. Trivedi, "Mc-cim: Compute-in-memory with monte-carlo dropouts for bayesian edge intelligence," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 70, no. 2, pp. 884–896, Nov 2021, DOI:10.1109/TCSI.2022.3224703.

[9] S. Nasrin, D. Badawi, A. E. Cetin, W. Gomes, and A. R. Trivedi, "Mf-net: Compute-in-memory sram for multibit precision inference using memory-immersed data conversion and multiplication-free operators," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 5, pp. 1966–1978, April 2021, DOI:10.1109/TCSI.2021.3064033.

[10] S. Nasrin, M. B. Hashem, N. Darabi, B. Parpillon, F. Fahim, W. Gomes, and A. R. Trivedi, "Memory-immersed collaborative digitization for area-efficient compute-in-memory deep learning," in *2023 IEEE 5th International Conference on Artificial Intelligence Circuits and Systems (AICAS)*. IEEE, July 2023, pp. 1–5, DOI:https://doi.org/10.48550/arXiv.2307.03863.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, p. 6000–6010, Dec 2017, DOI:https://doi.org/10.48550/arXiv.1706.03762.

[12] V. K. Sangwan, H.-S. Lee, H. Bergeron, I. Balla, M. E. Beck, K.-S. Chen, and M. C. Hersam, "Multi-terminal memtransistors from polycrystalline monolayer molybdenum disulfide," *Nature*, vol. 554, no. 7693, pp. 500–504, Feb 2018, DOI:10.1038/nature25747.

[13] X. Yan, J. H. Qian, V. K. Sangwan, and M. C. Hersam, "Progress and challenges for memtransistors in neuromorphic circuits and systems," *Advanced Materials*, vol. 34, no. 48, p. 2108025, December 2022, DOI:https://doi.org/10.1002/adma.202108025.

[14] J. Yuan, S. E. Liu, A. Shylendra, W. A. Gaviria Rojas, S. Guo, H. Bergeron, S. Li, H.-S. Lee, S. Nasrin, V. K. Sangwan *et al.*, "Reconfigurable mos2 memtransistors for continuous learning in spiking neural networks," *Nano letters*, vol. 21, no. 15, pp. 6432–6440, July 2021, DOI:https://doi.org/10.1021/acs.nanolett.1c00982.

[15] H.-S. Lee, V. K. Sangwan, W. A. G. Rojas, H. Bergeron, H. Y. Jeong, J. Yuan, K. Su, and M. C. Hersam, "Dual-gated mos2 memtransistor crossbar array," *Advanced Functional Materials*, vol. 30, no. 45, p. 2003683, Sep 2020, DOI:https://doi.org/10.1002/adfm.202003683.

[16] V. K. Sangwan, S. E. Liu, A. R. Trivedi, and M. C. Hersam, "Two-dimensional materials for bio-realistic neuronal computing networks," *Matter*, vol. 5, no. 12, pp. 4133–4152, December 2022, DOI:10.1016/j.matt.2022.10.017.

[17] A. Azizi, X. Zou, P. Ercius, Z. Zhang, A. L. Elías, N. Perea-López, G. Stone, M. Terrones, B. I. Yakobson, and N. Alem, "Dislocation motion and grain boundary migration in two-dimensional tungsten disulphide," *Nature communications*, vol. 5, no. 1, p. 4867, Sep 2014, DOI:10.1038/ncomms5867.

[18] L. Candanedo, "Appliances Energy Prediction," UCI Machine Learning Repository, 2017, DOI: https://doi.org/10.24432/C5VC8G.

[19] W. Kirchgässner, O. Wallscheid, and J. Böcker, "Estimating electric motor temperatures with deep residual machine learning," *IEEE Transactions on Power Electronics*, vol. 36, no. 7, pp. 7480–7488, 2020 Dec, DOI:10.1109/TPEL.2020.3045596.

[20] S. Lin, W. Lin, W. Wu, F. Zhao, R. Mo, and H. Zhang, "Segrnn: Segment recurrent neural network for long-term time series forecasting," pp. 1–11, 2023 Aug, DOI : https://doi.org/10.48550/arXiv.2308.11200.

[21] L. Wang, M.-A. LaCroix, and A. C. Carusone, "A 4-gs/s single channel reconfigurable folding flash adc for wireline applications in 16-nm finfet," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 64, no. 12, pp. 1367–1371, July 2017, DOI:10.1109/TCSII.2017.2726063.

[22] G. Agarwal and V. Dwivedi, "Low-power two-stage op-amp in 16 nm," in *Information and Communication Technology for Intelligent Systems: Proceedings of ICTIS 2020, Volume 1*. Springer, 2021, pp. 637–642, DOI:10.1007/978-981-15-7078-0_62.

[23] A. Dodda, N. Trainor, J. M. Redwing, and S. Das, "All-in-one, bio-inspired, and low-power crypto engines for near-sensor security based on two-dimensional memtransistors," *Nature communications*, vol. 13, no. 1, p. 3587, June 2022, DOI:10.1038/s41467-022-31148-z.

[24] Y. Zheng, H. Ravichandran, T. F. Schranghamer, N. Trainor, J. M. Redwing, and S. Das, "Hardware implementation of bayesian network based on two-dimensional memtransistors," *Nature communications*, vol. 13, no. 1, p. 5578, Sept 2022, DOI:10.1038/s41467-022-33053-x.