e-ISSN 2150-5551
CN 31-2103/TB

Nano-Micro Letters

REVIEW

https://doi.org/10.1007/s40820-025-01850-w

Cite as
Nano-Micro Lett.
(2026) 18:13

Received: 11 March 2025
Accepted: 20 June 2025

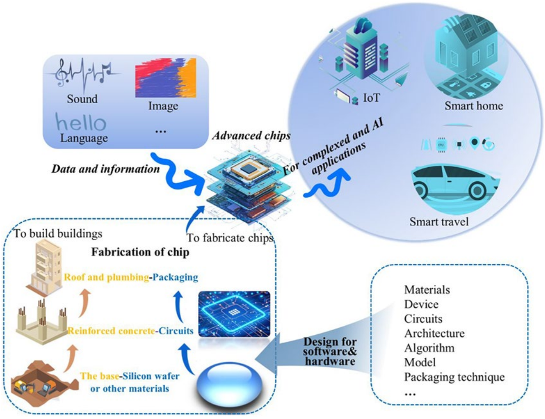
Advanced Design for High-Performance and AI Chips

Ying Cao[1,2], Yuejiao Chen[3], Xi Fan[4], Hong Fu[2] ✉, Bingang Xu[1] ✉

HIGHLIGHTS

- A comprehensive review focused on the recent advancement of the advanced and artificial intelligence (AI) chip is presented.
- The design tactics for the enhanced and AI chips can be conducted from a diversity of aspects, with materials, circuit, architecture, and packaging technique taken into considerations, for the pursuit of multimodal data processing abilities, robust reconfigurability, high energy efficiency, and enhanced computing power.
- A broad outlook on the future considerations of the advanced chip is put forward.

ABSTRACT Recent years have witnessed transformative changes brought about by artificial intelligence (AI) techniques with billions of parameters for the realization of high accuracy, proposing high demand for the advanced and AI chip to solve these AI tasks efficiently and powerfully. Rapid progress has been made in the field of advanced chips recently, such as the development of photonic computing, the advancement of the quantum processors, the boost of the biomimetic chips, and so on. Designs tactics of the advanced chips can be conducted with elaborated consideration of materials, algorithms, models, architectures, and so on. Though a few reviews present the development of the chips from their unique aspects, reviews in the view of the latest design for advanced and AI chips are few. Here, the newest development is systematically reviewed in the field of advanced chips. First, background and mechanisms are summarized, and subsequently most important considerations for co-design of the software and hardware are illustrated. Next, strategies are summed up to obtain advanced and AI chips with high excellent performance by taking the important information processing steps into consideration, after which the design thought for the advanced chips in the future is proposed. Finally, some perspectives are put forward.



KEYWORDS Artificial intelligence; Advanced chips; AI chips; Design tactics; Review and perspective

✉ Hong Fu, hfu@eduhk.hk; Bingang Xu, tcxubg@polyu.edu.hk
[1] Nanotechnology Center, School of Fashion and Textiles, The Hong Kong Polytechnic University, Hong Kong 999077, People's Republic of China
[2] Department of Mathematics and Information Technology, The Education University of Hong Kong, Hong Kong 999077, People's Republic of China
[3] State Key Laboratory for Powder Metallurgy, Central South University, Changsha 410083, People's Republic of China
[4] Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences, Ningbo 315201, People's Republic of China

Published online: 29 July 2025

Shanghai Jiao Tong University Press

Springer

1 Introduction

The past decade has witnessed the rapid progress of artificial intelligence (AI) techniques, which has revolutionized a wide range of fields, including the way to interpret information, the approach to discovery new materials, the method for creative work, and so on [1–9]. Particularly, great progress has been made in the functional materials and novel devices [10–12], which calls for AI to further promote these fields. Models of AI contain billions of parameters for the realization of high accuracy, which proposes high demand for the energy efficiency of processors. For instance, the deep neural network (DNN) model which contains many parameters can greatly promote the development of the recognition of images [13], the classification of videos, the transcription of speech [14, 15], and so on. To be specific, it has been verified that transformer and recurrent neural network transducer (RNNT) models with up to one billion parameters have shown a remarkable decrease in word error rate (WER) for the automated transcription of spoken English-language sentences. In addition to transcription, deep learning (DL) has also enhanced the performance of computer vision remarkably, which has been widely applied in the fields of autonomous driving [16], intelligent robotics [17], smart wearable devices [18, 19], and so on. Accordingly, new challenges have been put forward for the chips to handle these AI tasks. The advanced chips, which are featured with improved computing efficiency, reduced energy consumption, enhanced reliability, and excellent flexible expansion to be qualified for dealing with massive data, parallel tasks, and high concurrent requests proposed by the AI tasks, have drawn great attention, and significant progress of the advanced chips has been made by means of not only making improvements on the current silicon materials and silicon technologies, but also developing novel materials and modes [20]. For instance, data center chips, which are specifically designed for data centers, are featured with high performance and energy efficiency, and therefore, they are applied for cloud computing, AI training and inference, and big data analysis. Edge computing chips, which mainly pay attention to low latency, low power consumption, and miniaturization, have their advantages for the tasks required for real-time processing and environmental adaptability. Design thought for advanced chips referred to the process of transforming circuit structures and functions into physical layouts for the application of high-performance computing, covers wide aspects, which include but not limited to materials selection, device and circuit design, architecture optimization, and packaging technique development, and therefore, it is of importance for the rapid progress made in this field.

Many endeavors have been made to meet the challenges proposed by the AI tasks, with a lot of achievements and techniques emerging as the most promising approaches to address these issues [21–26]. For example, photonic computing makes it possible to process data faster and more energy efficiently [27]. At the meantime, the utilization of AI for optics can also improve the design and control of these optical systems [28–33]. Both the model training and inferential capability have been taken into considerations with the large-scale photonic chiplet and fully forward mode training being put forward. Computing-in-memory (CIM) which is inspired by the way in which human brain is used to process information has been put forward to resolve the von Neumann bottleneck [34]. Not only various synaptic arrays, but also efficient neuronal devices are developed. The advanced cognitive capabilities owned by the human brain have fueled a significant amount of AI research, which promote the development of sophisticated brain-inspired algorithms, as well as neuromorphic hardware with the pursuit to simulate various aspects of neural processing. Efforts have been made to develop efficient neuronal electronics. For instance, a novel dendrite function-like neuron has been developed [34]. Biocomputing, which is widely an interdisciplinary field combining biology and computer technology and uses other units instead of electrons or photons for information processing, has also emerged to address the existing issues. In addition to novel materials and new modes, improvements have also been made in areas of the conventional silicon-based chips, and more advanced preparation and packaging technology are proposed to deal with the increasing system complexity.

Significant progress has been made in both the hardware and the software of the advanced chips recently, which favors the fabrication of the chips. It is proposed that the fabrication of the chip bears some analogy to the construction of buildings. The fabricated chips can then be applied to handle various information to realize complexed and AI tasks, including computer vision, speech recognition and transcription, parallel imaging and all-optical classification, patients' gaits classification, and other various fields, with Internet of Things (IoT), smart

https://doi.org/10.1007/s40820-025-01850-w

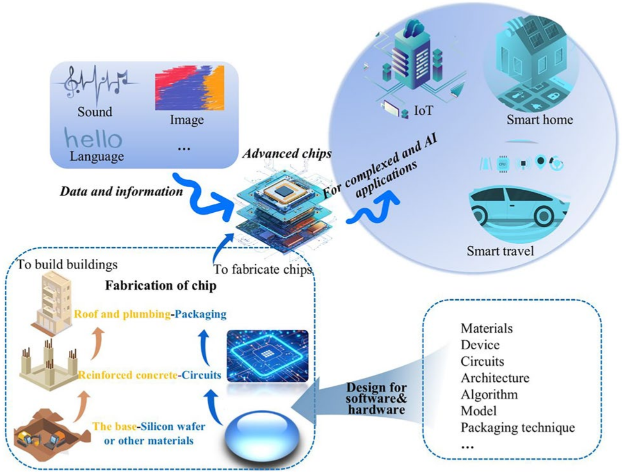


Fig. 1 Overview of the advanced and AI chip. The design for the software and hardware favors the fabrication of the chips, which bears some analogy to the construction of buildings. The fabricated chips can then be applied to handle various information to realize complexed and AI tasks

travel, smart robot, and smart home included (Fig. 1). An analog-AI chip with 35 million phase-change memory (PCM) devices has been developed [1]. A systemic energy efficiency of 74.8 peta-operations per second per watt is managed to be achieved by a type of all-analog photoelectronic chip [27]. Further to the inference chip, a fully forward mode (FFM) learning has been proposed for the training of optical neural networks, which is able to accomplish the compute-intensive training process on the physical system [35]. The fully hardware implementation of CIM has been experimentally realized by integrating neuron devices with a low accuracy loss [34]. Neuromorphic hardware equipped with associative learning abilities has been fabricated [36]. The low processor resting power of 0.42 mW has been achieved by a neuromorphic system on chip with the features of no-input calling for no energy, while a real-time power of as low as 0.70 mW can be realized for this system by the co-design of algorithm, software, and hardware [37]. The large-scale photonic chiplets, Taichi, which has millions-of-neurons capability with 160-tera-operations per second per watt (TOPS/W) energy efficiency, have been put forward. It has been verified that the high-fidelity AI-generated content can be realized by the photonic chiplet with up to two orders of magnitude of improvement in efficiency [38]. Publication number and the citation frequency of the papers concerning about the AI chip are counted from web of science. The data are collected with “AI chip” or “advanced chip” as topic words and are also filtered according to the actual relevance of the topic. As a result, an increasing number of original works have been published with high impact and sharply increasing citation frequency, which is demonstrated in Fig. 2. These results show that the research focused on the advanced chips has drawn great attention. The design strategies have been launched from various aspects, including materials, devices, circuits, architecture, and packaging techniques with the pursuit for multimodal data processing, reconfigurability, enhanced computing power, and high energy efficiency (Fig. 3). For instance, for multimodal data processing, which is required to handle different types of data, like images, sounds, and texts, proper packaging technology can facilitate the integration of different processing units more closely to enhance the processing speed, while reducing latency. Besides, the reconfigurable architecture which makes it possible for the hardware structure to be reconfigured according to different tasks also makes contribution to the multimodal data processing with the adjustment to different algorithm. However, reviews from the view of recent design tactics for AI chips are few. Herein, this review focused on the advanced design of the high-performance chips by means of not only making improvements on the current silicon materials and silicon technologies, but also developing novel materials and modes, like photonic computing, and the quantum processors,

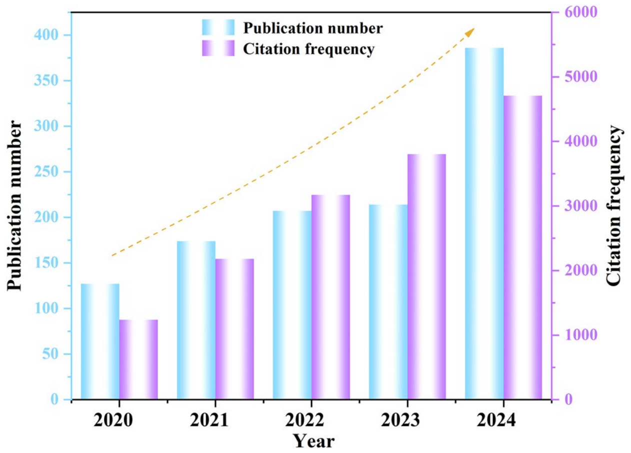


Fig. 2 Publication and the citation frequency of the papers concerning about the AI chip. The data are collected from web of science with “AI chip” or “advanced chip” as topic words, and are also filtered according to the actual relevance of the topic

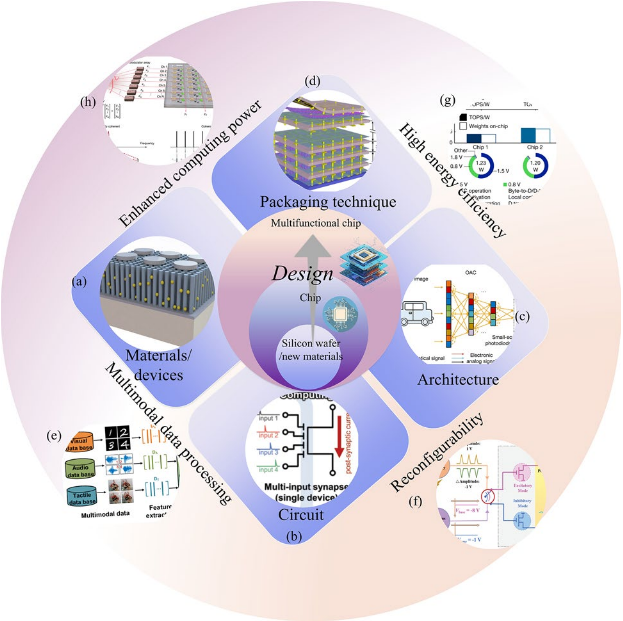


Fig. 3 Design strategies about the advanced chips. Design strategies carried for **a** material/device, reproduced with permission from Ref. [36] Copyright 2024, Springer, **b** circuit, reproduced with permission from Ref [39]. Copyright 2024, Wiley–VCH GmbH, **c** architecture, reproduced with permission from Ref. [27] Copyright 2023, Nature, and **d** packaging technique, reproduced with permission from Ref. [40] Copyright 2024, Nature. The design objective of realizing **e** multimodal data processing, reproduced with permission from Ref [41]. Copyright 2024, Nature, **f** reconfigurability, reproduced with permission from Ref [42]. Copyright 2023, Wiley–VCH GmbH, **g** high energy efficiency, reproduced with permission from Ref. [1] Copyright 2023, Nature, and **h** enhanced computing power, reproduced with permission from Ref. [43] Copyright 2024, Nature

among which many can meet the challenges proposed by the rapidly developing AI technology.

In this review, the basic background of AI chips was introduced first, as well as their working mechanisms, after which the design ideas in regard to software and hardware from the aspects of both the technique development for the conventional silicon-based chips, and the adoption of novel modes that extend the information processing from electrons, to photons, quantum, and biological elements, were demonstrated. Key factors which should be under consideration when designing the advanced chips were discussed from the view of the information processing procedures. Last but not least, we put forward some ideas with respect to the outlook of the advanced chips.

2 Mechanisms

The chips are applied to deal with various information and data. For instance, data can be collected from multimodal sensors. As for a typical task, the information is first captured by the sensors and is then digitized by a large number of analog-to-digital converters (ADCs) [27] (Fig. 4a). Data are then processed and transmitted (Fig. 4b, c). The neural network (NN) on a digital processing unit can then be made use of to process the information for recognition, classification, and other purposes. Edge computing can implement data processing at the sensors. In particular, as to a sensing-computing system on chip (SoC), the sensors can be integrated onto the chips to provide the information to be processed. For example, by leveraging the DVS as the eye of the chip, an asynchronous chip can be designed [44–46]. As the brightness of the scene changes, the DVS is managed to generate a stream of events asynchronously and sparsely, which can then be processed by the operation of the processor in the chip. However, it is proposed that not all sensors are solid state due to the diverse types of sensors, and therefore some are not suitable for integrated computing units. In addition to the sensing-computing system, there is also a high demand for large language model (LLM) acceleration, and therefore, how to provide strong computing power support should be taken into considerations.

The neuromorphic hardware learning from the information processing of human brain is a promising candidate for next-generation computer architectures because of its massive parallelism, robust fault tolerance, and high efficiency, which is different to the conventional architecture. The exploiting of the neuromorphic computing systems makes it possible to implement the parallel processing, which enables the execution of separate complex tasks by making use of several processors simultaneously, leading to the enhanced processing efficiency [39, 47–50]. Moreover, it is also expected for the neuromorphic systems to accomplish the processing of integrated signals from various inputs. The development of materials has promoted the realization of these functions greatly. The electrochemical artificial synapses can facilitate the simultaneous processing of multi-input signals via a unit device. The working mechanisms of the electrochemical artificial synapses composed of the electrolyte-based dielectric and

https://doi.org/10.1007/s40820-025-01850-w

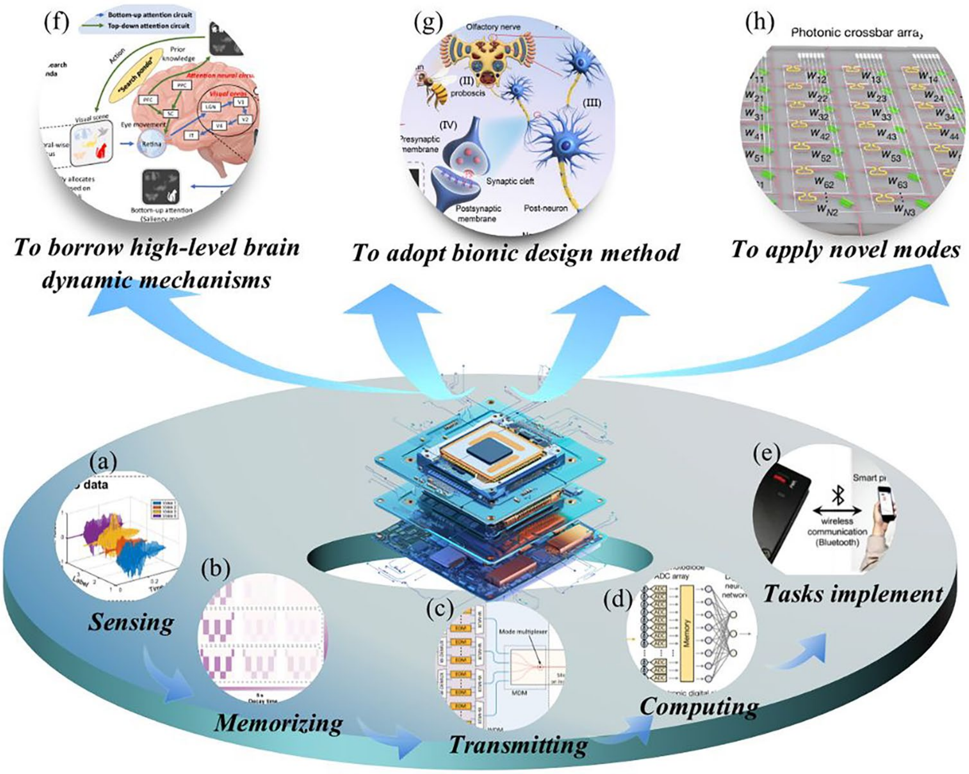


Fig. 4 Schematic illustration for the working mechanism of the advanced chips. Schematic illustration for the stage of **a** sensing, reproduced with permission from Ref [41]. Copyright 2024, Nature, **b** memorizing, reproduced with permission from Ref. [36] Copyright 2024, Springer, **c** transmitting, reproduced with permission from Ref. [59] Copyright 2022, Nature, **d** computing, reproduced with permission from Ref. [27] Copyright 2023, Nature, and **e** task implement, reproduced with permission from Ref [39]. Copyright 2024, Wiley–VCH GmbH. Schematic illustration for the method to improve the performance of chips by **f** borrowing high-level brain dynamic mechanisms, reproduced with permission from Ref. [37] Copyright 2024, Nature, **g** adopting bionic Design method, reproduced with permission from Ref. [36] Copyright 2024, Springer, and **h** applying novel modes, reproduced with permission from Ref. [43] Copyright 2024, Nature

ion-permeable semiconducting layer origin from the resistance tuning of the channel with penetrated ions and the retentive relaxation property.

Information is expected to be processed by the chips as the way of human brain, including learning, reasoning, and memorizing [39]. It turns out that human brain is managed to run even more complex neural networks with a total energy need of only 20 W [51, 52]. A variety of behaviors in the biological synapses, which are responsible for the information transmission between biological neurons, are simulated by the artificial neuromorphic electronics to handle the information collected by the sensors. Inspirations are also expected to be obtained from some high-level brain dynamic mechanisms in regard to the design of neuromorphic chips [53]. For the human brain, an important feature is to allocate its resources dynamically according to the required demand. To be specific, the salient stimuli can receive greater attention, which can be manifested by the heightened spiking activity in brain regions or the corresponding neurons associated with the stimulus. This high-level dynamic computing nature of the human brain is expected to be learned by the neuromorphic chips which are featured with minimal energy consumption for no input and significant variations for input changes. From the perspective of functional materials, some potential candidates, like two-terminal memristors which are featured with their compact synapse-like structures, have been extensively explored to equip the electronics with high complexity and improved completeness like the biological neurons for information transmission and processing [54, 55].

High capacity and high-throughput computing architectures are then required to handle the complex multimodality information collected from the environment [56] (Fig. 4d), and finally, the chips can be applied to implement various tasks (Fig. 4e). Great endeavors have been made to enhance these processes to improve the overall performance of the whole systems by a series of attempts, including borrowing high-level brain dynamic mechanisms (Fig. 4f), adopting bionic design approach (Fig. 4g), applying novel modes (Fig. 4h), and so on. Photonic processors are proposed to be a key to the hardware-based AI accelerators [23, 57, 58]. For the realization of in-memory photonic convolutional processing free of data movement between the memory and photonic processors, photonic tensor core incorporating phase-change-material photonic memories has been made use of. Generally, the data carried by each input coherent light at different wavelengths are weighted by the phase-change-material photonic memories. As a result, various tasks can be accomplished by the chips, ranging from computer vision, speech recognition, to gaits classification, which makes them to be qualified for a diversity of fields, including IoTs, smart homes, intelligent robotics, and so on.

3 Co-design of the Software and Hardware

AI relies on hardware and software to simulate human intelligence, and it is critical to carry out the co-design of both the software and the hardware for the advanced and AI chips. Specifically, software programming is of importance for the construction and training of NN, while hardware is crucial to process and handle the data for AI operation [60–62]. For example, although a highly programmable accelerator architecture for analog-AI has been proposed, it has yet to be demonstrated in hardware for the reason that the simulation study contains several design assumptions, among which one is the application of a dense and efficient circuit-switched 2D mesh for the exchange of massively parallel vectors of neuron-activation data over short distances, and another is the successful realization of DNN models which are large enough to be relevant for the commercial applications while maintaining high accuracy [1]. As a result, these issues should be solved for the design and fabrication of the analog-AI chips. Another case in point is that efforts have been made to design the CIM-based hardware systems in accordance with the requirements of the AI algorithm to successfully implement the extensive tasks of AI, promoting the commercial production of the CIM-based chips [34]. In this case, elaborate designs are essential in terms of both the optimized algorithms and innovative hardware for the neuromorphic computing systems. Besides, an algorithm-software-hardware co-design has also been put forward to realize the spike-based dynamic computing in the neuromorphic chip, with the hardware featured with no running energy for no-input, and the complete software toolchain for the efficient deployment of algorithms in a variety of dynamic vision applications [37].

3.1 Software

3.1.1 Some General Principles for Software Design

AI algorithms have been evolved rapidly. The intricate cognitive capabilities achieved by the human brain have sparked extensive research in AI with the promotion of sophisticated brain-inspired algorithms. It is worthwhile mentioning that the device-algorithm co-optimizations need to be carried out for the real-world application. Particularly, the software toolchain with data management, model simulation, and host management included is beneficial to deploy the algorithms and models efficiently for various applications [37]. Moreover, when developing different chips, the challenges and solutions at the software level are various, and design of the software is of important for all of these techniques, which lies in the aspects of model, algorithm adaptation, and toolchain. For instance, as to memristor, the integrated memory and computing architecture is required, while optical path programming is essential for photonic computing.

3.1.2 To Collaborate with the Hardware

The design of the software plays a crucial role in achieving various advantages of the advanced chips by working together with the hardware [37]. For instance, endeavors were made to combine the high-level dynamic computing nature of the brain with machine intelligence to equip the neuromorphic computing with energy advantages. The hardware was developed to meet the demand from dynamic computing, which indicated that no-input consumed no energy. Meanwhile, the design for an attention-based framework was also carried out to meet the challenge of dynamic computing which was featured with the fact that varied inputs consumed the energy with large variance. To accomplish this goal, inspirations for designing the dynamic spiking neural networks (SNNs) were gained from the understandings of visual attention in neuroscience. To be specific, since attention is a limited resource, the brain only processes a part of sensory input selectively. The neural related to attention can be divided into four structural levels, including circuit level, area level, neuron level, and synaptic level, and a general classification of attention neural circuits is the top-down versus bottom-up dichotomy (Fig. 5a). Top-down allocates the attention to internal behavioral goals of the brain, which can be presented through the priority map, while bottom-up deploys attention corresponding to the physical salience of a stimulus. As for the design of the framework for neuromorphic computing, a typical spiking neuron model and attention-based dynamic SNNs were illustrated as Fig. 5b, c. It was worthwhile mentioning that the dynamic framework acted as plug-and-play attention modules with the membrane potential optimized in a data-dependent manner, and combinable strategies of refinement and masking were provided by this dynamic framework. It was verified that a real-time power as low as 0.70 mW was successfully achieved by this neuromorphic system.

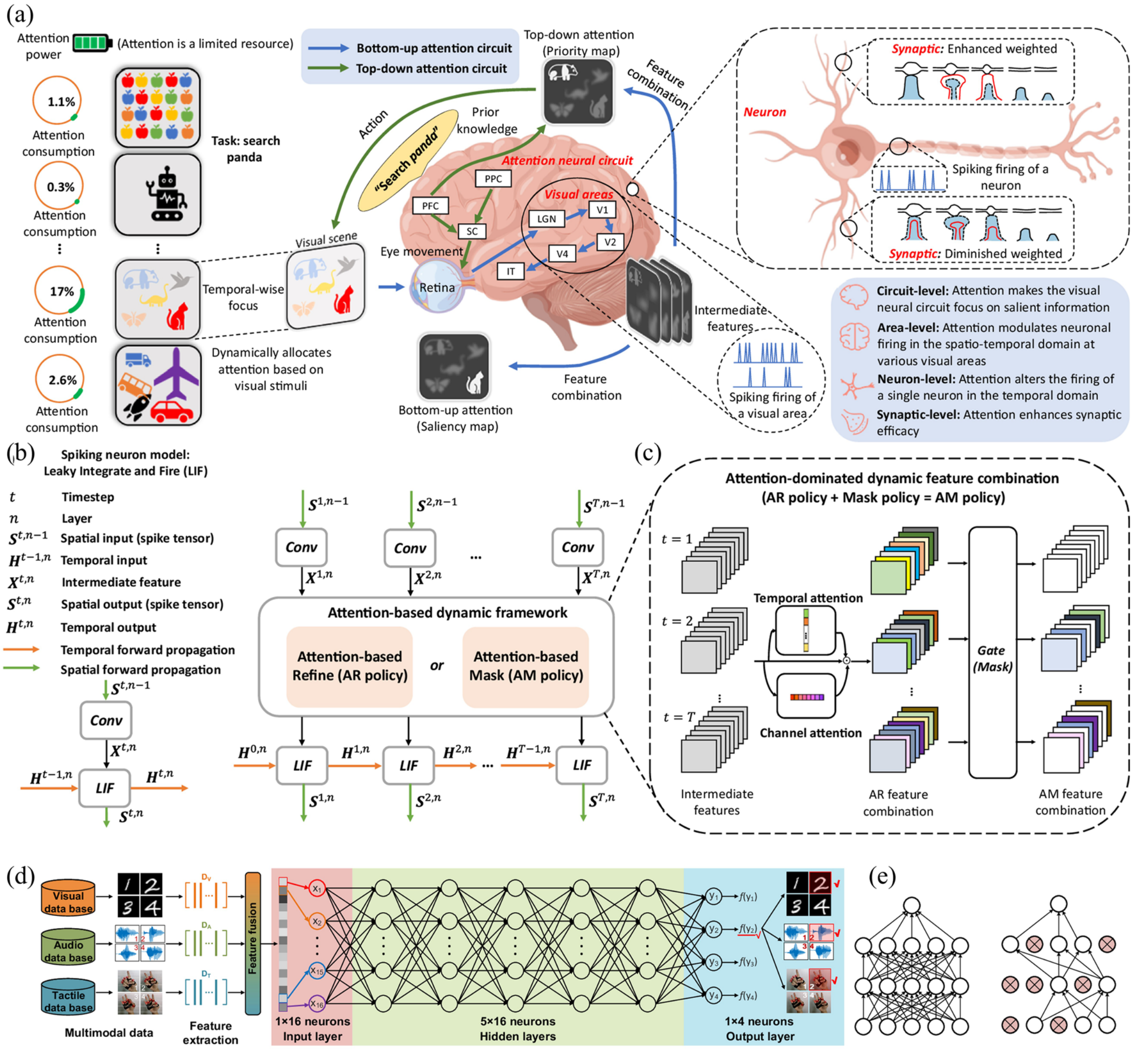


Fig. 5 Schematic of how software designs facilitate the development of advanced chips. **a** Schematic diagram for the attention-based dynamic response in neuroscience. Illustration for **b** a typical spiking neuron model and **c** attention-based dynamic SNNs. **a**–**c** Reproduced with permission from Ref. [37] Copyright 2024, Nature. **d** Schematic diagram of the optical neural network model for multimodal classification. **e** Schematic diagram of the drop-out algorithm. **d**–**e** Reproduced with permission from Ref [41]. Copyright 2024, Nature

3.1.3 To Conduct the Design of Algorithm

Some challenges brought by the explosive growth of the AI can be met by the design of algorithm, like the issue that multiple types of data are needed to be handled along with the boost development of the artificial intelligence generated content (AIGC) [63–66]. For example, it was pointed out that the majority of photonic neuromorphic processors for DL were able to handle only a single data modality for the reason that abundant parameters for training in optical domain were lack. To address this issue, a trainable diffractive optical neural network (TDONN) chip weas developed. In particular, the optical neural network model designed for the multimodal classification tasks was formed by three parts with an input layer, five hidden layers, and an output layer included (Fig. 5d). After the procedures of feature extraction and feature fusion, a feature vector was got from the datasets of different modalities, which was then applied as the input of the NN with the size of the feature vector matching the number of neurons. Each of the vector element was encoded into the optical signal by intensity modulation. In the hidden layers, the neurons were arranged in accordance with a multi-layer layout. The connection weights between each neuron were adjusted during training, and therefore trainable neurons were deemed as a critical prerequisite for reconfigurable TDONN, since the strong reconfigurability was essential for the multimodal DL. It took two steps for training of the TDONN chip, with the first one to extract the features and the second step to train the tunable diffractive units to accomplish the target tasks. It was worthwhile mentioning that customized gradient descent algorithm and drop-out mechanism of optical neurons were designed for the realization of the function. Firstly, an iteration threshold Titer was set for each neuron in the hidden layer of TDONN. During the iteration process, for the condition where the neuron could not increase CF after T adjustments, the neuron was set to be inactivated, and in the following iterations, this inactivated neuron would not be adjusted. As the training progresses, the number of deactivated neurons increased, and only the activated neurons needed to be tuned, leading to the reduce of the workload (Fig. 5e).

3.2 Hardware

3.2.1 Some General Principles for Hardware Design

Hardware design is imperative for promoting the development of different types of chips, the reason that it can solve the problems of different chips, making full use of these chips in various fields. To be specific, memristor, which can simulate the plasticity of biological synapses, plays a critical role in the brain-inspired computing. Photonic computing is featured with ultra high-speed, while it is also encountered with the problem of poor compatibility with silicon-based electronic chip. The computing power of quantum computing to deal with specific problems far exceeds that of classical computers, but the extremely low-temperature requirement is usually a challenge. Neuromorphic computing is managed to mimic the structure of human brain, and it can realize event-driven computing by means of asynchronous SNN, which is qualified for real-time perception and IoT. Accordingly, new circuit layout or material structure design is carried out to meet these challenges.

3.2.2 To Develop the Materials

The development of materials is served as one of the most important supports for the thriving chip industry. For instance, CIM-based hardware systems are designed according to the requirements from AI algorithm to accelerate the extensive computations by means of eliminating frequent data transfers between memory and processing units [67–69]. Accordingly, many endeavors have been made on the development of non-volatile memories (eNVMs) for the purpose of storing the weights in neural networks, with the PCM, RRAM, ferroelectric field effect transistor (FeFET), and other eNVMs included. Besides, more advanced functions are expected to be realized with high-efficiency algorithm while maintaining low hardware costs and high flexibility for the accomplishment of different application scenarios. As for the design of the hardware, a series of factors, like the stability, uniformity, and feasibility for large-scale realization, should be taken into consideration. Accordingly, efforts have been made not only by adopting novel modes, like the neuromorphic

https://doi.org/10.1007/s40820-025-01850-w

computing, photonic computing, and quantum computing, but also by improving the existing silicon chips, like the development of the package technique.

3.2.3 To Exploit New Mode: Neuromorphic Computing

Much efforts have been made on mapping the biological behavior in the nervous system to the electrical behavior in various devices, and many techniques have been emerged as the most promising approaches to meet the challenges brought by the AI tasks. It turns out that excessive energy consumption occurs with a significant amount of data moving between memory and processor, which is known as the von Neumann bottleneck [1]. CIM is proposed to be a promising approach to meet the challenge of increasing computational tasks brought about by the rapidly booming AI [34]. For the DNN models containing many large fully connected (FC) layers for the natural language processing (NLP), enormous movements of data are required in conventional digital implementation, while amortization over the subsequent computing is lacking. Analog-AI hardware is managed to meet this challenge by means of leveraging arrays of non-volatile memory (NVM) to perform the multiply–accumulate (MAC) operations, so that these workloads can be dominated directly in the memory [70–73]. When neuron-excitation data are moved to the location of the weight data, where the computation is executed, both the time and the energy are promising to be reduced. When taking the finite endurance and the power-hungry programming of NVM devices into consideration, it is inevitable that such analog-AI systems should be fully weight stationary. A highly heterogeneous and programmable accelerator architecture for analog-AI has been developed with the energy efficiencies 40–140 times higher than those of cutting-edge graphics processing units, but it has yet to be demonstrated in hardware due to the fact that several design assumptions are included [74].

Although the rapid progress has been made in CIM technology, it is crucial to recognize that the majority of the non-linear computations for the results after linear matrix–vector multiplying relies on conventional complementary metal oxide semiconductor (CMOS) circuits, with ADCs and digital circuits for complex arithmetic included, leading to excessive area and energy costs [75, 76] (Fig. 6a). It is crucial to make exploration for hardware implementation of activation functions on the basis of emerging devices and functional materials. Inspiration was obtained from dendritic computation of the pyramid neurons in the brain cortex to deal with the overhead in the hardware implementation of activation functions [34]. The distinguished calcium-mediated dendritic action potentials (dCaAPs) were brought into focus of the researchers which were in the pyramid neurons of the human layer 2 and 3 cortex. When compared to conventional all-or-none action potentials (APs), it was observed that the amplitude of dCaAPs becomes maximal for a certain threshold-level stimuli and was dampened for stronger stimuli (Fig. 6b), and therefore it was proposed that this distinctive dCaAP made it possible for a single neuron to implement XOR classification which typically required multilayered neural networks because of its inherent linear non-separability. It was pointed out that the electronic elements featured with negative differential resistance (NDR) were promising candidates of such mimicry, for which the measured response decreased as the stimulus intensity increased (Fig. 6c). NDR characteristics could be found in a wide range of electronics, among which Mott materials were one of the best candidates. As a well-studied Mott material, vanadium oxide (VO_2) was investigated as a potential substitute for conventional activation units of NN. Moreover, this novel activation unit was managed to be integrated within a non-von Neumann architecture, which was verified by co-implementing 1T1R arrays and these neurons on a single hardware platform (Fig. 6d).

In addition to the imitating of the essential synaptic functions, the in-depth study of the underlying learning and memory mechanisms in the biological brain is also vital for the realization of intelligent information processing at the hardware level [77]. For instance, it is proposed that the hardware realization of associative learning makes contribution to improving the functionality of NN, enhancing the performance of machine learning (ML) algorithms [78, 79]. Furthermore, it can also promote the development of more autonomous machines which are featured with the ability to adapt and learn in dynamical environments without the requirement for pre-programming [80–82].

3.2.4 To Exploit New Mode: Photonic Computing

In the post-Moore era, greater challenges have been proposed for the continuous demand of higher performance [38]. Photonic computing has offered significant advantages

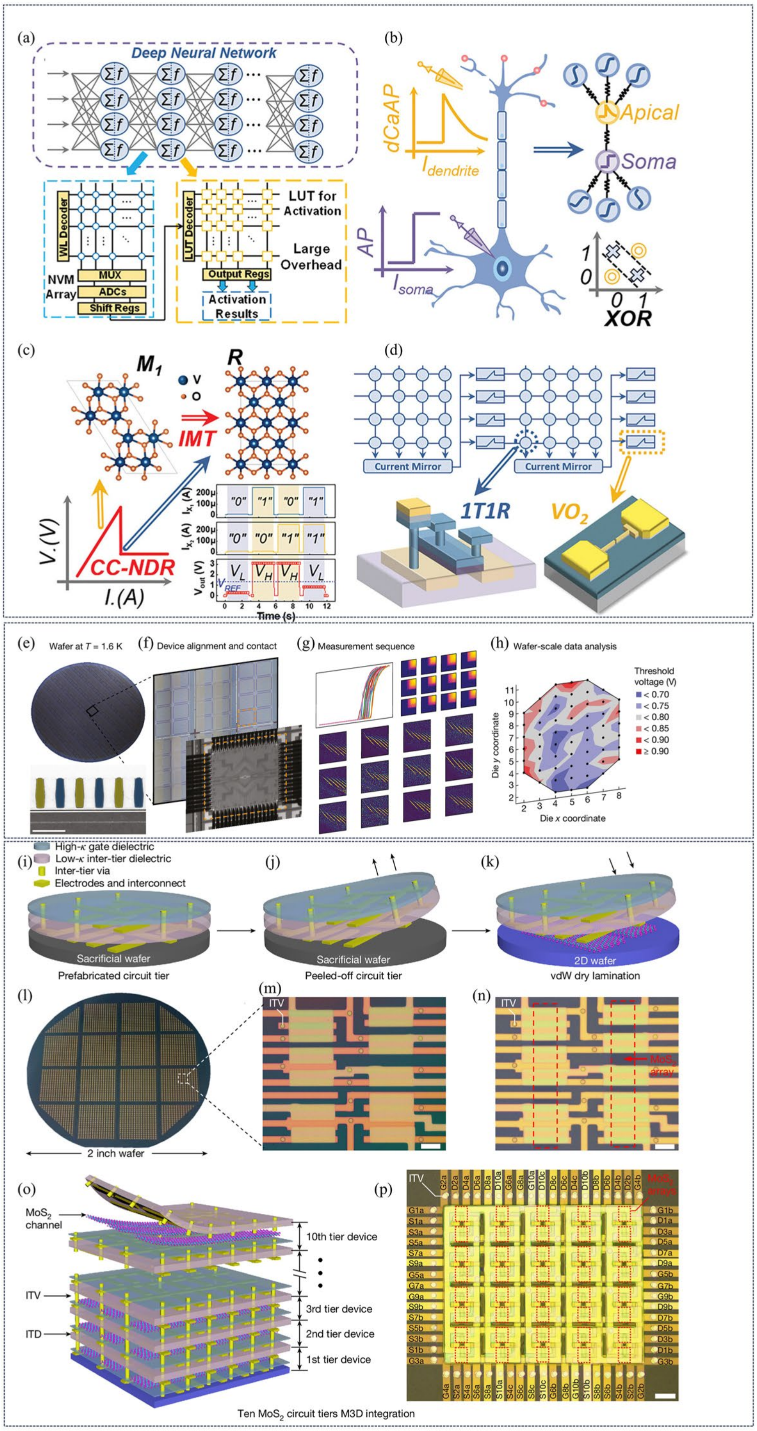


Fig. 6 Schematic of how hardware design promotes the development of different types of chips. **a** Schematic of the DNN structure and how to be realized by conventional hardware. **b** Schematic illustration of the calcium-mediated dendritic action potentials (dCaAPs) and the conventional all-or-none APs. **c** Schematic of NDR, insulator–metal transition (IMT), and the XOR operation realized in a single device. **d** Schematic illustration for the fully-hardware implementation of DNN. **a–d** Reproduced with permission from Ref. [34] Copyright 2024, Wiley–VCH GmbH. **e** Optical image of the completed spin qubit wafer. **f** Schematic of the device alignment and contact. **g** Various measurements used to extract the data. **h** The data used for statistical analysis. **e–h** Reproduced with permission from Ref. [100] Copyright 2024, Nature. **i** Circuit tier prefabrication on a sacrificial substrate. **j** Physically peeling off circuit tier, and **k** van der Waals dry lamination. **l** Optical images and **m** the zoomed-in image of prefabricated circuit tier on 2 inch sacrificial substrate. **n** Optical image of the final device. **o** Schematic diagram and **p** optical image of a 10-tier M3D system. **i–p** Reproduced with permission from Ref. [40] Copyright 2024, Nature

for the unprecedented light-speed and low-consumption computing [21, 22], which empowers much faster and more energy-efficient processing of data. In this case, the features of light are made use of to represent the information, and propagation and interference are taken advantaged of for computing [57, 83–95]. Meanwhile, the utilization of AI for optics can promote the design and control of optical systems. Recently, both the photons and the electrons have been used in an all-analog way to come up with a practical solution for the intelligent computing [27]. Moreover, the development of integrated photonics also makes contribution for the implementation of intelligent tasks by the photonic computing chips [25, 96–99].

3.2.5 To Exploit New Mode: Quantum Computing

In addition to the neuromorphic computing and photonic computing, quantum computing has been emerged as another advanced type of computing [100]. To promote the applications of spin qubit technology, physical qubit count is required to increase substantially, which makes it essential to fabricate spin qubit devices with the density, volume, and uniformity comparable with those of classical computing chips composed of billions of transistors [101]. The spin qubit technology is featured with its inherent advantages for scaling due to the qubit size, and another advantage is the native compatibility with CMOS manufacturing infrastructure. As a result, it is pointed out that manufacturing spin qubit devices with the same infrastructure as classical computing chips is managed to release the potential of spin qubits for scaling, and it is possible for them to offer an approach for building the fault-tolerant quantum computers. Furthermore, the scale of cryogenic device testing must be launched to enable efficient device screening [102, 103]. Spin qubits based on electrons in Si have demonstrated impressive control fidelities, but the challenges exist in the aspects of yield and process variation. Recently, some progress has been made to address this issue. One case in point was that a testing technique taking advantages of the cryogenic 300-mm wafer prober for collecting the data in high volume on the performance of hundreds of industry-manufactured spin qubit devices at 1.6 K was developed. It took about 2 h to cool 300-mm wafers to an electron temperature of 1.6 K [100], and the transmission electron micrograph of a Si/SiGe quantum dot qubit device cross section is shown in Fig. 6e. As is demonstrated in Fig. 6f, the device pads were then aligned to the probe pins, and devices were connected to measurement electronics at room temperature. A diversity of measurements could then be used to extract the data (Fig. 6g), and when this process on many devices across the wafer was repeated, the statistical analysis of wafer-scale trends was managed to be implemented by making use of the device data, which is illustrated in Fig. 6h.

3.2.6 To Promote the Integrating Technique

Besides the new materials and novel modes for the development of the advanced chips, progress has also been made in the aspect of integrating technique [40]. Monolithic three-dimensional (M3D) integration, for which multiple stacked tiers are fabricated sequentially on the same wafer by deposition of the upper tiers, has been proposed to overcome the scaling limitation with higher device density, and it enables new 3D computation systems, in which case various tiers, like the logic, memory, and sensor, are managed to be vertically interconnected [104–106]. As to the silicon-based M3D integration, challenges exist in the aspect of the low thermal budget, for which the process temperature of upper tiers should not exceed the back-end-of-line temperature to get rid of the performance degradation. It has been pointed out that two-dimensional (2D) semiconductors are promising for M3D integration, which is attributed to their dangling-bonds-free surface and the ability to be integrated to various substrates [107–113]. Recently, an alternative low-temperature M3D integration method by van der Waals lamination of entire prefabricated circuit tiers has been developed. The detailed integration processes included the procedures of circuit tier prefabrication on a sacrificial substrate, physically peeling off circuit tier and van der Waals dry lamination, which is demonstrated in Fig. 6i–k. It was noticeable that the prefabrication of all circuit stacks was based on standard photolithography processes, and it was compatible with wafer-scale M3D integration, which is demonstrated in Fig. 6l–n. A 10-tier M3D circuit within a total thickness of approximately 8 μm could be realized to verify the high-density M3D systems with multiple circuit tiers in the vertical direction, which is shown in Fig. 6o, p.

4 Strategies to Design Advanced and AI Chip with Enhanced Overall Performance

4.1 For Memory Purpose

The complexed and comprehensive simulations about the functions of the biological learning and memory are expected to be accomplished by the artificial neuromorphic devices [36]. A large amount of research has been launched focused on the neuromorphic electronics featured with massive parallelism, high efficiency, and capability. In particularly, as a form of associative learning, classical conditioning generally comprised of conditional stimuli (CS) and unconditioned stimuli (US) contains four features, including acquisition, extinction, recovery, and generalization, which are relevant to information storage, elimination of outdated information, rememorizing, and storage of new information in a cycle [114]. Accordingly, synaptic electronics equipped with associative learning capabilities are potential candidates for next-generation AI. Light has been used to coordinate with electrical devices to fully realize the aforementioned four features of classical conditional when taking the shortcomings of crosstalk, poor sustainability, and complex circuits for purely electrical signals with into account [36]. What is more, the difference in the aspect of relaxation times between and electrical stimuli and light endows the devices an inherent advantage to realize the characteristics of classic conditioning. The associative learning was accomplished by optoelectronic memristors based on Ag/TiO_2 nanowires (NWs): ZnO quantum dots (QDs)/FTO (ATZ-based device). As is shown in Fig. 7a, the flower nectar was served as the US that caused the proboscis extension, while the flower odor was served as CS which must be trained through the coordination of the olfactory and proboscis nerves to trigger the proboscis extension directly. A two-port ATZ-based memristive device was designed to simulate the synaptic behavior with a structure of the vertical arrangement similar to that of the synapses (Fig. 7b), and the SEM of the as-prepared device is demonstrated in Fig. 7c. It was verified that in addition to the basic synaptic behaviors, more advanced synaptic functions like learning-forgetting-relearning functions could also be achieved.

4.2 During Transmitted Process

The issue of data transfer limit for high-performance silicon chips has drawn a lot of attention, for which several schemes have been proposed [59]. Optical computing has great potential in improving the speed of a diversity of ML applications, which is attributed to its enhanced data transfer, low latency, and fast computation rate when taking the fact that light travels much faster than electrical signal under considerations [13, 23, 58, 115, 116]. Besides, the use of optical interconnects has become as a potential technology that can address this problem. It is pointed out that the chip-scale optical interconnects are promoted by the development of wavelength-division multiplexing (WDM) technique, which makes it possible to realize the parallel signal transmission by means of encoding data independently carried on multiple frequencies of light [117, 118]. After that, in order to further increase the link bandwidth, attentions have been paid on the other promising dimension of signal encoding for multiplexing, like the spatial domain. To be specific, the light can be decomposed into a series of optical beams with orthogonal spatial cross sections, and these orthogonal spatial modes can act as independent communication channels [119–126]. It is possible for each of them to support a full WDM link, leading to the multiplicative effect on the bandwidth of an optical link provided by the mode-division multiplexing (MDM). Latterly, progress has been made focused on the integrating mode and WDM on a chip [127–133].

In an attempt to offer new dimensions of data transfer with the aim of fulfilling the growing need for speed, an integrated multi-dimensional system that integrated wavelength and mode multiplexing on a silicon photonic circuit for the on-chip and chip-to-chip interconnects was put forward [59] (Fig. 7d). A multi-wavelength laser source was evenly distributed into multiple WDM transmitter circuits with each WDM circuit encoding data independently onto different frequencies of light. An inverse-designed MDM multiplexer took the overlapping modes from the multiple WDM transmitters, and after that they were transformed into copropagating spatially orthogonal modes. The data could then be transmitted through chip-to-fiber couplers and multimode fiber to the receiver. The MDM-WDM demultiplexers were used to

https://doi.org/10.1007/s40820-025-01850-w

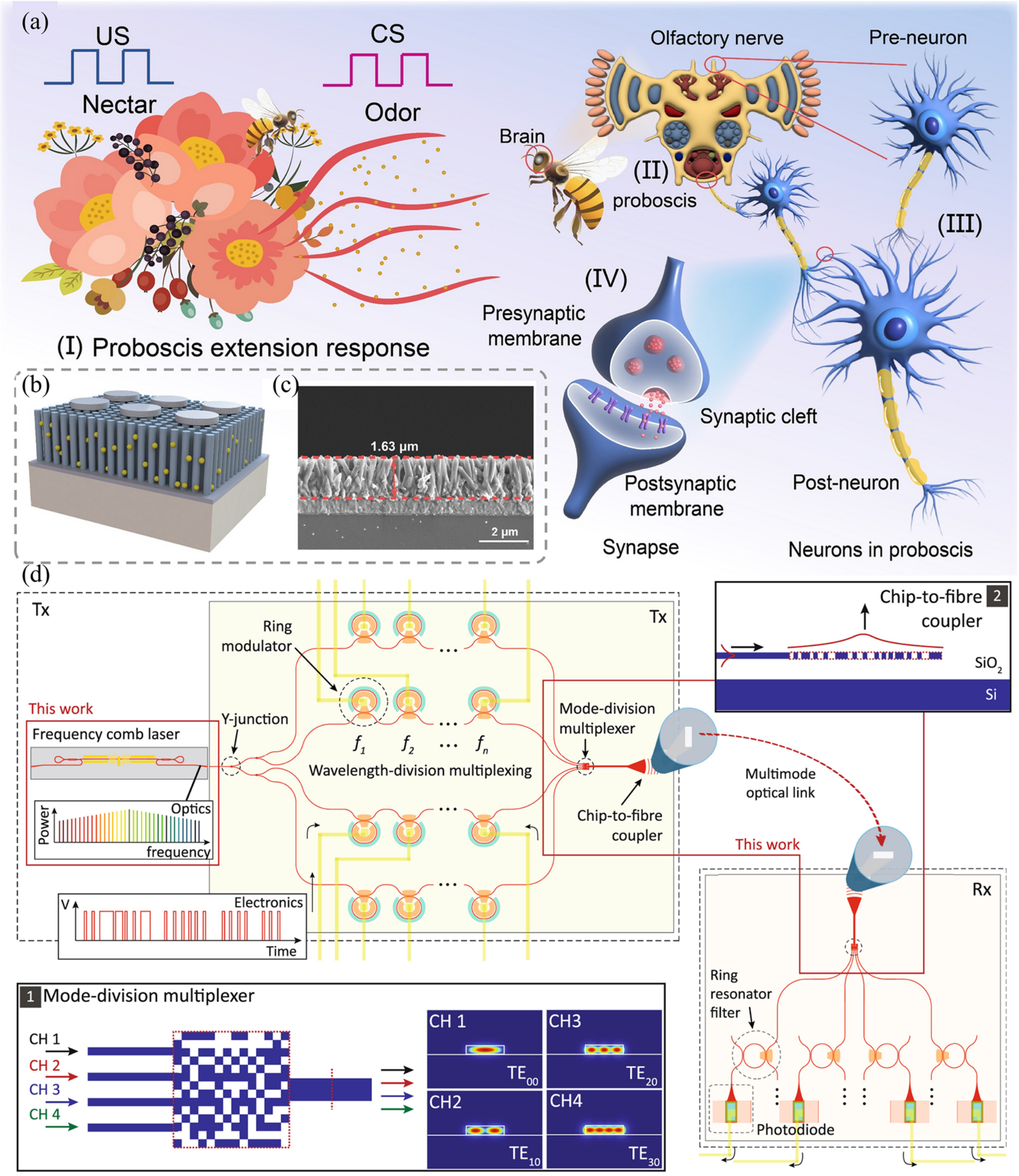


Fig. 7 Schematic for the design strategies of AI chips in regard to data memory and transfer. **a** Schematic illustration of the proboscis extension response. **b** Schematic of the ATZ-based device. **c** SEM of the as-prepared device. **a**–**c** Reproduced with permission from Ref. [36] Copyright 2024, Springer. **d** Schematic illustration of the multi-dimensional communication. **d** Reproduced with permission from Ref. [59] Copyright 2022, Nature

separate the mode and wavelength channels, and photodiodes were taken advantages of for detection. It was verified that a 1.12-Tb/s natively errorfree data transmission could be fulfilled.

4.3 At the Computing Stage

Dynamic computing is a promising approach in DL, and the dynamic neural networks are managed to adapt the computational graphs to the input in the inference stage, showing the attractive properties in many aspects [134]. The neuromorphic and traditional AI systems are two typical paradigms for dynamic computing [37]. Particularly, neurons in SNNs communicate through spike trains, and the spike-based neuromorphic computing is naturally featured with a dynamic computational graph, with only a small portion of the overall spiking neurons being active at any moment and the rest being idle. In contrast, the neurons in traditional Artificial Neural Networks (ANNs) exchange information via continuous values and are controlled by static computational graphs. As a result, dynamic algorithms are developed to implement dynamic computing (Fig. 8a, b).

The energy constraints become a major restriction to deploy traditional AI methods, and therefore high demand for the energy efficiency has also been proposed for the computing. Correspondingly, much efforts have been made to come up with the schemes for energy-efficient computing. For example, better energy efficiency can be offered by analog in-memory computing (analog-AI) as it can perform matrix–vector multiplications (MVM) in parallel on 'memory tiles' [1]. Besides, the neuromorphic computing provides a promising way for energy-efficient machine intelligence by learning from the way by which information is processed via brain, taking advantages of artificial neurons and the SNNs on neuromorphic chips. The neuromorphic computing meets the challenges of how to learn from the high-level brain dynamic mechanisms to realize the excellent computational efficiency [37].

In addition to the requirement from dynamic computing and energy constraints, high demand has also been put forward for the weight-reconfigurable capacity of computing for some fields, like the healthcare monitoring, on which occasion it is essential to finely reconfigure the relative intensity of weight from each input. In an attempt to achieve the precise and independent modification of each input, a neuromorphic computing system that was managed to integrate two different environmental information with reconfigurable weights by making use of a simple circuitry based on electrochemical artificial synapses was designed [39]. From the perspective of dealing with various environmental information, a complex logic circuit was essential with the increased complexity of the processor, since more environmental factors need to be taken into consideration for a conventional CMOS-based processor, while a single device was managed to handle these environmental information by neuromorphic computing with an electrolyte-based multi-input synapse, which is demonstrated in Fig. 8c. Schematic illustration of the neuromorphic signal integration system is shown in Fig. 8d. To be specific, the sensors were responsible for the transform of the raw data into electrical signals, and then a weight control circuit was made use of to assign weights to the signals. The processing synapse could then integrate the signals, and finally a logical decision could be made by the artificial neuron. Correspondingly, the schematic signal flow of this system is demonstrated in Fig. 8e. Action was executed if the synapse output exceeded the level of the criteria. It was noticeable that the potentiation of the processing synapse was modulated with the different weights for signals, leading to the different final action state even for the same environmental signals. A hydrogen explosion risk assessment system was designed accordingly, with the schematic circuit diagram shown in Fig. 8f and the photographic image demonstrated in Fig. 8g. Hydrogen concentration and temperature were used as the inputs, and the signals were then updated by the weight control circuit, after which procedure they were converted into a postsynaptic current to represent the hydrogen explosion risk by taken advantages of the multi-input artificial synapse.

5 Design Considerations for Future Advanced and AI Chip

A sharply increased calculations have been brought about with the development of AI technology [39]. The prosperity of AI is largely empowered by a significant amount of parameters and improved computing powers [34]. As to many vision tasks, short exposure time is essential to complete the tasks with ultra-low latency, calling for extremely high computing power [27]. In addition, the computing capability and energy efficiency are critical issues which need to be balanced for high-performance computing [135].

5.1 For High-Performance Computing

5.1.1 To Accelerate Computing Speed

The computing speed should be further accelerated to cooperate with the improved performance of various tasks at the algorithmic level [13, 136]. Large-bandwidth and high energy efficiency computing can be achieved by optical AI for which optics and photonics are fully leveraged. A fact that cannot be ignored is that digital devices remain to be

https://doi.org/10.1007/s40820-025-01850-w

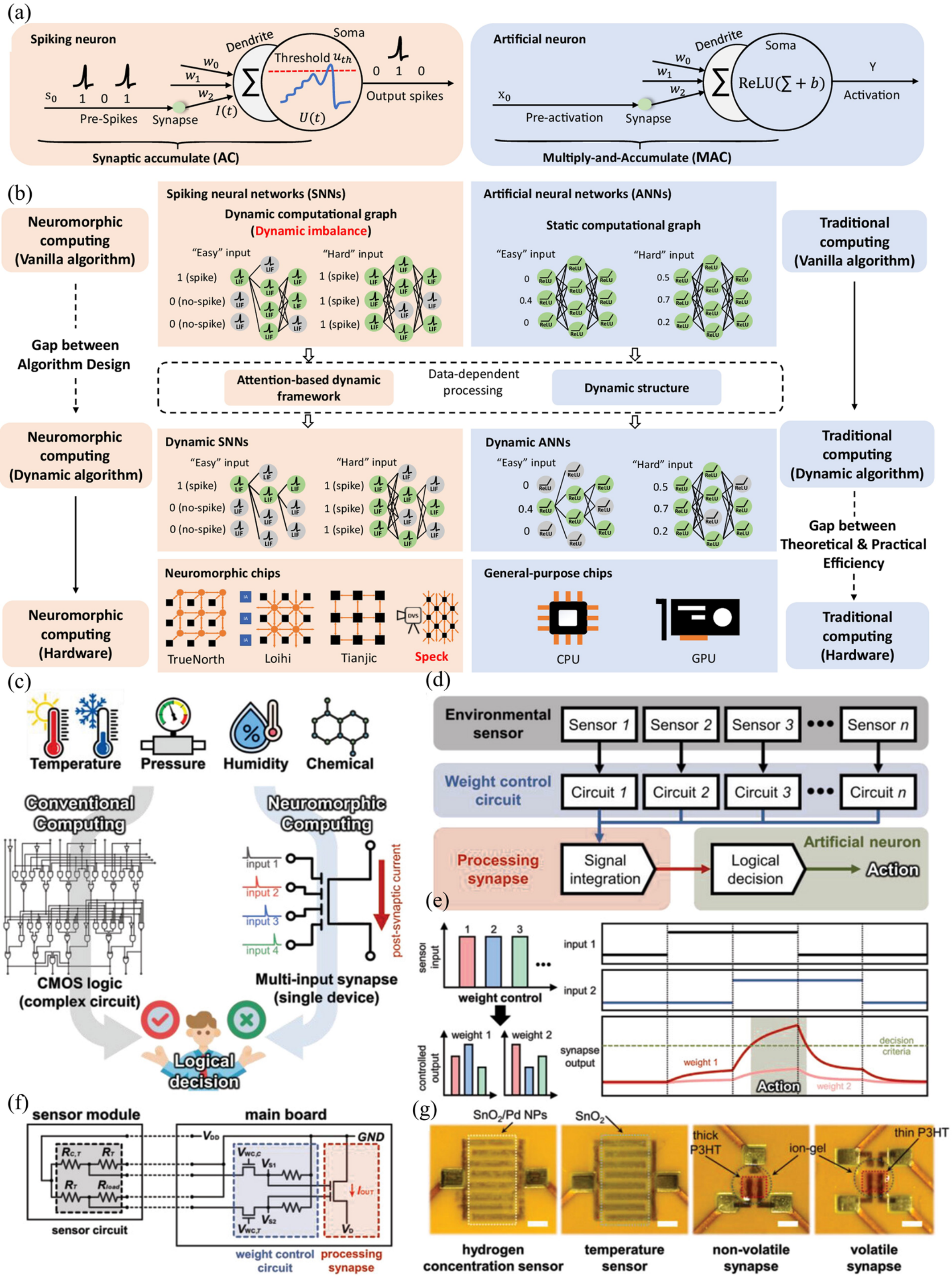
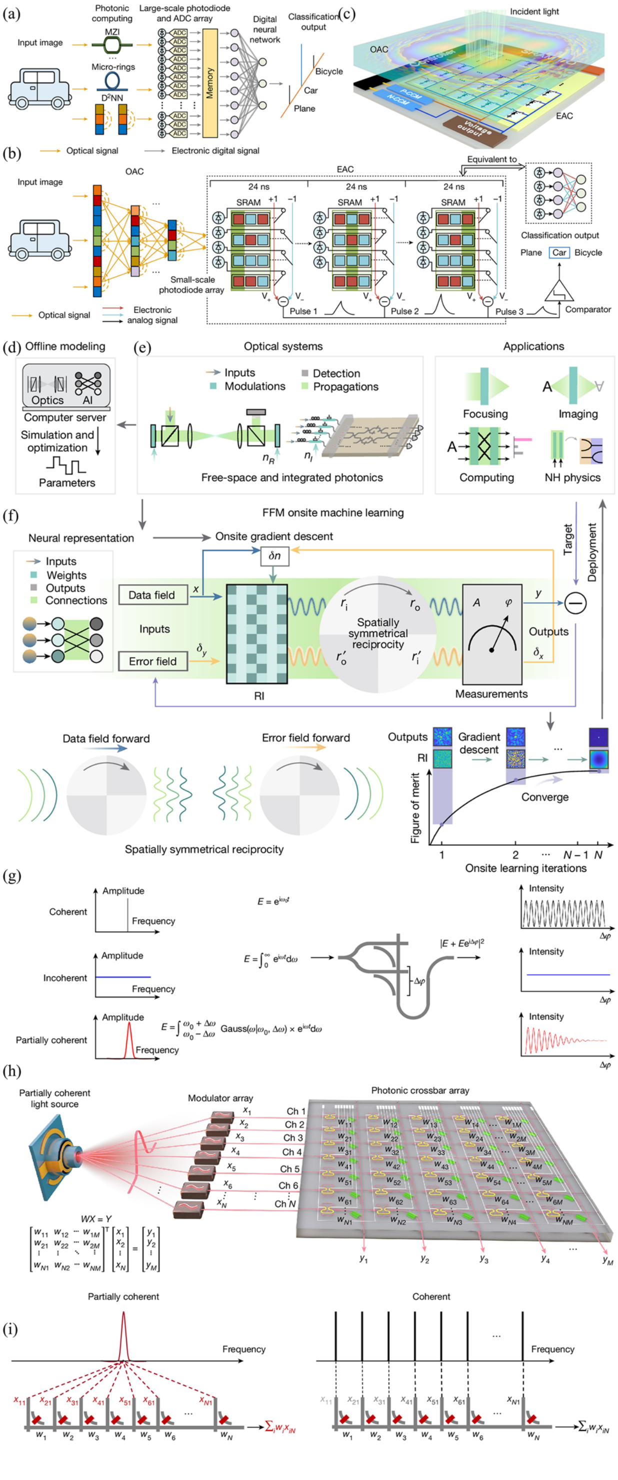


Fig. 8 Schematic for the design strategies of AI chips in regard to computing. Comparison between **a** spiking neuron and artificial neuron, and **b** the neuromorphic and traditional computing for a dynamic computing. **a**, **b** Reproduced with permission from Ref. [37] Copyright 2024, Nature. **c** Comparison between conventional and neuromorphic computing. **d** Schematic illustration and **e** schematic signal flow of the neuromorphic signal integration system. **f** The circuit diagram and **g** photographic image of the hydrogen explosion risk assessment system. **c–g** Reproduced with permission from Ref [39]. Copyright 2024, Wiley–VCH GmbH

Fig. 9 Design considerations of AI chips for high-performance computing. The workflow of **a** traditional optoelectronic computing, and **b** all-analog photoelectronic computing. **c** Schematic diagram of the all-analog photoelectronic chip. **a–c** Reproduced with permission from Ref. [27] Copyright 2023, Nature. **d** Schematic diagram for the conventional optics-related AI and **e** the general optical systems. **f** Schematic illustration of FFM onsite ML. **d–f** Reproduced with permission from Ref. [35] Copyright 2024, Nature. Schematic illustration of **g** a generalized unit cell with coherent light sources, and **h** the proposed photonic convolutional processing system with partially coherent light. **i** Schematic illustration for the *N*-fold enhancement in regard to parallelism. **g–i** Reproduced with permission from Ref. [43] Copyright 2024, Nature ▸



the mainstream, and therefore it is essential to convert the optical signals into digital ones for vision tasks even after optical computing by means of large-scale photodiodes and power-hungry ADCs to conduct the necessary postprocessing procedures [27] (Fig. 9a). In an effort to address this issue, an optoelectronic hybrid architecture was designed, which was managed to reduce massive ADCs, and therefore vision tasks could be accomplished in a power-efficient and high-speed manner (Fig. 9b). To be specific, the information was encoded into light fields. The features of high-resolution images were extracted by using a multi-layer diffractive optical computing module at light speed, which was optical analog computing (OAC). It was worthwhile mentioning that the demand for optoelectronic conversion could be reduced by dimension reduction all optically. The electronic analog computing (EAC) with a 32 × 32 photodiode array was then introduced to convert optical signals into analog electronic ones due to the photoelectric effect, working as a nonlinear activation. These photodiodes are either connected to the V_+ positive line or V_- negative line according to the weights in the static random-access memory (SRAM). Based on Kirchhoff's law, the generated photocurrents were summed up on both lines, after which process the differential voltage of the computing lines V_+ and V_- was calculated by the analog subtractor as the output node. It was noticeable that by means of resetting the computing lines and updating weights, this system can output another pulse with different connections of photodiodes. The output could be used either as predicted labels of classification categories or as inputs of another digital neural network. Schematic diagram of the all-analog photoelectronic chip is demonstrated as Fig. 9c.

Another challenge met by the optical computing is that they are implemented in silico on electronic computers, and therefore both strict modeling and large amounts of training data are essential (Fig. 9d). In particularly, optical AI primarily includes the optical emulation of electronic ANNs, and the photonic architecture design is conducted on electronic computers [24, 137]. Accordingly, it proposes the challenge of correcting the experimental system error which calls for extensive work to characterize the optical propagation spatially and temporally [83, 96, 98, 138]. As to AI empowered optical design, the system must also be modeled analytically or implicitly [139–141]. It consumes more time for analytical and numerical modeling with the increase of the system complexity. It is pointed out that the precise modeling of a general optical system is difficult to be achieved due to the system imperfections and the complexity of light-wave propagation. Some efforts have been made to address these issues [35]. FFM learning was developed, which mapped optical systems to parameterized onsite neural networks. It was worthwhile mentioning that by taking advantages of spatial symmetry and Lorentz reciprocity, the necessity of backward propagation in the gradient descent training was eliminated. Specifically, as for general optical systems, free-space lens optics and integrated photonics were contained, with the modulation regions marked as dark green and propagation regions demonstrated as light green, in which occasion the refractive indexes were respectively, tunable and fixed (Fig. 9e). These regions in the optical system could be mapped to weights and neuron connections, which made it possible to construct a differentiable onsite neural network between the input and output (Fig. 9f).

5.1.2 To Realize High-Capacity Signal Processing

In addition to the method mentioned above, parallel multithread processing is also one of the key approaches to achieve high-speed and high-capacity signal processing, which is a promising way to meet the increasing demand for high-capacity datasets processing [142]. Recently, a photonic convolutional processing system using partially coherent light to realize boost computing parallelism without substantially sacrificing the accuracy has been proposed [43]. It was pointed out that a variety of system architectures for photonic convolutional processing was developed with the coherent light sources being applied in all of these cases. However, the operation of the coherent nanophotonic circuits needed the precise control of numerous phase shifters so that the desired coherent interference in the circuit could be achieved. A generalized unit cell to perform multiply-and-accumulate operations is illustrated in Fig. 9g, while a system with partially coherent light for parallelized photonic computing is proposed as Fig. 9h. It was worthwhile mentioning that for the system with partially coherent light for parallelized photonic computing, the coherent light source was not necessary, leading to less rigorous feedback control and thermal-management requirements. As for the partially coherent system, a Gaussian-shaped optical carrier could be sent to all input channels and summed in a bus waveguide, while for a coherent system, different input channels should receive optical carriers at distinct wavelengths to avoid intensity fluctuation. As a result, one MVM operation for input vectors of dimension *N* called for only one optical band for partially coherent system, while *N* optical bands were required with coherent light being applied, making it possible for the *N*-fold enhancement in parallelism as using partially coherent light (Fig. 9i).

5.2 With Improved Energy Efficiency

5.2.1 General Approaches to Improve the Energy Efficiency

In addition to the enhanced computing performance, the high energy efficiency is another important requirement for the advanced chips. For example, in regard to many vision tasks, the ADCs with high throughput and high precision reduce the imaging frame rate on account of limited data bandwidth, causing remarkable energy consumption [27]. Accordingly, efforts have been made on the design of an optoelectronic hybrid architecture in an all-analog way, to reduce the massive ADCs for the accomplishment of power-efficient vision tasks. Furthermore, neuromorphic computing tends to be a promising approach for energy-efficient machine intelligence by simulating the neurons of the human brain and using spiking neural networks [37]. It is proposed that the human brain is managed to allocate its resources dynamically according to the required demand [143, 144]. As a result, greater attention is paid to salient stimuli, which is proved via the heightened spiking activity of the brain regions or neurons associated with the stimulus. Additionally, endeavors have also been made to design the neuromorphic chip with no needs for the global or local clock signal, which efficiently avoids the redundant power consumed by the clock empty flips [37]. Furthermore, it is worthwhile mentioning that CIM is important in the field of AI, for which both the memory and processing functions can be integrated within the same module, leading to the enhanced efficiency. Memristors, which are featured with their striking similarity with biological counterparts in the aspect of device dynamics, play an important role in this field [145].

5.2.2 Analog In-memory Computing

The vast amounts of data transferred between memory and processor lead to the unessential energy consumption. Both the time and the energy are expected to be saved by the Analog-AI hardware with the function to apply arrays of non-volatile memory (NVM) to execute the MAC operations. One case in point was that an analog-AI chip was designed to recognize and transcript speech energy efficiently. It was noticeable that not only the fully end-to-end SW_{eq} accuracy for a small keyword-spotting network but also the near-SW_{eq} accuracy on the much larger MLPerf RNNT was verified [1]. Particularly, the tiny-model task of keyword-spotting network (KWS) on the Google speech-commands dataset was targeted. The MLPerf version of RNNT, which was a large data center network, was implemented on Librispeech. It was worthwhile mentioning that the model contained 45 million weights, which was implemented by more than 140 million PCM devices across five chips. This system demonstrated excellent power performance. To be specific, Chip 4 showed the best power performance of 12.40 TOPS/W, which was attributed to the most on-chip weights (Fig. 10a). It was proposed that there existed a correlation between the reported TOPS/W and the number of weights that were encoded on-chip. Another 25% improvement in TOPS/W could be achieved for chip 4 caused by the reducing the maximum input duration without large WER degradation, which is illustrated in Fig. 10b. Energy efficiency at different levels is illustrated in Fig. 10c, which reflected how the costs of data communication, incomplete tile usage, as well as the inefficient digital computing resulted to the fact that the large peak TOPS/W of the analog tile itself was down to the final sustained value of 6.94 TOPS/W. The full processing time of the overall system was estimated (Fig. 10d). It was noticeable that the average processing time for each sample was more than 10^4 times faster than the actual speech time, leading to a real-time factor of only 8×10^{-5}. Number of operations performed on-chip versus off-chip in the RNNT experiment is shown in Fig. 10e. In contrast to the MLPerf submissions, a 14-fold improvement was managed to be realized by this system in regard to the samples per second per watt and TOPS/W (Fig. 10f).

5.2.3 Dynamic Computing with Asynchronous Chip

To reach the goal of energy efficiency, the composition of different power consumption should be taken into considerations. The power that is required to operate an AI system is usually composed of two aspects, resting power which is determined by the hardware design, and running power which relies on the model as the hardware is fixed [37] (Fig. 10g). It is proposed that for the great majority of hardware a significant amount of energy is consumed even when no computing is being done, leading to very high ratio of the resting power to the overall power. Consequently, it is difficult to reduce the overall power only by reducing the running power (Fig. 10h, i). To be specific, the chip architecture (asynchronous/synchronous) can leave an impact on the power consumption, and it has been proposed that the asynchronous architecture, for which the change of the circuit state is only caused by the change of the external input, is featured with the advantage of low power consumption compared with synchronous circuits. The event-driven mechanism is an approach for asynchronous chips to coordinate the work of each module. When taking the design strategies for sensing-computing chip into considerations, event-driven chips can be made use of, since the sensor can only wake up the chip when the environmental changes (such as temperature changes or motion triggers) are detected to complete data collection and transmission, leading to the improved energy efficiency and low latency.

In contrast to the most common neuromorphic hardware design which begins with the bottom of the compute stack, elaborated design can be conducted for the customization of the neuromorphic hardware which is to be applied at the edge for the specific purposes with low power consumption taken into consideration. One case in point was that a sensing-computing neuromorphic chip, Speck, was designed with a 128 × 128-pixel DVS integrated onto an asynchronous spike-based AI chip, which is shown in Fig. 10j. Speck was a sensing-computing end-to-end SoC with the always-on hardware applicable to various scenarios, such as Internet of things, smart travel, smart home, intelligent robotic, and so on (Fig. 10k, l). It was worthwhile mentioning that its processing pipeline was built with asynchronous digital logic,

https://doi.org/10.1007/s40820-025-01850-w

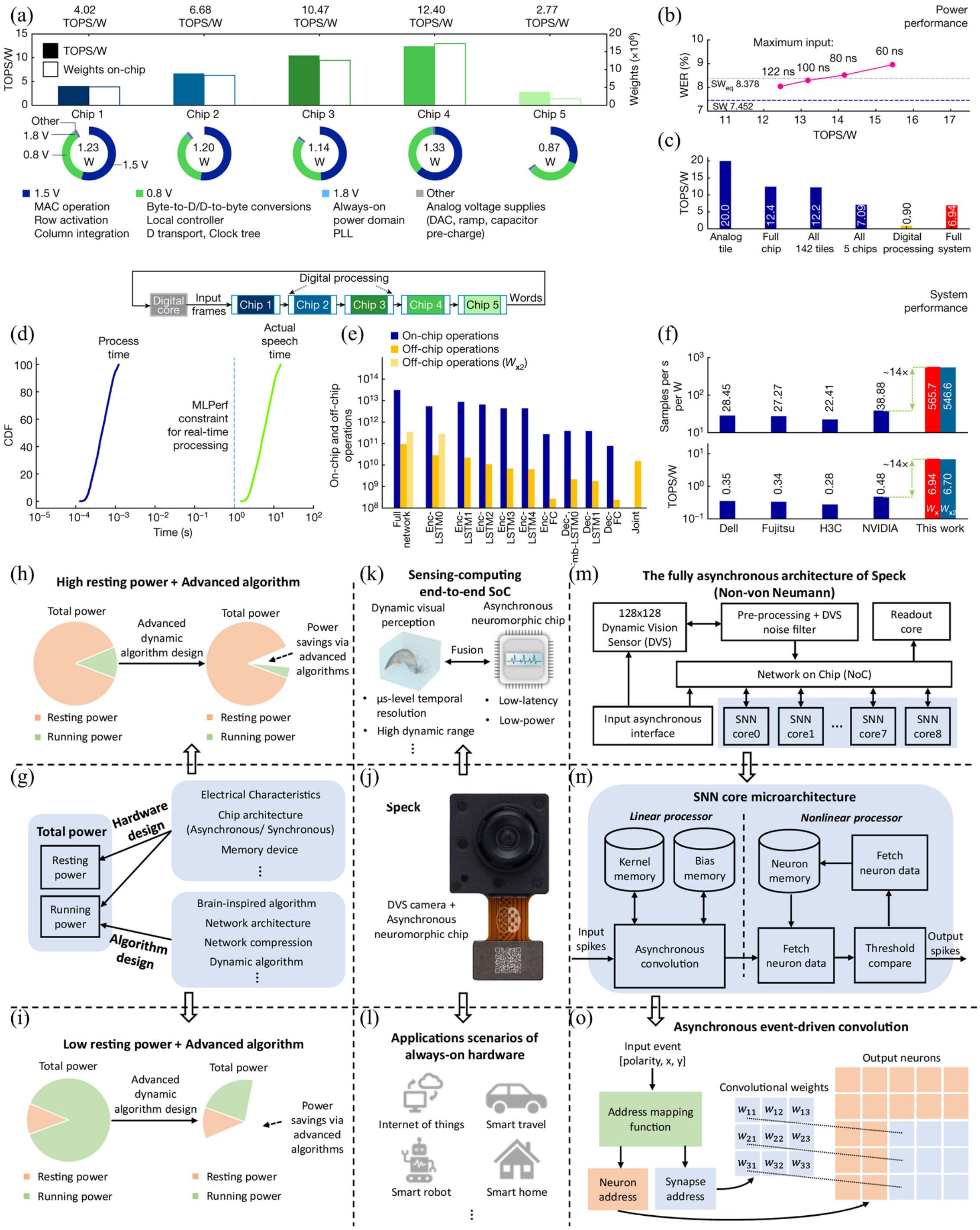


Fig. 10 Design considerations of AI chips with improved energy efficiency. **a** Measured power and TOPS/W corresponding to each chip. **b** An improvement in TOPS/W caused by the reducing the maximum input duration. **c** Energy efficiency at different levels. **d** Processing time and actual speech time. **e** Number of operations performed on-chip versus off-chip in the RNNT experiment. **f** Samples per second per watt and TOPS/W compared with MLPerf submissions. **a**–**f** Reproduced with permission from Ref. [1] Copyright 2023, Nature. **g** Power composition of AI systems. The case of **h** high resting power and **i** low resting power. **j** Physical display of Speck. **k** Illustration for the sensing-computing end-to-end SoC, and **l** its application scenarios. **m** Fully asynchronous architecture of Speck. The design of **n** SNN core, and **o** the asynchronous event-driven convolution. **g**–**o** Reproduced with permission from Ref. [37] Copyright 2024, Nature

which made it possible for the chip to realize always-on low resting power consumption and optimum latency. To address the issue that the implementation of asynchronous circuits is complicated, the overall sensing to computing strategy was optimized. There was a central event router which is able to be configured to route events from any to any of the 9-SNN cores, and every core was managed to work independently and asynchronously, which was illustrated in Fig. 10m. As a result, the design effort could be limited to a single SNN core (Fig. 10n). Additionally, the asynchronous event-driven convolution was included as one of the core designs for the improvement of the computational efficiency as well (Fig. 10o).

6 Perspectives

Overall, the recent development, including but not limited to the co-design strategies for the software and hardware, the realization of enhanced overall performance, and the potential for broader application have been reviewed in depth. Great progress has been made in the field of advanced chips due to the high challenges brought by AI, which has revolutionized various aspects, ranging from information industry to material science. To execute the complex algorithmic programs and advanced tasks proposed by these new challenges, the elaborate design of chips covers every aspect, including materials, algorithm, architectures, processing technology, integrating method, and so on. Progress has been made on developing novel materials and models, as well as overcoming the shortcomings of the existing conventional materials and architectures for chips. New fabrication processes for both the production and the package of the devices have been developed, aiming to induce the cost and develop complex chips. The advanced chips are qualified to be applied for video recognition tasks, speech recognition and transcription, visual memory and many other fields, offering fast and efficient information processing functions (Fig. 11).

Summary for the state-of-the-art advanced and AI chips is illustrated in Table 1 with the performance, scales, other properties, and applications included. The quantitative indicators of the chips are critical to the systems. To be specific, energy efficiency refers to the effective amount of work completed by a chip with per unit of energy consumed when implementing a task, which makes sense for the environmental sustainability. The computing speed of a chip is the core indicator for measuring its data processing capability, which is important for shortening the task processing time and supporting complex tasks. For AI training which needs to handle large amounts of parameters, chips with high computing speed are managed to shorten the training cycle, accelerating technological iteration. The latency of a chip refers to the time interval from the triggering of an input to the generation of an effective output, which is a key indicator for measuring the response speed of a chip. While ensuring high energy efficiency and computing speed, reducing latency has become another challenge in chip design, which is especially essential for some real-time tasks. Besides, the abilities of integrating more transistors, realizing a larger area, or expanding to more application fields are also imperative for these systems. For example, the scale expansion of chips is in relevant to the change from achieving a single function to multi-functions or from small-scale to large-scale applications, which can leave impacts on a series of factors, like cost, power consumption, design complexity, and so on.

Significant improvements of the advanced chips have happened and accompanied by the discovery of novel modes, the improvement of the package techniques, the accelerating of the efficiency, as well as the enhancement of computing

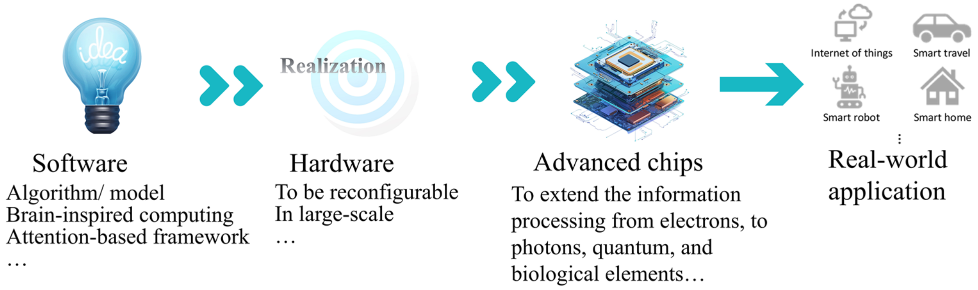


Fig. 11 Outlook of the advanced chips

https://doi.org/10.1007/s40820-025-01850-w

Table 1 Summary for the performance of the state-of-the-art advanced chips

	Type	Energy efficiency (TOPS/W)	Computing speed (TOPS)	Latency	Scale	Accuracy	Other key features	Application	References
Neuromorphic chip	Analog-AI chip	12.4	–	2.4 μs for each audio frame	To combines 35 million phase-change memory devices across 34 tiles and	With fully end-to-end SWeq accuracy for a small keyword-spotting network	To show a 14-fold improvement compared with traditional ones, and to demonstrated a WER of 9.258%	Speech recognition and transcription	[1]
	Integration of trainable dendritic neurons and high-density RRAM chip	–	–	380 ns	–	~90%	To realize 516× and 1.3×10^5× improvements on the LAE (LA E = Latency^{-1} × Area^{-1} × Energy^{-1}) FoM when compared to digital and analog CMOS activation circuits	For CIM-based neuromorphic computing	[34]
	Neuromorphic hardware	–	–	–	A 3×7 memristor array	An accuracy of 88.9% for handwriting digit recognition	To realize complex biological associative learning behaviors	Visual memory application	[36]
	Sensing-computing neuromorphic chip	–	–	Less than 0.1 ms	To be an efficient medium-scale neuromorphic sensing-computing edge hardware	92%	With the low processor resting power of 0.42 mW and real-time power as low as 0.70 mW	As edge computing devices for smart home application scenarios	[37]
	Neuromorphic computing systems	–	–	–	–	–	To be weight-reconfigurable	For hydrogen explosion risk assessment	[39]
	Neuromorphic optoelectronic computing system	1.58	240.1	–	–	With a blind-testing accuracy of 97.6% on 10,000 digit images	To be reconfigurable	For high-speed image and video recognition	[83]
Photonic chip	Photonic computing	1	0.108		With a 3×3 photonic tensor core, using phase-change-material photonic memories	92.2% accuracy (92.7% theoretically)	To boost computing parallelism while maintaining the accuracy	To classify the gaits of ten patients with Parkinson's disease with	[43]

Table 1 (continued)

Type	Energy efficiency (TOPS/W)	Computing speed (TOPS)	Latency	Scale	Accuracy	Other key features	Application	References
A silicon photonic circuit	–	–	–	Multimode optical transmission between separate silicon chips	–	With a 1.12-Tb/s natively errorfree data transmission	Silicon photonic transmitters	[59]
A trainable diffractive optical neural network	7.28	217.6	30.2 ps	With 1 × 16 neurons input, 5 × 16 neurons hidden and 1 × 4 neurons output layers	85.7% accuracy for multi-modal test sets	With high computing density (447.7 TOPS/mm^2)	To accomplish four-class classification in different modalities	[41]
Photonic chiplet	160	–	3.79 ms	With 4256 total neurons and a net scale of 13.96 million	Testing at 91.89% accuracy in the 1623-category Omniglot dataset	To experimentally achieve on-chip 1000-category-level classification and high-fidelity AI-generated content with up to two orders of magnitude of improvement in efficiency	For large-scale photonic computing and artificial general intelligence (AGI)	[38]
Photonic convolutional accelerator	–	11.3	< 200 ps	–	With an accuracy of 88% for recognition of handwritten digit images	For generating convolutions of images with 250,000 pixels	For real-time video recognition	[22]
An integrated photonic tensor core	0.4	–	–	With the matrix size being easily be scaled up to 40 × 40	With an accuracy of 95.3%	With computing densities of more than 400 TOPS per mm^2	For parallel convolutional processing	[24]
An on-chip photonic DNN	–	0.27	570 ps	To be scaled to a classifier with a larger number of pixels	With an accuracy of 93.8% for two-class classification of handwritten letters	With a classification time of under 570 ps	For image classification	[25]
Photonic processing unit	0.2	–	–	–	With the accuracy 96.6% of recognition	With a preeminent photonic-core compute density of over 1 TOPS mm^{-2}	For image reconstruction, video action recognition, and autonomous driving	[149]

Table 1 (continued)

	Type	Energy efficiency (TOPS/W)	Computing speed (TOPS)	Latency	Scale	Accuracy	Other key features	Application	References
	All-analog photoelectronic chip	7.48×10^{4}	4.6×10^{3}	72 ns for each frame	With two 400 × 400 SiO_2 OAC layers and a 1,024 × 3 EAC layer	92.6% for time-lapse video recognition task	With superior system robustness in lowlight conditions (0.14 fJ μm-^{2} each frame)	Time-lapse video recognition task	[27]
	All-optical processing	5.40×10^{6}	–	–	–	94.5%	To facilitate orders-of-magnitude-faster learning processes	To design non-conventional imaging modalities	[35]
	Chip integrated meta surfaces	–	–	–	–	–	With the potential to be compatible with on-chip optical systems and to independently encode multiple optical parameters	For multidimensional encryption	[150]
Quantum chip	Quantum simulator	–	–	–	–	–	To realize the stable trapping of 512 ions in a 2D Wigner crystal	To run noisy intermediate-scale quantum algorithms	[151]
Silicon-based chip	Biomimetic olfactory chips	–	–	–	With 10,000 individually addressable sensors per chip	With a prediction accuracy of up to 99.04%	With distinguishability of mixed gases and 24 distinct odors	To be integrated with vision sensors on a robot dog	[152]
	Si-based optical memristive crossbar array	–	–	–	With a 5 × 5 optoelectronic synapse array	With a classification accuracy of 98.02%	To enables an ultralow power (2.8 × 10–^{13} J) fine-tuning process	For patient-specific issues	[153]

power. This review offers a keen insight into the design strategies for the advanced and AI chips, with some perspectives for the chips applied in the future proposed as follows:

1. Endeavors have been made to equip the AI chips with more intelligent performance learning from biology. a) Efforts have been made focused on mapping the biological behavior to the electrical behavior in devices. It is expected for the systems to realize more complex biological performances. The associative learning behavior, which is commonly found in the cranial nerves of insects and is featured with the acquisition, extinction, restoration, and generalization, has been simulated by ZnO QDs-based optoelectronic memristors, which provide novel scheme for the field of machine self-learning. It is desirable to develop chips learning from more advanced behaviors of the creatures. b) Extensive investigations have been carried out on neuromorphic devices based on the human brain, which is a potential candidate for the next-generation computer architecture. The method of how to learn from the high-level brain dynamic mechanisms to equip neuromorphic computing with more energy advantages is always in high demand. Endeavors have been made from both the software and the hardware aspects to address this issue. Moreover, chips used for dealing with image information are expected to be managed to handle the dynamic, diverse, and unpredictable scenes in real application scenarios, like autonomous driving. It is desirable to design the chips that are efficient in various fields to percept and address even the difficult issues existing in the real world. In particular, the dynamic computing, which is a critical feature of human brain, has been simulated by this system. In the future, more advanced strategies can be adopted for the realization of high-level brain dynamic mechanisms to fully achieve the brain advantages in many aspects.
2. Efforts can be made to make full use of the novel modes that extend the information processing from electrons, to photons, quantum, and biological elements, by taking advantages of the strengths and overcoming their weaknesses. a) Photonics-based systems are managed to provide high-speed computing units, and therefore efforts have been made focused on the algorithms design to exploit their unique advantages. For instance, approaches have been developed to realize the high throughput and precision by the successful application of cellular automata [146]. Ultrafast silicon photonic reservoir computing engine has been developed, which paves the way for high-speed photonic computing [147]. For photonic computing, to truly become a leading technology in the field of AI, a series of key challenges still need to be meet which mainly lies in the aspect of integration, dynamic reconfiguration capability, standardization, and cost issues. In particular, the compatibility of silicon-based photonic chips with the existing CMOS processes needs to be optimized, and the capacity of photonic chips to dynamically adapt to different tasks is expected, since the hardware of photonic chips is relatively fixed. b) Low power consumption and real-time requirements have promoted the application of CIM in many fields, like intelligent sensors and IoT. For example, some progress has been made for cryogenic in-Memory Computing recently [148]. In the future, more endeavors can be made to enhance the computing abilities of the memory by making use of new materials, such as two-dimensional materials and oxide semiconductors, and optimizing the circuit architectures. Besides, 3D packaging can also be applied for CIM to obtain the systems with excellent overall performance. c) Additionally, cellular computing has emerged focused on the analysis and modeling of real cellular processes to implement computing with the aspects of information processing and adaptation. Attempt has been made on the reprogrammable circuits that are managed to increase circuit flexibility and realize the scalability of complex cell-based computing devices. The feasibility of proposing several circuits by making use of only a small set of engineered cells that can be externally reprogrammed to implement simple logics in response to the specific inputs has successfully been proved. In the future, more efforts can be made focused on taking advantages of biological circuits to implement logics and meet numerous biological challenges.
3. The advanced chips that are qualified for real-world applications are always in high demand. Multi-input signals are usually needed to be processed properly by the advanced processors suitable for diverse external information in the open-world applications. The integrated signals from different input are needed to be handled accurately and timely. The version of GPT-4 has successfully accomplished the processing of multimodal data, like images and audio. A neuromorphic computing system applied for the risk assessment has been developed with several kinds of factors taking into considerations. In the future work, more work focused in the development of algorithms and hardware tailored for open-world applications can be conducted. The overall performances are expected to be enhanced for the chips to meet the high requirement proposed by the real-world applications.
4. The reconfigurable behavior is an important aim for computing hardware. For the chips with reconfigurable

https://doi.org/10.1007/s40820-025-01850-w

capacities, their function can be changed even after the accomplishment of the fabrication, and therefore multimodal data and different tasks can be dealt with, making the high flexibility in adapting to different tasks feasible. It is especially critical to the chips used for some specific purposes like healthcare monitoring, for which it is imperative to finely reconfigure the relative intensity of weight updates from each input. Explorations have been made to equip different types of chips with strong reconfigurability. The reconfigurability and multimodal capability have been achieved for a TDONN chip by taking advantages of on-chip diffractive optics with massive tunable elements. The reconfigurability has also been available for the diffractive-interference hybrid photonic chiplet, which is acted as the fundamental building block for a diversity of advanced ML tasks, with 1000-category classification and content generation included. An all-analog chip combining electronic and light computing (ACCEL) is also equipped with the reconfigurability for different tasks without changing the OAC module. The integration of two different information with reconfigurable weights has been accomplished by a neuromorphic computing system. In the future, the high degree of adaptability to different assignments empowered by reconfiguration is expected to be accessible for more chiplet when it is necessary.

5. More explorations on large-scale integrations are expected to be made for chips. With the increasing of information, chips are required to be integrated to an ever-growing level to process the booming signals. The large-scale integrations of various chips are indispensable to getting rid of the shortcomings of each chip. For inorganic counterparts, like CMOS chips, an integration level in ultra-large-scale has been realized, while poor mechanical compatibility with organisms exists. It is ideal for the devices to overcome inherent shortcomings and accomplish the large-scale integration. Moreover, the integrations are closely related to the technologies. A diversity of techniques like photolithography, screening, printing, and shadow-mask evaporation has been developed. In the future, the continuous progress of the techniques is expected to be made in order to miniaturize these devices.
6. The application of sustainable materials in AI chips is one of the most important trends in this field with the aim of reducing the environmental impact and improving energy efficiency. Efforts can be made from various aspects, such as selecting degradable substrates, developing environmentally friendly manufacturing process, preparing environmentally friendly heat dissipation materials, and so on. Some bio-elastomers with active-controllable degradation rates have been designed, which can be applied as the bio-electronic substrates and encapsulation layers. In the future, more endeavors can be made to make a balance between meeting the high-performance requirements of AI chips and controlling the costs when using sustainable materials.

Acknowledgements This work was supported by the Hong Kong Polytechnic University (1-WZ1Y, 1-W34U, 4-YWER).

Author Contributions Ying Cao helped in methodology, investigation, writing—original draft. Yuejiao Chen and Xi Fan contributed to methodology.. Hong Fu helped in resources, methodology, writing—review & editing. Bingang Xu was involved in conceptualization, funding acquisition, methodology, supervision, writing—review & editing.

Declarations

Conflict of interest The authors declare no interest conflict. They have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. S. Ambrogio, P. Narayanan, A. Okazaki, A. Fasoli, C. Mackin et al., An analog-AI chip for energy-efficient speech recognition and transcription. Nature **620**(7975), 768–775 (2023). https://doi.org/10.1038/s41586-023-06337-5
2. L. Gao, J. Lin, L. Wang et al., Machine learning-assisted design of advanced polymeric materials. Acc. Mater. Res. **5**(5), 571–584 (2024). https://doi.org/10.1021/accountsmr.3c00288
3. J. Benavides-Hernández, F. Dumeignil, From characterization to discovery: artificial intelligence, machine learning and high-throughput experiments for heterogeneous catalyst design. ACS Catal. **14**(15), 11749–11779 (2024). https://doi.org/10.1021/acscatal.3c06293
4. Y. Fu, A. Howard, C. Zeng, Y. Chen, P. Gao et al., Physics-guided continual learning for predicting emerging aqueous

organic redox flow battery material performance. ACS Energy Lett. **9**(6), 2767–2774 (2024). https://doi.org/10.1021/acsenergylett.4c00493

5. P. Akbari, M. Zamani, A. Mostafaei, Machine learning prediction of mechanical properties in metal additive manufacturing. Addit. Manuf. **91**, 104320 (2024). https://doi.org/10.1016/j.addma.2024.104320
6. Q. Liu, W. Chen, V. Yakubov, J.J. Kruzic, C.H. Wang et al., Interpretable machine learning approach for exploring process-structure-property relationships in metal additive manufacturing. Addit. Manuf. **85**, 104187 (2024). https://doi.org/10.1016/j.addma.2024.104187
7. J. Li, M. Zhou, H.-H. Wu, L. Wang, J. Zhang et al., Machine learning-assisted property prediction of solid-state electrolyte. Adv. Energy Mater. **14**(20), 2304480 (2024). https://doi.org/10.1002/aenm.202304480
8. X. Zhou, C. Xu, X. Guo, P. Apostol, A. Vlad et al., Computational and machine learning-assisted discovery and experimental validation of conjugated sulfonamide cathodes for lithium-ion batteries. Adv. Energy Mater. **15**, 2401658 (2024). https://doi.org/10.1002/aenm.202401658
9. E. Alibagheri, A. Ranjbar, M. Khazaei, T.D. Kühne, S.M. Vaez Allaei, Remarkable optoelectronic characteristics of synthesizable square-octagon haeckelite structures: machine learning materials discovery. Adv. Funct. Mater. **34**(27), 2470150 (2024). https://doi.org/10.1002/adfm.202470150
10. T. Jing, B. Xu, Y. Yang, M. Li, Y. Gao, Organogel electrode enables highly transparent and stretchable triboelectric nanogenerators of high power density for robust and reliable energy harvesting. Nano Energy **78**, 105373 (2020). https://doi.org/10.1016/j.nanoen.2020.105373
11. Y. Liu, B. Xie, Q. Hu, R. Zhao, Q. Zheng et al., Regulating the Helmholtz plane by trace polarity additive for long-life Zn ion batteries. Energy Storage Mater. **66**, 103202 (2024). https://doi.org/10.1016/j.ensm.2024.103202
12. J. Wen, B. Xu, J. Zhou, Toward flexible and wearable embroidered supercapacitors from cobalt phosphides-decorated conductive fibers. Nano-Micro Lett. **11**(1), 89 (2019). https://doi.org/10.1007/s40820-019-0321-x
13. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. Nature **521**(7553), 436–444 (2015). https://doi.org/10.1038/nature14539
14. G.E. Dahl, D. Yu, L. Deng, A. Acero, Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. Audio Speech Lang. Process. **20**(1), 30–42 (2012). https://doi.org/10.1109/TASL.2011.2134090
15. W.-N. Hsu, B. Bolte, Y.H. Tsai, K. Lakhotia, R. Salakhutdinov et al., HuBERT: self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 3451–3460 (2021). https://doi.org/10.1109/TASLP.2021.3122291
16. J. Wu, Y. Guo, C. Deng, A. Zhang, H. Qiao et al., An integrated imaging sensor for aberration-corrected 3D photography. Nature **612**(7938), 62–71 (2022). https://doi.org/10.1038/s41586-022-05306-8
17. A. Suleiman, Z. Zhang, L. Carlone, S. Karaman, V. Sze, Navion: a 2-mW fully integrated real-time visual-inertial odometry accelerator for autonomous navigation of nano drones. IEEE J. Solid State Circuits **54**(4), 1106–1119 (2019). https://doi.org/10.1109/JSSC.2018.2886342
18. J. Bai, S. Lian, Z. Liu, K. Wang, D. Liu, Smart guiding glasses for visually impaired people in indoor environment. IEEE Trans. Consum. Electron. **63**(3), 258–266 (2017). https://doi.org/10.1109/TCE.2017.014980
19. T. Starner, Project glass: an extension of the self. IEEE Pervasive Comput. **12**(2), 14–16 (2013). https://doi.org/10.1109/MPRV.2013.35
20. J. Si, P. Zhang, C. Zhao, D. Lin, L. Xu et al., A carbon-nanotube-based tensor processing unit. Nat. Electron. **7**(8), 684–693 (2024). https://doi.org/10.1038/s41928-024-01211-2
21. X. Lin, Y. Rivenson, N.T. Yardimci, M. Veli, Y. Luo et al., All-optical machine learning using diffractive deep neural networks. Science **361**(6406), 1004–1008 (2018). https://doi.org/10.1126/science.aat8084
22. X. Xu, M. Tan, B. Corcoran, J. Wu, A. Boes et al., 11 TOPS photonic convolutional accelerator for optical neural networks. Nature **589**(7840), 44–51 (2021). https://doi.org/10.1038/s41586-020-03063-0
23. G. Wetzstein, A. Ozcan, S. Gigan, S. Fan, D. Englund et al., Inference in artificial intelligence with deep optics and photonics. Nature **588**(7836), 39–47 (2020). https://doi.org/10.1038/s41586-020-2973-6
24. J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, X. Li et al., Parallel convolutional processing using an integrated photonic tensor core. Nature **589**(7840), 52–58 (2021). https://doi.org/10.1038/s41586-020-03070-1
25. F. Ashtiani, A.J. Geers, F. Aflatouni, An on-chip photonic deep neural network for image classification. Nature **606**(7914), 501–506 (2022). https://doi.org/10.1038/s41586-022-04714-0
26. F. Zangeneh-Nejad, D.L. Sounas, A. Alù, R. Fleury, Analogue computing with metamaterials. Nat. Rev. Mater. **6**(3), 207–225 (2021). https://doi.org/10.1038/s41578-020-00243-2
27. Y. Chen, M. Nazhamaiti, H. Xu, Y. Meng, T. Zhou et al., All-analog photoelectronic chip for high-speed vision tasks. Nature **623**(7985), 48–57 (2023). https://doi.org/10.1038/s41586-023-06558-8
28. G. Genty, L. Salmela, J.M. Dudley, D. Brunner, A. Kokhanovskiy et al., Machine learning and applications in ultrafast photonics. Nat. Photonics **15**(2), 91–101 (2021). https://doi.org/10.1038/s41566-020-00716-4
29. S. Molesky, Z. Lin, A.Y. Piggott, W. Jin, J. Vucković et al., Inverse design in nanophotonics. Nat. Photonics **12**(11), 659–670 (2018). https://doi.org/10.1038/s41566-018-0246-9
30. A.M. Palmieri, E. Kovlakov, F. Bianchi, D. Yudin, S. Straupe et al., Experimental neural network enhanced quantum tomography. NPJ Quantum Inf. **6**, 20 (2020). https://doi.org/10.1038/s41534-020-0248-6
31. J. Peurifoy, Y. Shen, L. Jing, Y. Yang, F. Cano-Renteria et al., Nanophotonic particle simulation and inverse design using

artificial neural networks. Sci. Adv. **4**(6), eaar4206 (2018). https://doi.org/10.1126/sciadv.aar4206

32. T.W. Hughes, M. Minkov, I.A.D. Williamson, S. Fan, Adjoint method and inverse design for nonlinear nanophotonic devices. ACS Photonics **5**(12), 4781–4787 (2018). https://doi.org/10.1021/acsphotonics.8b01522
33. A.Y. Piggott, J. Lu, K.G. Lagoudakis, J. Petykiewicz, T.M. Babinec et al., Inverse design and demonstration of a compact and broadband on-chip wavelength demultiplexer. Nat. Photonics **9**(6), 374–377 (2015). https://doi.org/10.1038/nphoton.2015.69
34. Z. Yang, W. Yue, C. Liu, Y. Tao, P.J. Tiw et al., Fully hardware memristive neuromorphic computing enabled by the integration of trainable dendritic neurons and high-density RRAM chip. Adv. Funct. Mater. **34**(44), 2405618 (2024). https://doi.org/10.1002/adfm.202405618
35. Z. Xue, T. Zhou, Z. Xu, S. Yu, Q. Dai et al., Fully forward mode training for optical neural networks. Nature **632**(8024), 280–286 (2024). https://doi.org/10.1038/s41586-024-07687-4
36. W. Wang, Y. Wang, F. Yin, H. Niu, Y.-K. Shin et al., Tailoring classical conditioning behavior in TiO$_2$ nanowires: ZnO QDs-based optoelectronic memristors for neuromorphic hardware. Nano-Micro Lett. **16**(1), 133 (2024). https://doi.org/10.1007/s40820-024-01338-z
37. M. Yao, O. Richter, G. Zhao, N. Qiao, Y. Xing et al., Spike-based dynamic computing with asynchronous sensing-computing neuromorphic chip. Nat. Commun. **15**(1), 4464 (2024). https://doi.org/10.1038/s41467-024-47811-6
38. Z. Xu, T. Zhou, M. Ma, C. Deng, Q. Dai et al., Large-scale photonic chiplet Taichi empowers 160-TOPS/W artificial general intelligence. Science **384**(6692), 202–209 (2024). https://doi.org/10.1126/science.adl1203
39. Y.J. Choi, D.G. Roe, Z. Li, Y.Y. Choi, B. Lim et al., Weight-reconfigurable neuromorphic computing systems for analog signal integration. Adv. Funct. Mater. **34**(33), 2316664 (2024). https://doi.org/10.1002/adfm.202316664
40. D. Lu, Y. Chen, Z. Lu, L. Ma, Q. Tao et al., Monolithic three-dimensional tier-by-tier integration *via* van der Waals lamination. Nature **630**(8016), 340–345 (2024). https://doi.org/10.1038/s41586-024-07406-z
41. J. Cheng, C. Huang, J. Zhang, B. Wu, W. Zhang et al., Multimodal deep learning using on-chip diffractive optics with *in situ* training capability. Nat. Commun. **15**(1), 6189 (2024). https://doi.org/10.1038/s41467-024-50677-3
42. Y. Bu, T. Xu, S. Geng, S. Fan, Q. Li et al., Ferroelectrics-electret synergetic organic artificial synapses with single-polarity driven dynamic reconfigurable modulation. Adv. Funct. Mater. **33**(20), 2213741 (2023). https://doi.org/10.1002/adfm.202213741
43. B. Dong, F. Brückerhoff-Plückelmann, L. Meyer, J. Dijkstra, I. Bente et al., Partial coherence enhances parallelized photonic computing. Nature **632**(8023), 55–62 (2024). https://doi.org/10.1038/s41586-024-07590-y
44. G. Indiveri, R. Douglas, Neuromorphic vision sensors. Science **288**(5469), 1189–1190 (2000). https://doi.org/10.1126/science.288.5469.1189
45. P. Lichtsteiner, C. Posch, T. Delbruck, A 128× 128 120 dB 15 μs latency asynchronous temporal contrast vision sensor. IEEE J. Solid State Circuits **43**(2), 566–576 (2008). https://doi.org/10.1109/JSSC.2007.914337
46. G. Gallego, T. Delbruck, G. Orchard, C. Bartolozzi, B. Taba et al., Event-based vision: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **44**(1), 154–180 (2022). https://doi.org/10.1109/tpami.2020.3008413
47. Y. van de Burgt, A. Melianas, S.T. Keene, G. Malliaras, A. Salleo, Organic electronics for neuromorphic computing. Nat. Electron. **1**(7), 386–397 (2018). https://doi.org/10.1038/s41928-018-0103-3
48. M.E. Beck, A. Shylendra, V.K. Sangwan, S. Guo, W.A. Gaviria Rojas et al., Spiking neurons from tunable Gaussian heterojunction transistors. Nat. Commun. **11**(1), 1565 (2020). https://doi.org/10.1038/s41467-020-15378-7
49. G.-T. Go, Y. Lee, D.-G. Seo, T.-W. Lee, Organic neuroelectronics: from neural interfaces to neuroprosthetics. Adv. Mater. **34**(45), e2201864 (2022). https://doi.org/10.1002/adma.202201864
50. C. Qian, Y. Choi, S. Kim, S. Kim, Y.J. Choi et al., Risk-perceptional and feedback-controlled response system based on NO$_2^-$detecting artificial sensory synapse. Adv. Funct. Mater. **32**(18), 2112490 (2022). https://doi.org/10.1002/adfm.202112490
51. K. Roy, A. Jaiswal, P. Panda, Towards spike-based machine intelligence with neuromorphic computing. Nature **575**(7784), 607–617 (2019). https://doi.org/10.1038/s41586-019-1677-2
52. A. Mehonic, A.J. Kenyon, Brain-inspired computing needs a master plan. Nature **604**(7905), 255–260 (2022). https://doi.org/10.1038/s41586-021-04362-w
53. J.H.R. Maunsell, Neuronal mechanisms of visual attention. Annu. Rev. Vis. Sci. **1**, 373–391 (2015). https://doi.org/10.1146/annurev-vision-082114-035431
54. Q. Liu, S. Gao, L. Xu, W. Yue, C. Zhang et al., Nanostructured perovskites for nonvolatile memory devices. Chem. Soc. Rev. **51**(9), 3341–3379 (2022). https://doi.org/10.1039/d1cs00886b
55. M. Chen, M. Sun, H. Bao, Y. Hu, B. Bao, Flux–charge analysis of two-memristor-based chua's circuit: dimensionality decreasing model for detecting extreme multistability. IEEE Trans. Ind. Electron. **67**(3), 2197–2206 (2019). https://doi.org/10.1109/TIE.2019.2907444
56. N. Fei, Z. Lu, Y. Gao, G. Yang, Y. Huo et al., Towards artificial general intelligence *via* a multimodal foundation model. Nat. Commun. **13**(1), 3094 (2022). https://doi.org/10.1038/s41467-022-30761-2
57. H. Zhou, J. Dong, J. Cheng, W. Dong, C. Huang et al., Photonic matrix multiplication lights up photonic accelerator and beyond. Light Sci. Appl. **11**(1), 30 (2022). https://doi.org/10.1038/s41377-022-00717-8

58. B.J. Shastri, A.N. Tait, T. Ferreira de Lima, W.H.P. Pernice, H. Bhaskaran et al., Photonics for artificial intelligence and neuromorphic computing. Nat. Photonics **15**(2), 102–114 (2021). https://doi.org/10.1038/s41566-020-00754-y
59. K.Y. Yang, C. Shirpurkar, A.D. White, J. Zang, L. Chang et al., Multi-dimensional data transmission using inverse-designed silicon photonics and microcombs. Nat. Commun. **13**(1), 7862 (2022). https://doi.org/10.1038/s41467-022-35446-4
60. Z. Wang, S. Joshi, S. Savel'ev, W. Song, R. Midya et al., Fully memristive neural networks for pattern classification with unsupervised learning. Nat. Electron. **1**(2), 137–145 (2018). https://doi.org/10.1038/s41928-018-0023-2
61. E. Peterson, A. Lavin, Physical computing for materials acceleration platforms. Matter **5**(11), 3586–3596 (2022). https://doi.org/10.1016/j.matt.2022.09.022
62. K. Hippalgaonkar, Q. Li, X. Wang, J.W. Fisher III., J. Kirkpatrick et al., Knowledge-integrated machine learning for materials: lessons from gameplaying and robotics. Nat. Rev. Mater. **8**(4), 241–260 (2023). https://doi.org/10.1038/s41578-022-00513-1
63. H. Huang, O. Zheng, D. Wang, J. Yin, Z. Wang et al., ChatGPT for shaping the future of dentistry: the potential of multi-modal large language model. Int. J. Oral Sci. **15**(1), 29 (2023). https://doi.org/10.1038/s41368-023-00239-y
64. B. Meskó, The impact of multimodal large language models on health care's future. J. Med. Internet Res. **25**, e52865 (2023). https://doi.org/10.2196/52865
65. M. Moor, O. Banerjee, Z.S.H. Abad, H.M. Krumholz, J. Leskovec et al., Foundation models for generalist medical artificial intelligence. Nature **616**(7956), 259–265 (2023). https://doi.org/10.1038/s41586-023-05881-4
66. X. Wang, G. Chen, G. Qian, P. Gao, X.-Y. Wei et al., Large-scale multi-modal pre-trained models: a comprehensive survey. Mach. Intell. Res. **20**(4), 447–482 (2023). https://doi.org/10.1007/s11633-022-1410-8
67. M.A. Zidan, J.P. Strachan, W.D. Lu, The future of electronics based on memristive systems. Nat. Electron. **1**(1), 22–29 (2018). https://doi.org/10.1038/s41928-017-0006-8
68. B. Yan, Y. Yang, R. Huang, Memristive dynamics enabled neuromorphic computing systems. Sci. China Inf. Sci. **66**(10), 200401 (2023). https://doi.org/10.1007/s11432-023-3739-0
69. A. Sebastian, M. Le Gallo, R. Khaddam-Aljameh, E. Eleftheriou, Memory devices and applications for in-memory computing. Nat. Nanotechnol. **15**(7), 529–544 (2020). https://doi.org/10.1038/s41565-020-0655-z
70. S. Ambrogio, P. Narayanan, H. Tsai, R.M. Shelby, I. Boybat et al., Equivalent-accuracy accelerated neural-network training using analogue memory. Nature **558**(7708), 60–67 (2018). https://doi.org/10.1038/s41586-018-0180-5
71. R. Khaddam-Aljameh, M. Stanisavljevic, J. Fornt Mas, G. Karunaratne, M. Brandli et al., HERMES-core: a 1.59-TOPS/mm^2 PCM on 14-nm CMOS in-memory compute core using 300-ps/LSB linearized CCO-based ADCs. IEEE J. Solid State Circuits **57**(4), 1027–1038 (2022). https://doi.org/10.1109/jssc.2022.3140414
72. P. Yao, H. Wu, B. Gao, J. Tang, Q. Zhang et al., Fully hardware-implemented memristor convolutional neural network. Nature **577**(7792), 641–646 (2020). https://doi.org/10.1038/s41586-020-1942-4
73. W. Wan, R. Kubendran, C. Schaefer, S.B. Eryilmaz, W. Zhang et al., A compute-in-memory chip based on resistive random-access memory. Nature **608**(7923), 504–512 (2022). https://doi.org/10.1038/s41586-022-04992-8
74. S. Jain, H. Tsai, C.-T. Chen, R. Muralidhar, I. Boybat et al., A heterogeneous and programmable compute-In-memory accelerator architecture for analog-AI using dense 2-D mesh. IEEE Trans. VLSI Syst. **31**(1), 114–127 (2023). https://doi.org/10.1109/tvlsi.2022.3221390
75. O. Krestinskaya, K.N. Salama, A.P. James, Learning in memristive neural network architectures using analog backpropagation circuits. IEEE Trans. Circuits Syst. I Regul. Pap. **66**(2), 719–732 (2019). https://doi.org/10.1109/TCSI.2018.2866510
76. M. Giordano, G. Cristiano, K. Ishibashi, S. Ambrogio, H. Tsai et al., Analog-to-digital conversion with reconfigurable function mapping for neural networks activation function acceleration. IEEE J. Emerg. Sel. Top. Circuits Syst. **9**(2), 367–376 (2019). https://doi.org/10.1109/JETCAS.2019.2911537
77. J. Hochstetter, R. Zhu, A. Loeffler, A. Diaz-Alvarez, T. Nakayama et al., Avalanches and edge-of-chaos learning in neuromorphic nanowire networks. Nat. Commun. **12**(1), 4008 (2021). https://doi.org/10.1038/s41467-021-24260-z
78. W. Schultz, A. Dickinson, Neuronal coding of prediction errors. Annu. Rev. Neurosci. **23**, 473–500 (2000). https://doi.org/10.1146/annurev.neuro.23.1.473
79. R.A. Poldrack, J. Clark, E.J. Paré-Blagoev, D. Shohamy, J. Creso Moyano et al., Interactive memory systems in the human brain. Nature **414**(6863), 546–550 (2001). https://doi.org/10.1038/35107080
80. Z. Wang, C. Li, W. Song, M. Rao, D. Belkin et al., Reinforcement learning with analogue memristor arrays. Nat. Electron. **2**(3), 115–124 (2019). https://doi.org/10.1038/s41928-019-0221-6
81. J.H. Baek, K.J. Kwak, S.J. Kim, J. Kim, J.Y. Kim et al., Two-terminal lithium-mediated artificial synapses with enhanced weight modulation for feasible hardware neural networks. Nano-Micro Lett. **15**(1), 69 (2023). https://doi.org/10.1007/s40820-023-01035-3
82. K. He, Y. Liu, J. Yu, X. Guo, M. Wang et al., Artificial neural pathway based on a memristor synapse for optically mediated motion learning. ACS Nano **16**(6), 9691–9700 (2022). https://doi.org/10.1021/acsnano.2c03100
83. T. Zhou, X. Lin, J. Wu, Y. Chen, H. Xie et al., Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit. Nat. Photonics **15**(5), 367–373 (2021). https://doi.org/10.1038/s41566-021-00796-w
84. T. Fu, Y. Zang, Y. Huang, Z. Du, H. Huang et al., Photonic machine learning with on-chip diffractive optics. Nat. Commun. **14**(1), 70 (2023). https://doi.org/10.1038/s41467-022-35772-7

https://doi.org/10.1007/s40820-025-01850-w

85. E. Goi, X. Chen, Q. Zhang, B.P. Cumming, S. Schoenhardt et al., Nanoprinted high-neuron-density optical linear perceptrons performing near-infrared inference on a CMOS chip. Light Sci. Appl. **10**(1), 40 (2021). https://doi.org/10.1038/s41377-021-00483-z
86. H. Zhang, M. Gu, X.D. Jiang, J. Thompson, H. Cai et al., An optical neural chip for implementing complex-valued neural network. Nat. Commun. **12**, 457 (2021). https://doi.org/10.1038/s41467-020-20719-7
87. T. Wang, S.-Y. Ma, L.G. Wright, T. Onodera, B.C. Richard et al., An optical neural network using less than 1 photon per multiplication. Nat. Commun. **13**(1), 123 (2022). https://doi.org/10.1038/s41467-021-27774-8
88. Z. Wang, G. Hu, X. Wang, X. Ding, K. Zhang et al., Single-layer spatial analog meta-processor for imaging processing. Nat. Commun. **13**(1), 2188 (2022). https://doi.org/10.1038/s41467-022-29732-4
89. J. Li, D. Mengu, N.T. Yardimci, Y. Luo, X. Li et al., Spectrally encoded single-pixel machine vision using diffractive networks. Sci. Adv. **7**(13), eabd7690 (2021). https://doi.org/10.1126/sciadv.abd7690
90. M.S.S. Rahman, J. Li, D. Mengu, Y. Rivenson, A. Ozcan, Ensemble learning of diffractive optical networks. Light Sci. Appl. **10**(1), 14 (2021). https://doi.org/10.1038/s41377-020-00446-w
91. J. Feldmann, N. Youngblood, C.D. Wright, H. Bhaskaran, W.P. Pernice, All-optical spiking neurosynaptic networks with self-learning capabilities. Nature **569**(7755), 208–214 (2019). https://doi.org/10.1038/s41586-019-1157-8
92. W. Shi, Z. Huang, H. Huang, C. Hu, M. Chen et al., LOEN: lensless opto-electronic neural network empowered machine vision. Light Sci. Appl. **11**(1), 121 (2022). https://doi.org/10.1038/s41377-022-00809-5
93. J. Chang, V. Sitzmann, X. Dun, W. Heidrich, G. Wetzstein, Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification. Sci. Rep. **8**(1), 12324 (2018). https://doi.org/10.1038/s41598-018-30619-y
94. J. Bueno, S. Maktoobi, L. Froehly, I. Fischer, M. Jacquot et al., Reinforcement learning in a large-scale photonic recurrent neural network. Optica **5**(6), 756 (2018). https://doi.org/10.1364/optica.5.000756
95. A. Silva, F. Monticone, G. Castaldi, V. Galdi, A. Alù et al., Performing mathematical operations with metamaterials. Science **343**(6167), 160–163 (2014). https://doi.org/10.1126/science.1242818
96. Y. Shen, N.C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones et al., Deep learning with coherent nanophotonic circuits. Nat. Photonics **11**(7), 441–446 (2017). https://doi.org/10.1038/nphoton.2017.93
97. X. Yuan, Y. Wang, Z. Xu, T. Zhou, L. Fang, Training large-scale optoelectronic neural networks with dual-neuron optical-artificial learning. Nat. Commun. **14**(1), 7110 (2023). https://doi.org/10.1038/s41467-023-42984-y
98. T. Zhou, W. Wu, J. Zhang, S. Yu, L. Fang, Ultrafast dynamic machine vision with spatiotemporal photonic computing. Sci. Adv. **9**(23), eadg4391 (2023). https://doi.org/10.1126/sciadv.adg4391
99. Z. Xu, X. Yuan, T. Zhou, L. Fang, A multichannel optical computing architecture for advanced machine vision. Light Sci. Appl. **11**(1), 255 (2022). https://doi.org/10.1038/s41377-022-00945-y
100. S. Neyens, O.K. Zietz, T.F. Watson, F. Luthi, A. Nethwewala et al., Probing single electrons across 300-mm spin qubit wafers. Nature **629**(8010), 80–85 (2024). https://doi.org/10.1038/s41586-024-07275-6
101. D. Wecker, B. Bauer, B.K. Clark, M.B. Hastings, M. Troyer, Gate-count estimates for performing quantum chemistry on small quantum computers. Phys. Rev. A **90**(2), 022305 (2014). https://doi.org/10.1103/physreva.90.022305
102. M. Brauns, S.V. Amitonov, P.-C. Spruijtenburg, F.A. Zwanenburg, Palladium gates for reproducible quantum dots in silicon. Sci. Rep. **8**(1), 5690 (2018). https://doi.org/10.1038/s41598-018-24004-y
103. J.P. Dodson, N. Holman, B. Thorgrimsson, S.F. Neyens, E.R. MacQuarrie et al., Fabrication process and failure analysis for robust quantum dots in silicon. Nanotechnology **31**(50), 505001 (2020). https://doi.org/10.1088/1361-6528/abb559
104. M.M. Shulaker, G. Hills, R.S. Park, R.T. Howe, K. Saraswat et al., Three-dimensional integration of nanotechnologies for computing and data storage on a single chip. Nature **547**(7661), 74–78 (2017). https://doi.org/10.1038/nature22994
105. K. Zhu, S. Pazos, F. Aguirre, Y. Shen, Y. Yuan et al., Hybrid 2D-CMOS microchips for memristive applications. Nature **618**(7963), 57–62 (2023). https://doi.org/10.1038/s41586-023-05973-1
106. S. Kim, J. Seo, J. Choi, H. Yoo, Vertically integrated electronics: new opportunities from emerging materials and devices. Nano-Micro Lett. **14**(1), 201 (2022). https://doi.org/10.1007/s40820-022-00942-1
107. J. Jiang, K. Parto, W. Cao, K. Banerjee, Ultimate monolithic-3D integration with 2D materials: rationale, prospects, and challenges. IEEE J. Electron Devices Soc. **7**, 878–887 (2019). https://doi.org/10.1109/JEDS.2019.2925150
108. L. Tong, J. Wan, K. Xiao, J. Liu, J. Ma et al., Heterogeneous complementary field-effect transistors based on silicon and molybdenum disulfide. Nat. Electron. **6**(1), 37–44 (2022). https://doi.org/10.1038/s41928-022-00881-0
109. W. Meng, F. Xu, Z. Yu, T. Tao, L. Shao et al., Three-dimensional monolithic micro-LED display driven by atomically thin transistor matrix. Nat. Nanotechnol. **16**(11), 1231–1236 (2021). https://doi.org/10.1038/s41565-021-00966-5
110. J.-H. Kang, H. Shin, K.S. Kim, M.-K. Song, D. Lee et al., Monolithic 3D integration of 2D materials-based electronics towards ultimate edge computing solutions. Nat. Mater. **22**(12), 1470–1477 (2023). https://doi.org/10.1038/s41563-023-01704-z
111. D. Jayachandran, R. Pendurthi, M.U.K. Sadaf, N.U. Sakib, A. Pannone et al., Three-dimensional integration of two-dimensional field-effect transistors. Nature

625(7994), 276–281 (2024). https://doi.org/10.1038/s41586-023-06860-5

112. J. Tang, Q. Wang, Z. Wei, C. Shen, X. Lu et al., Vertical integration of 2D building blocks for all-2D electronics. Adv. Electron. Mater. **6**(12), 2000550 (2020). https://doi.org/10.1002/aelm.202000550
113. Y. Liu, Y. Huang, X. Duan, Van der waals integration before and beyond two-dimensional materials. Nature **567**(7748), 323–333 (2019). https://doi.org/10.1038/s41586-019-1013-x
114. H. Zhang, H. Zeng, A. Priimagi, O. Ikkala, Viewpoint: Pavlovian materials-functional biomimetics inspired by classical conditioning. Adv. Mater. **32**(20), e1906619 (2020). https://doi.org/10.1002/adma.201906619
115. M.I. Jordan, T.M. Mitchell, Machine learning: trends, perspectives, and prospects. Science **349**(6245), 255–260 (2015). https://doi.org/10.1126/science.aaa8415
116. D. Woods, T.J. Naughton, Photonic neural networks. Nat. Phys. **8**(4), 257–259 (2012). https://doi.org/10.1038/nphys2283
117. D.A.B. Miller, Device requirements for optical interconnects to silicon chips. Proc. IEEE **97**(7), 1166–1185 (2009). https://doi.org/10.1109/JPROC.2009.2014298
118. D.A.B. Miller, Attojoule optoelectronics for low-energy information processing and communications. J. Light. Technol. **35**(3), 346–396 (2017). https://doi.org/10.1109/JLT.2017.2647779
119. D.A.B. Miller, Waves, modes, communications, and optics: a tutorial. Adv. Opt. Photon. **11**(3), 679 (2019). https://doi.org/10.1364/aop.11.000679
120. J. Wang, J.-Y. Yang, I.M. Fazal, N. Ahmed, Y. Yan et al., Terabit free-space data transmission employing orbital angular momentum multiplexing. Nat. Photonics **6**(7), 488–496 (2012). https://doi.org/10.1038/nphoton.2012.138
121. D.J. Richardson, J.M. Fini, L.E. Nelson, Space-division multiplexing in optical fibres. Nat. Photonics **7**(5), 354–362 (2013). https://doi.org/10.1038/nphoton.2013.94
122. N. Bozinovic, Y. Yue, Y. Ren, M. Tur, P. Kristensen et al., Terabit-scale orbital angular momentum mode division multiplexing in fibers. Science **340**(6140), 1545–1548 (2013). https://doi.org/10.1126/science.1237861
123. R. Ryf, S. Randel, A.H. Gnauck, C. Bolle, A. Sierra et al., Mode-division multiplexing over 96 km of few-mode fiber using coherent 6 × 6 MIMO processing. J. Lightwave Technol. **30**(4), 521–531 (2012). https://doi.org/10.1109/jlt.2011.2174336
124. R.G.H. van Uden, R.A. Correa, E.A. Lopez, F.M. Huijskens, C. Xia et al., Ultra-high-density spatial division multiplexing with a few-mode multicore fibre. Nat. Photonics **8**(11), 865–870 (2014). https://doi.org/10.1038/nphoton.2014.243
125. J.M. Kahn, D.A.B. Miller, Communications expands its space. Nat. Photonics **11**(1), 5–8 (2017). https://doi.org/10.1038/nphoton.2016.256
126. G. Rademacher, B.J. Puttnam, R.S. Luís, T.A. Eriksson, N.K. Fontaine et al., Peta-bit-per-second optical communications system using a standard cladding diameter 15-mode fiber. Nat. Commun. **12**(1), 4238 (2021). https://doi.org/10.1038/s41467-021-24409-w
127. L.H. Gabrielli, D. Liu, S.G. Johnson, M. Lipson, On-chip transformation optics for multimode waveguide bends. Nat. Commun. **3**, 1217 (2012). https://doi.org/10.1038/ncomms2232
128. L.-W. Luo, N. Ophir, C.P. Chen, L.H. Gabrielli, C.B. Poitras et al., WDM-compatible mode-division multiplexing on a silicon chip. Nat. Commun. **5**, 3069 (2014). https://doi.org/10.1038/ncomms4069
129. S.A. Miller, Y.-C. Chang, C.T. Phare, M.C. Shin, M. Zadka et al., Large-scale optical phased array using a low-power multi-pass silicon photonic platform. Optica **7**(1), 3 (2020). https://doi.org/10.1364/optica.7.000003
130. L.F. Frellsen, Y. Ding, O. Sigmund, L.H. Frandsen, Topology optimized mode multiplexing in silicon-on-insulator photonic wire waveguides. Opt. Express **24**(15), 16866–16873 (2016). https://doi.org/10.1364/OE.24.016866
131. W. Chang, L. Lu, X. Ren, D. Li, Z. Pan et al., Ultra-compact mode (de multiplexer based on subwavelength asymmetric Y-junction. Opt. Express **26**(7), 8162–8170 (2018). https://doi.org/10.1364/OE.26.008162
132. Y. Tong, W. Zhou, X. Wu, H.K. Tsang, Efficient mode multiplexer for few-mode fibers using integrated silicon-on-insulator waveguide grating coupler. IEEE J. Quantum Electron. **56**(1), 8400107 (2020). https://doi.org/10.1109/JQE.2019.2950126
133. H. Hu, F. Da Ros, M. Pu, F. Ye, K. Ingerslev et al., Single-source chip-based frequency comb enabling extreme parallel data transmission. Nat. Photonics **12**(8), 469–473 (2018). https://doi.org/10.1038/s41566-018-0205-5
134. Y. Han, G. Huang, S. Song, L. Yang, H. Wang et al., Dynamic neural networks: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **44**(11), 7436–7456 (2021). https://doi.org/10.1109/TPAMI.2021.3117837
135. J.A. Ang, D.J. Mountain, New horizons for high-performance computing. Computer **55**(12), 156–162 (2022). https://doi.org/10.1109/MC.2022.3200859
136. J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre et al., Mastering Atari, Go, chess and shogi by planning with a learned model. Nature **588**(7839), 604–609 (2020). https://doi.org/10.1038/s41586-020-03051-4
137. A. Sludds, S. Bandyopadhyay, Z. Chen, Z. Zhong, J. Cochrane et al., Delocalized photonic deep learning on the Internet's edge. Science **378**(6617), 270–276 (2022). https://doi.org/10.1126/science.abq8271
138. L.G. Wright, T. Onodera, M.M. Stein, T. Wang, D.T. Schachter et al., Deep physical neural networks trained with backpropagation. Nature **601**(7894), 549–555 (2022). https://doi.org/10.1038/s41586-021-04223-6
139. W. Ma, Z. Liu, Z.A. Kudyshev, A. Boltasseva, W. Cai et al., Deep learning for the design of photonic structures. Nat.

Photonics **15**(2), 77–90 (2021). https://doi.org/10.1038/s41566-020-0685-y

140. N. Mohammadi Estakhri, B. Edwards, N. Engheta, Inverse-designed metastructures that solve equations. Science **363**(6433), 1333–1338 (2019). https://doi.org/10.1126/science.aaw2498
141. A. McNamara, A. Treuille, Z. Popović, J. Stam, Fluid control using the adjoint method. ACM Trans. Graph. **23**(3), 449–456 (2004). https://doi.org/10.1145/1015706.1015744
142. X. Wang, P. Xie, B. Chen, X. Zhang, Chip-based high-dimensional optical neural network. Nano-Micro Lett. **14**(1), 221 (2022). https://doi.org/10.1007/s40820-022-00957-8
143. J. Moran, R. Desimone, Selective attention gates visual processing in the extrastriate cortex. Science **229**(4715), 782–784 (1985). https://doi.org/10.1126/science.4023713
144. T. Moore, M. Zirnsak, Neural mechanisms of selective visual attention. Annu. Rev. Psychol. **68**, 47–72 (2017). https://doi.org/10.1146/annurev-psych-122414-033400
145. X. Duan, Z. Cao, K. Gao, W. Yan, S. Sun et al., Memristor-based neuromorphic chips. Adv. Mater. **36**(14), 2310704 (2024). https://doi.org/10.1002/adma.202310704
146. I. Oguz, M. Yildirim, J.L. Hsieh, N.U. Dinc, C. Moser et al., Resource-efficient photonic networks for next-generation AI computing. Light Sci. Appl. **14**(1), 34 (2025). https://doi.org/10.1038/s41377-024-01717-6
147. D. Wang, Y. Nie, G. Hu, H.K. Tsang, C. Huang, Ultrafast silicon photonic reservoir computing engine delivering over 200 TOPS. Nat. Commun. **15**(1), 10841 (2024). https://doi.org/10.1038/s41467-024-55172-3
148. T. Günkel, J. Alcalà, A. Fernández, A. Barrera, L. Balcells et al., Field-induced phase transitions in cuprate superconductors for cryogenic in-memory computing. Small **21**(14), e2411908 (2025). https://doi.org/10.1002/smll.202411908
149. B. Bai, Q. Yang, H. Shu, L. Chang, F. Yang et al., Microcomb-based integrated photonic processing unit. Nat. Commun. **14**, 66 (2023). https://doi.org/10.1038/s41467-022-35506-9
150. S. Wan, K. Qu, Y. Shi, Z. Li, Z. Wang et al., Multidimensional encryption by chip-integrated metasurfaces. ACS Nano **18**(28), 18693–18700 (2024). https://doi.org/10.1021/acsnano.4c05724
151. S.-A. Guo, Y.-K. Wu, J. Ye, L. Zhang, W.-Q. Lian et al., A site-resolved two-dimensional quantum simulator with hundreds of trapped ions. Nature **630**(8017), 613–618 (2024). https://doi.org/10.1038/s41586-024-07459-0
152. C. Wang, Z. Chen, C.L.J. Chan, Z.-A. Wan, W. Ye et al., Biomimetic olfactory chips based on large-scale monolithically integrated nanotube sensor arrays. Nat. Electron. **7**(2), 157–167 (2024). https://doi.org/10.1038/s41928-023-01107-7
153. D. Kumar, H. Li, D.D. Kumbhar, M.K. Rajbhar, U.K. Das et al., Highly efficient back-end-of-line compatible flexible Si-based optical memristive crossbar array for edge neuromorphic physiological signal processing and bionic machine vision. Nano-Micro Lett. **16**(1), 238 (2024). https://doi.org/10.1007/s40820-024-01456-8

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