Article

Universal photonic artificial intelligence acceleration

https://doi.org/10.1038/s41586-025-08854-x

Received: 18 November 2024

Accepted: 3 March 2025

Published online: 9 April 2025

Check for updates

Sufi R. Ahmed[1], Reza Baghdadi[1], Mikhail Bernadskiy[1], Nate Bowman[1], Ryan Braid[1], Jim Carr[1], Chen Chen[1], Pietro Ciccarella[1], Matthew Cole[1], John Cooke[1], Kishor Desai[1], Carlos Dorta[1], Jonathan Elmhurst[1], Bryce Gardiner[1], Elliot Greenwald[1], Shashank Gupta[1], Parry Husbands[1], Brian Jones[1], Anthony Kopa[1], Ho John Lee[1], Arulselvan Madhavan[1], Adam Mendrela[1], Nicholas Moore[1], Lakshmi Nair[1], Aditya Om[1], Subie Patel[1], Rutayan Patro[1], Rob Pellowski[1], Esha Radhakrishnani[1], Sandeep Sane[1], Nicholas Sarkis[1], Joe Stadolnik[1], Mykhailo Tymchenko[1], Gongyu Wang[1], Kurt Winikka[1], Alexandra Wleklinski[1], Josh Zelman[1], Richard Ho[2], Ritesh Jain[1], Ayon Basumallik[1]✉, Darius Bunandar[1]✉ & Nicholas C. Harris[1]✉

Over the past decade, photonics research has explored accelerated tensor operations, foundational to artificial intelligence (AI) and deep learning[1–4], as a path towards enhanced energy efficiency and performance[5–14]. The field is centrally motivated by finding alternative technologies to extend computational progress in a post-Moore's law and Dennard scaling era[15–19]. Despite these advances, no photonic chip has achieved the precision necessary for practical AI applications, and demonstrations have been limited to simplified benchmark tasks. Here we introduce a photonic AI processor that executes advanced AI models, including ResNet[3] and BERT[20,21], along with the Atari deep reinforcement learning algorithm originally demonstrated by DeepMind[22]. This processor achieves near-electronic precision for many workloads, marking a notable entry for photonic computing into competition with established electronic AI accelerators[23] and an essential step towards developing post-transistor computing technologies.

With the exponential growth in AI model complexity driven by large language models, reinforcement learning and convolutional neural networks, electronic computers are now fundamentally bound by Moore's law and Dennard scaling[15]. Photonics offers an alternative by exploiting the high bandwidth, low latency and energy efficiency of light-based computation[5,6]. Recent developments, including photonic accelerators based on interleaved time-wavelength modulation and photoelectric multiplication, underline the progress towards photonic processors for AI[7,24–28]. Although these systems have demonstrated essential linear algebra operations, such as matrix multiplication, challenges in achieving precision, scalability, system integration and compatibility with advanced AI architectures remain[29,30]. Addressing these challenges is crucial for positioning photonic processors as a viable alternative to electronic accelerators, with the potential for substantial gains in computational speed and energy efficiency[5,8–10,24,29,31].

Here we report the first photonic processor, to our knowledge, capable of executing state-of-the-art neural networks, including transformers, convolutional networks classification and segmentation, and reinforcement learning algorithms. Critically, this photonic processor achieves accuracies approaching those of 32-bit digital floating-point systems on advanced tasks, validating its computational integrity even without requiring advanced techniques such as fine-tuning and quantization-aware training[32–34]. The design integrates six chips in a single package, using high-speed interconnects between vertically aligned photonic tensor cores (PTCs) and control dies, thus achieving high efficiency and scalability for AI computation[7]. This work focuses on accurately executing state-of-the-art neural networks—even with several hardware non-idealities discussed in the Supplementary information, the photonic processor generates 65.5 trillion adaptive block floating-point[35] (ABFP) 16-bit operations per second at 78 W of electrical power and 1.6 W of optical power. This work represents the highest level of integration achieved in photonic processing.

The photonic processor, depicted in Fig. 1a, integrates four 128 × 128 PTCs fabricated using GlobalFoundries' 90-nm photonics process[36], as illustrated in Fig. 1b. Each PTC occupies 14.00 × 24.96 mm and contains all photonic components and analogue mixed-signal circuits required to operate them, except for high-speed analogue-to-digital converters (ADCs). The processor also incorporates two digital control interface (DCI) chips, manufactured with GlobalFoundries' 12-nm process, each measuring 31.4 × 25.0 mm. High-speed ADCs are included in the DCI. This innovative package assembles six chips: two full-reticle, 25-billion-transistor DCI dies on an organic interposer (54 × 56 mm) with four PTC dies underneath (see Supplementary information section III).

Figure 1c shows the DCI die floor plan. During a tensor operation, vectors flow from the input pipeline through a PTC to the output pipeline. The PTC floor plan is mirrored about the *y* axis to match two PTCs to one DCI. Each DCI includes 64 reduced instruction set computer (RISC) cores, implemented using quad-core 32-bit RISC-V SiFive E76-MC complexes and a 268-MB unified buffer for storing activations, weight parameters and other data. The PTC floor plan in Fig. 1d includes 128 10-bit photonic vector units and 128 × 128 7-bit weight unit cells.

[1]Lightmatter, Mountain View, CA, USA. [2]OpenAI, San Francisco, CA, USA. ✉e-mail: ayon@lightmatter.co; darius@lightmatter.co; nick@lightmatter.co

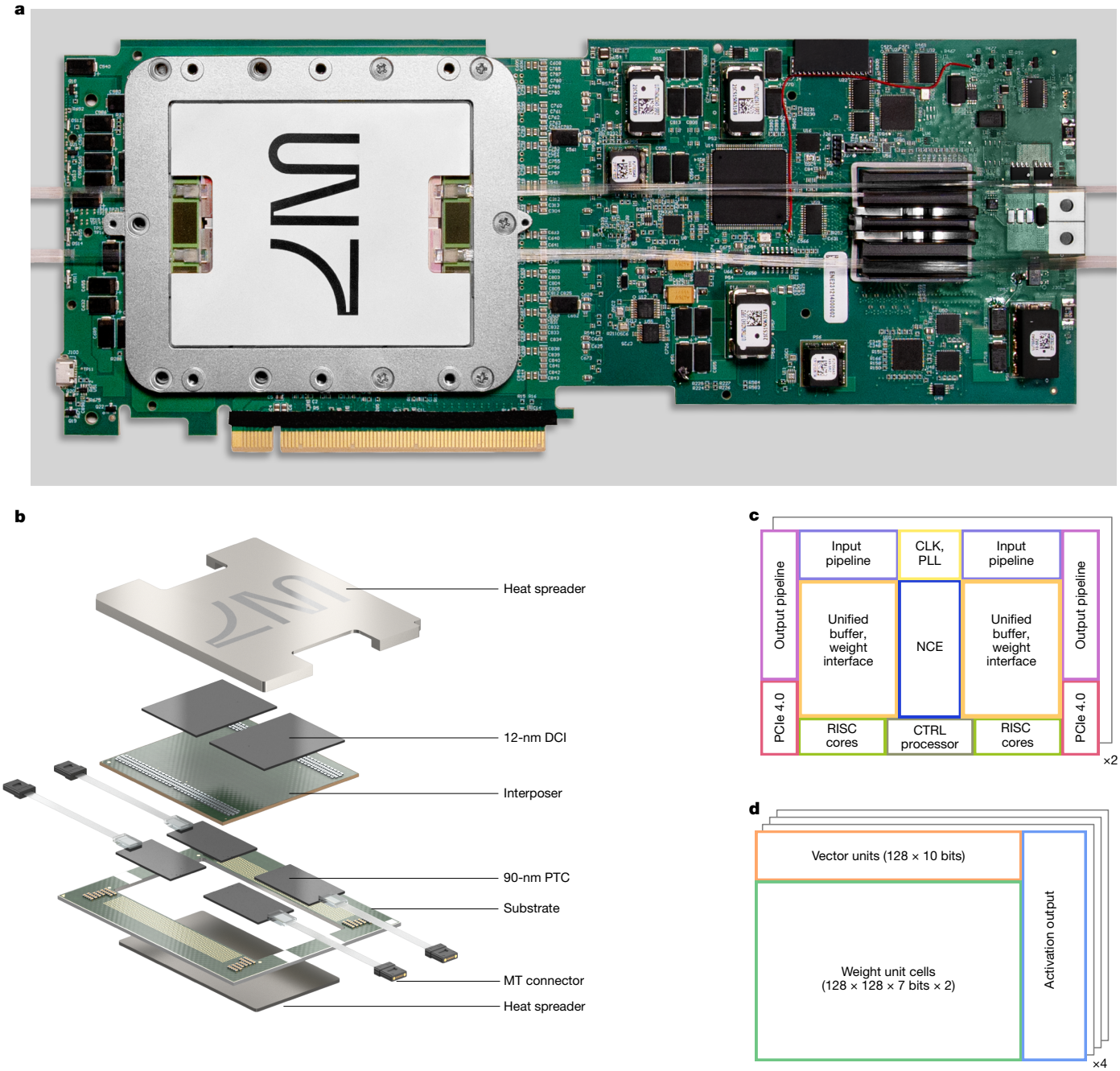

Fig. 1 | Quad-core photonic processor. a, Photograph of processor on a 16-lane PCIe card with four PTCs, each with a 12-fibre array attached. **b**, Packaging architecture rendering showing the integrated heat spreader, two 12-nm control chiplets measuring approximately 780 mm^2 on an organic substrate and four 349-mm^2 photonic chiplets on the back side, a bottom substrate for socket contact and a silicon heat spreader spanning the four photonic cores. **c**, Floor plan of the 12-nm digital chiplets with the input–output interfaces, memory, RISC cores and data converters. **d**, Floor plan of the photonic chiplets with the weight transfer interface, vector units and weight units.

Execution model and digital architecture

The block diagram of the photonic processor, shown in Fig. 2a, illustrates the integration of a DCI with PTCs. In this architecture, the RISC control code and device instruction set architecture (ISA) are transmitted from the host system by means of a PCIe Gen4 x16 bus to shared memory accessible by the RISC cores and a dedicated hardware instruction sequencer (CTRL in Fig. 2a). The CTRL module runs NuttX, an open-source real-time operating system optimized for embedded systems. This configuration enables efficient execution of calibration routines and other control operations. This sequencer directs commands across various subsystems through a real-time command network. Each PTC executes a 128 × 128 matrix-vector product (MVP), drawing weights from the unified buffer and transferring them to the weight interface of the tensor core at 1,024 GB s^{-1}.

The system uses double-buffered weights to sustain uninterrupted data flow, enabling continuous streaming of new weights during MVP computation. Integrated input and output pipelines facilitate processing and output of 128-element vectors per cycle (256 GB s^{-1} each). Adjustments of active weights, with a settling time of approximately 10 ns, align with digital pipeline delays to maximize efficiency. The output pipeline also supports read–modify–write operations, accumulating tensor core outputs into the unified buffer. This unified

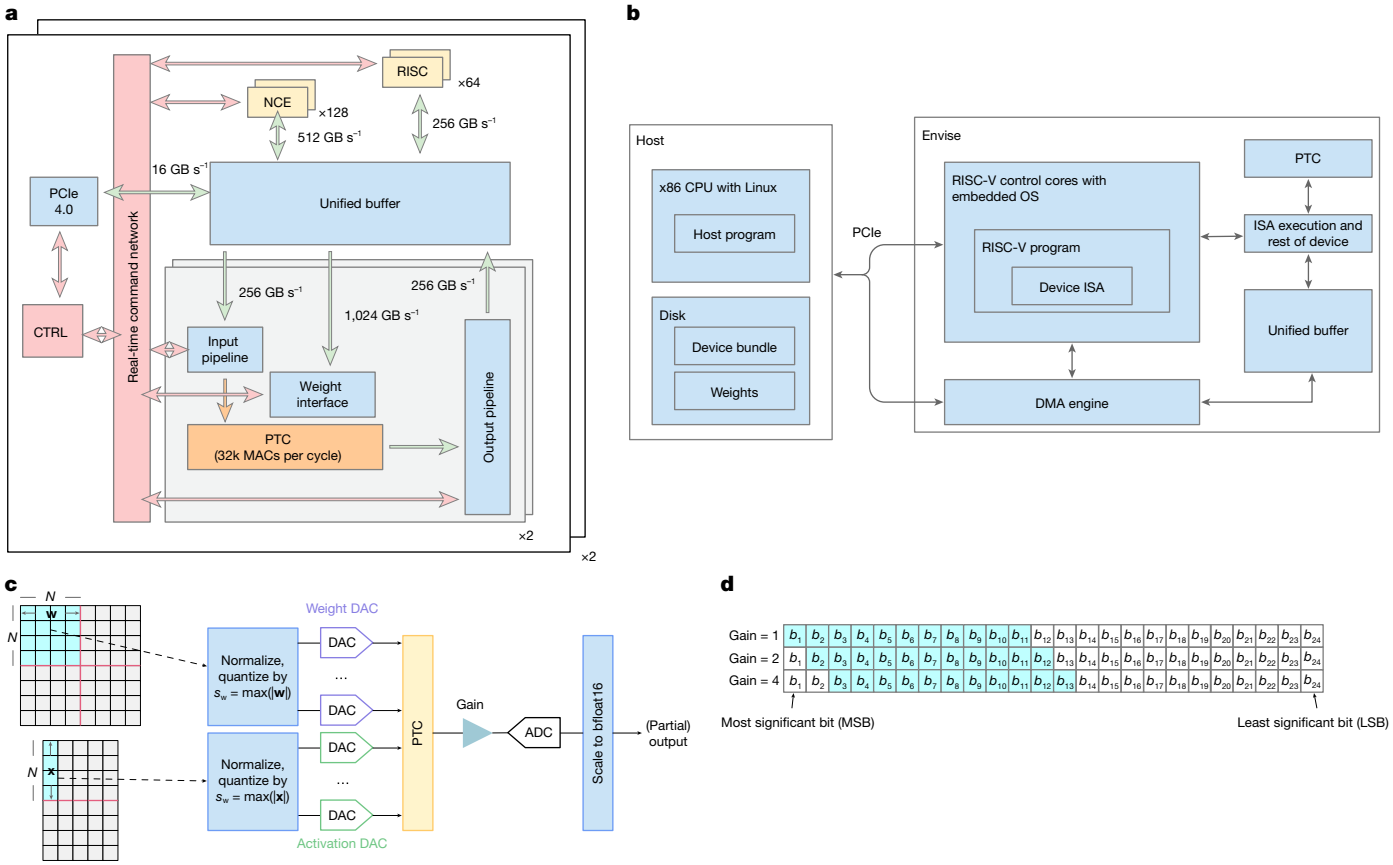

Fig. 2 | Processor functionality and execution model. a, Photonic processor block diagram including a PCIe 4.0 interface, unified buffer, RISC processors, programmable activation functions and PTCs. **b**, Software execution model. A host processor communicates with the photonic processor over PCIe, issues instructions and transfers data. **c**, Number representation in the photonic processor. **d**, Increasing gain allows lower bits to be read by the ADC, shown as highlighted, whereas upper bits may saturate. The example shows a total of 24 bits output (from 7-bit weights, 10-bit activations and accumulation of 128 elements per vector element). MACs, multiply–accumulate operations.

buffer spans the control die and manages several concurrent internal transfers, reaching speeds of 2,048 GB s^{-1}.

The neural compute engine (NCE) serves as a 128-way single instruction, multiple data (SIMD) load/store streaming processor, executing bfloat16 operations across 16 matrix registers (each holding 128 vectors of 128 elements) and 16 vector registers (each holding a 128-element vector). Operating concurrently on several execution streams at 256 GB s^{-1}, the NCE makes use of matrix registers as intermediate FIFOs for temporary data storage. Reductions occur in bfloat24, whereas nonlinear functions execute through a flexible, piecewise-linear lookup table. A further 64 RISC-V cores support operations not natively handled by the NCE, primarily assisting with PTC calibration.

The software execution model is shown in Fig. 2b. The input to the compiler consists of a machine learning network defined in PyTorch or TensorFlow 2.0, which the compiler converts into three distinct components: (1) a host x86 binary that coordinates network execution across several devices and handles operators designated to the host; (2) a RISC-V binary that interfaces with the embedded operating system, facilitating operations such as ISA execution and off-device direct memory access (DMA); and (3) ISA, which the instruction sequencer processes. Python bindings encapsulate network execution on top of the host x86 code, offering a PyTorch-like interface.

At load time, trained weights transfer into the unified buffer, whereas the RISC-V program and ISA are loaded into control-plane memory. Most ISA instructions follow a SIMD-like streaming structure capable of processing a dynamic number of vectors. For parallel execution of longer instructions, the compiler assigns dependency tags to the ISA, allowing the sequencer to execute instructions out of order. Alongside essential digital operations required for non-GEMM and non-convolutional network functions, the NCE generates PTC weights on the fly for operands whose values remain undetermined until runtime. The compiler autonomously manages partitioning and pipelining across several DCIs, supporting neural networks that exceed the capacity of the onboard SRAM.

Weights and activations in the photonic processor are represented in an ABFP format, as shown in Fig. 2c. Weights and activations in the NCE are represented in bfloat16 and are converted to ABFP by normalizing each vector of length 128 by the absolute value of the maximum vector entry. Each 128-long vector, therefore, shares a single bfloat16 scale: s_w and s_x for the weights and activations, respectively. The normalized vector is then quantized to the bit precision of the unit cells: 7 bits for the weights and 10 bits for the activations. Matrix multiplication occurs in the PTC, gain is applied to the output by the transimpedance amplifier (TIA) and the ADC reads 11 bits out. The output is scaled back to bfloat16 by multiplying back the two scales: $s_w \times s_x$. When the tensors to be multiplied are larger than the PTC, partial outputs are stored in bfloat16 and reduced to the final output through accumulation support in the output pipeline. Only in operations involving the PTC do we use ABFP; all other operations—including nonlinearities and accumulations—are performed using bfloat16 to maintain the accuracy of the network.

Photonic processor architecture

The PTC architecture is outlined in Fig. 3a–c using a 4 × 4 example array for simplicity. Many photonic AI processor architectures presented so far cannot efficiently run data-dependent AI models such as transformers,

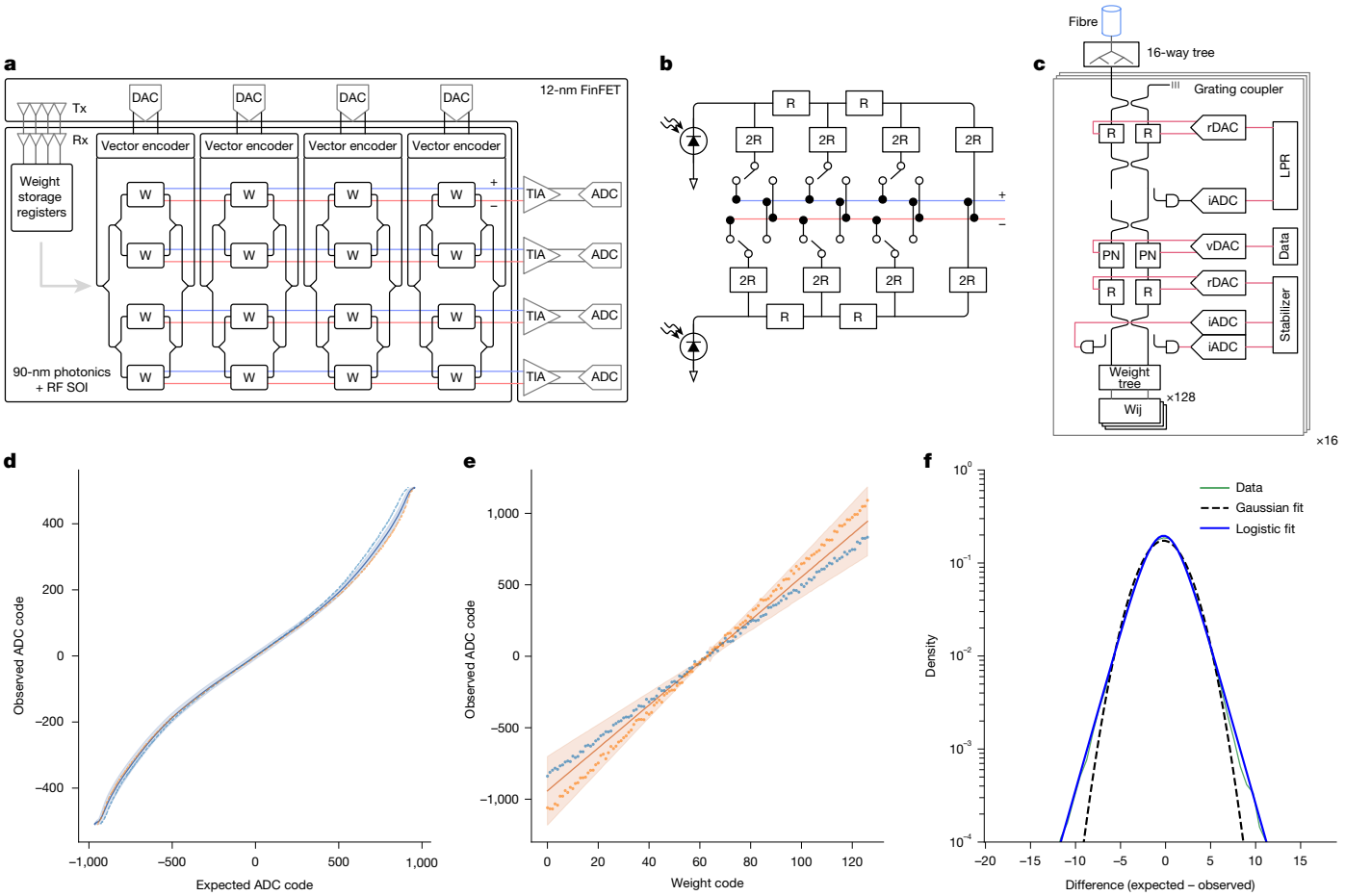

Fig. 3 | Unit cell operation and characterization. a, PTC architecture for a miniature 4 × 4 PTC. **b**, Weight unit cell consists of a differential photodetection unit coupled to a resistive, differential DAC. Example 3-bit circuit shown. **c**, Vector encoder unit cell comprises a MZI and temperature stabilization feedback loop. The grating coupler is used during wafer sort. **d**, Measured linearizing function for vector encoders in a photonic processor used as a lookup table to normalize nonlinearity with two randomly selected vector encoder linearizing functions plotted as dashed lines. **e**, Measured weight unit cell transfer function for 4,096 sample unit cells in a photonic processor with two randomly selected transfer functions plotted as points. **f**, Normalized error distribution between a set of 4,096 expected and observed MVPs with vectors and weights sampled from a normal distribution in bfloat16.

in which the next set of weights depends on the previous calculation—requiring symmetry between weight programming rates and vector programming rates while maintaining a degree of weight-stationarity. Some of the PTC's architectural choices are driven by the need to balance weight and vector programming rates for effective data-dependent AI computations while maximizing computational density.

Differential vector encoding and distribution are performed in the photonics domain with weight end points implemented as hybrid unit cells. Summation occurs by means of current addition along the differential metal wires. High-speed data conversion (including digital-to-analogue and analogue-to-digital) and amplification are performed in the DCI, using a high-performance, 12-nm-transistor process. Tensor operations, stabilization, local register storage and monitoring are conducted in the PTC. Weights are transferred across the weight transfer interface, a 3D-stacked electronic interconnect with the PTC directly below the DCI in the package. The weight unit cell is composed of differential photodetectors connected to a 7-bit segmented binary/R-2R ladder-based resistor digital-to-analogue converter (DAC) with an integral nonlinearity of 0.4 least significant bits (LSBs), a differential nonlinearity of 0.08 LSBs and a bandwidth of 1.5 GHz. The architecture of the weight unit cell is shown in Fig. 3b using an example 3-bit R-2R ladder for simplicity. The digital weight code that sets the switches effectively multiplies or scales the input photocurrent (representing the activation) by a number between −1 and +1. Each weight unit cell has a programmable slope parameter that enables calibration against fabrication errors. The mean weight transfer function for the PTCs with a ±1 standard deviation shaded region is shown in Fig. 3e. Saturation at the extremes is attributed to nonlinearity in the TIA preceding the ADC. The error at the mid-range code (zero value) is minimized by design, as the distribution of weight tensors in deep learning is typically centred around zero.

After weight programming, vector data are modulated onto a p–n junction-based Mach–Zehnder interferometer (MZI) using a 10-bit pseudo-differential resistive DAC[37]. The DAC has a segmented architecture with 6 bits encoded with binary weighting by means of an R-2R ladder and 4 bits with a thermometer encoding. The DAC can operate up to a 2 GHz update rate with an effective number of bits (ENOB) of 8.3 and a 3.2 V_{pp} differential output swing (1.6 V_{pp} singled-ended) at 72 mW. To account for thermal drift, the MZI is stabilized (as shown in Fig. 3c) using 1% tap monitor photodetectors, dual-slope integrating ADCs and a logic feedback controller. Light is delivered to the vector modulators through a 16-way photonic binary tree. The input power level is regulated to be constant within a fraction of a LSB using a laser power regulator, as shown in Fig. 3c. The mean transfer function for the vector modulators with a ±1 standard deviation shaded region is shown in Fig. 3d. Nonlinearity is attributed to the sinusoidal transfer function of the MZI.

Fig. 4 | Neural network tasks running on the photonic processor. a, Sample of the Oxford-IIIT Pet[48] dataset used to train SegNet and segmentation examples, as well as the segmentation results from the processor. **b**,**c**, Atari Pacman and DQN Atari Beamrider running on the photonic processor. **d**, Sample sentiment analyses performed by the photonic processor while executing BERT-Tiny on the IMDb sentiment analysis task. The examples provided illustrate the ability of the model to accurately classify movie reviews as positive or negative sentiment, demonstrating its proficiency in this task. **e**, Sample transcript generated by the photonic processor executing NanoGPT on the TinyShakespeare[44] text generation task. The transcript demonstrates the ability of the photonic processor to generate human-like text, highlighting its potential for generative language models.

A programmable TIA, consisting of a CMOS inverter-based pseudo-differential shunt-feedback TIA front end and Cherry–Hooper main amplifier[38], has a nominal transimpedance gain of 2.8 kΩ, bandwidth of 0.8 GHz, input-referred current noise of 0.4 μA_{rms}, maximum differential output voltage swing of 1.5 V_{pp} and power consumption of 41 mW. The TIA is followed by a 4× time-interleaved, two-stage pipelined successive approximation ADC[39–42]. The first pipeline stage resolves the first four (most significant) bits and the second stage resolves nine. There is one redundancy bit between the two pipeline stages and a second redundancy bit in the second successive approximation stage. The 11-bit ADC operates with a sampling rate of up to 2 GSPS with an ENOB of 9.8 bits while dissipating 66 mW, setting the peak tensor multiplication rate for the PTC.

The PTCs were programmed to implement 4,096 random MVPs with entries of the vectors and weights sampled from a normal distribution. The difference between the expected output codes and measured output codes is shown in Fig. 3f, along with a logistic distribution fit and a Gaussian distribution fit reference. The bfloat16 error in Fig. 3f takes into account the quantization error of the multiplication operands, as shown in Fig. 3d,e. The result of the multiplication of the operands also has analogue noise, which can be represented as a Gaussian error[43], and the output (with noise) is quantized again. The overall impact of the error propagation resembles a logistic distribution (see Supplementary information section IV).

The power consumption of each PTC was measured to be 3.2 W, including the stabilizer DACs and ADCs, weight unit DACs and all other components required to drive, control and stabilize the PTC, which are integrated on-chip. The DCI power consumption during AI workload execution was measured to be 65 W, including all components. For the complete photonic processor presented here, 43.7% of the power is spent on logic, the network-on-chip and SRAM, 35.6% on data converters, 4.2% on PCIe and 16.4% on the PTCs. At the time of publication, there are no comparable results that account for all of the practical energy costs associated with a full, end-to-end photonic processor operating with advanced neural networks. Number representation (for example, ABFP16) and computational precision substantially affect energy cost per calculation. We discuss this at length in the Supplementary information.

AI workload execution

In Fig. 4, we illustrate a subset of tasks executed on the photonic processor, using identical model weights and input data for both digital and photonic systems to ensure consistent evaluation. Linear and convolutional models, notably ResNet18, achieve accuracy levels comparable with those on digital FP32 platforms, even as dataset complexity increases. For classification tasks, the photonic processor consistently performs well, whereas regression tasks show a slight reduction in performance. All output values within a decision boundary are treated as the same result in classification tasks. By contrast, regression tasks rely on precise output values, which makes them more noise-sensitive. This distinction becomes evident in the BERT-Tiny transformer model, which performs robustly on the IMDb classification task but exhibits reduced accuracy on the SQuAD task, a regression-oriented task. Inference throughput, measured in inferences per second, is discussed in Supplementary information section II. Workload accuracy is listed in Table 1.

For reinforcement learning tasks, including the Atari game models[22], the photonic processor plays game environments such as Beamrider and Pacman, albeit with performance below FP32 levels. Specifically, Beamrider runs 6,430 steps on the photonic processor versus 30,304 steps on an FP32 processor and Pacman runs 1,825 steps on the photonic processor versus 3,329 steps on an FP32 processor.

Table 1 | Listing of neural network architectures and tasks run on the photonic processor compared with 32-bit floating point test accuracy

Model type	Model	Task	FP32 (%)	Here (%)	% of FP32
Convolutional/linear	ResNet18	CIFAR-10	88.3	86.4	97.8
Convolutional/linear	ResNet18	Imagenette	85.0	79.3	93.3
Convolutional/linear	ResNet18	ImageWoof	84.2	79.7	94.6
Convolutional/linear	ResNet18	MNIST	99.4	99.3	99.8
Multilayer perceptron	Three-layer linear	MNIST	76.7	74.0	96.5
Transformer	BERT-Tiny	IMDb	86.2	83.2	96.5
Transformer	BERT-Tiny	SQuAD (F1)	43.5	12.0	27.5
Convolutional/encoder–decoder	SegNet	Oxford-IIIT Pet	82.3	63.7	77.4
Convolutional/linear	Traffic light	Traffic light	100.0	99.0	99.0

Across these AI workloads, we attribute accuracy degradation primarily to limited optical power in the photonic processor owing to nonlinear absorption in silicon waveguides in the vector encoder circuit. Silicon waveguides are only used in regions of the chip that require thermo-optic modulation or high-speed modulation—all other waveguide routing in the chip is performed in silicon nitride. This nonlinear absorption limits the gain of the system to 1.86 rather than the design target gain of four (discussed in Fig. 2d). Larger models—such as ResNet34 with 21.7 million parameters compared with ResNet18's 11.6 million—demonstrate reduced accuracy, attributed to reduced gain. The impact of gain on neural network accuracy is further explored in the Supplementary information. Owing to an unoptimized DCI clock tree, these workloads were executed at a clock rate of 500 MHz, as discussed in the Supplementary information. However, the photonic processor can operate (with intermittent errors) at 2 GHz clock rate (262 trillion ABFP16 operations per second).

Techniques such as quantization-aware training and fine-tuning[32,33] show substantial potential for enhancing the accuracy of AI models executed on photonic hardware by addressing noise and precision limitations inherent to analogue computations. Quantization-aware training, in particular, allows models to adapt to low-precision representations by simulating quantization effects during training. This often results in performance closer to 32-bit digital precision. To evaluate this, we retrained ResNet18 with the ImageWoof dataset, achieving a marked improvement in test accuracy, increasing it from 64.4% to 79.7%. All other AI models were executed out-of-the-box, without retraining of any kind. This result illustrates the applicability of training techniques to improving accuracy in the presence of analogue noise.

Generative models, such as NanoGPT[44], exhibit output variability attributed to the inherent randomness in the computations of the photonic processor, which closely resembles outputs from digital hardware under similar constraints. Fine-tuning and quantization-aware training would probably further enhance performance on reinforcement learning tasks, such as those involving the Atari Deep Q-Network, by enabling the photonic processor to handle more complex decision boundaries and reducing noise impact on precise reward prediction. Given their sensitivity to numerical precision, regression tasks stand to gain considerably from these techniques, as they would enable finer control over weight adjustments and increase output stability.

Conclusion

This work represents a substantial advancement in photonic computing for AI, showcasing a photonic processor that achieves near-digital precision and performance on complex, out-of-the-box AI models. Successfully implementing models such as ResNet and BERT and demonstrating proficiency with reinforcement learning algorithms, including the DeepMind Atari deep learning algorithm, this photonic processor marks an essential step in post-transistor computing. These results validate the potential of photonics as a competitive alternative to conventional AI accelerators and set the stage for future developments in photonics-based AI technology. As research in this field progresses, photonic processors may play an important role in meeting the growing computational demands of AI and machine learning applications.

Although the photonic processor presented here demonstrates notable advancements in computational capability, several challenges and opportunities remain for the field to address. The PTC occupies 350 mm^2—less than half of a full-reticle die area (typically 850 mm^2). The PTC footprint can be further reduced by making use of advanced packaging technologies that enable reduced interconnect pitch between the DCI and the PTC (for example, hybrid bonding or microbumps). By implementing these changes, we believe that the current architecture could be scaled to 512 × 512 compute units in a single, full-reticle die. 3D stacking of PTCs and DCIs could enable arrays far beyond the four-core implementation shown here.

The energy efficiency of photonic computation scales nonlinearly with the size of the tensor core, with larger cores offering much greater efficiency[5]. However, realizing this advantage requires new neural network architectures designed to minimize memory lookups by performing more computation per data load. This shift could substantially reduce energy spent on data movement and memory access, addressing dominant bottlenecks in AI systems optimized for graphics and tensor processing units[23]. The photonic processor architecture demonstrated here can be augmented using wavelength division multiplexing by adding input multiplexers and output demultiplexers. This offers a path towards increasing the computational density of the photonic processor presented here.

The optical power required for higher clock rates is limited by nonlinear absorption in silicon photonic devices, particularly in p–n junctions and thermo-optic phase shifters. Future designs must overcome this limitation through innovations in device engineering, adaptive power control or hybrid materials that complement the properties of silicon.

Advances in materials science hold transformative potential for photonic processors. New materials, such as lithium niobate[45], barium titanate[46] and emerging two-dimensional systems[47], could reduce optical nonlinearity, increase modulation speeds and enable more compact unit cells. These innovations promise to enhance computational capability, energy efficiency and scalability, opening the door to hybrid photonic-electronic platforms.

The interdisciplinary nature of these challenges demands collaboration across photonics, materials science and machine learning. Addressing these hurdles will drive the use of photonic processors as a next-generation AI hardware platform.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information,

acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-025-08854-x.

1. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
2. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2017).
3. He, K., Zhang, X., Ren, S. & Sun, J. in *Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (IEEE, 2016).
4. Vinyals, O. et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* **575**, 350–354 (2019).
5. Shen, Y. et al. Deep learning with coherent nanophotonic circuits. *Nat. Photonics* **11**, 441–446 (2017).
6. Lin, X. et al. All-optical machine learning using diffractive deep neural networks. *Science* **361**, 1004–1008 (2018).
7. Hamerly, R., Bernstei, L., Sludds, A., Soljačić, M. & Englund, D. Large-scale optical neural networks based on photoelectric multiplication. *Phys. Rev. X* **9**, 021032 (2019).
8. Feldmann, J. et al. Parallel convolutional processing using an integrated photonic tensor core. *Nature* **589**, 52–58 (2021).
9. Dong, B. et al. Partial coherence enhances parallelized photonic computing. *Nature* **632**, 55–62 (2024).
10. Dong, B. et al. Higher-dimensional processing using a photonic tensor core with continuous-time data. *Nat. Photonics* **17**, 1080–1088 (2023).
11. Becker, S., Englund, D. & Stiller, B. An optoacoustic field-programmable perceptron for recurrent neural networks. *Nat. Commun.* **15**, 3020 (2024).
12. Chen, Z. et al. Deep learning with coherent VCSEL neural networks. *Nat. Photonics* **17**, 723–730 (2023).
13. Wang, T. et al. An optical neural network using less than 1 photon per multiplication. *Nat. Commun.* **13**, 123 (2022).
14. Sludds, A. et al. Delocalized photonic deep learning on the internet's edge. *Science* **378**, 270–276 (2022).
15. Shalf, J. The future of computing beyond Moore's Law. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **378**, 20190061 (2020).
16. Schwierz, F. & Liou, J. J. in *Proc. 2020 IEEE Latin America Electron Devices Conference (LAEDC)* 1–4 (IEEE, 2020).
17. Leiserson, C. E. et al. There's plenty of room at the Top: what will drive computer performance after Moore's law? *Science* **368**, eaam9744 (2020).
18. Moore, G. E. Cramming more components onto integrated circuits. *Proc. IEEE* **86**, 82–85 (1998).
19. Waldrop, M. M. The chips are down for Moore's law. *Nature* **530**, 144–147 (2016).
20. Vaswani, A. et al. in *Proc. Advances in Neural Information Processing Systems 30* (eds Guyon, I. et al.) 5998–6008 (Curran Associates, 2017).
21. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. in *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (eds Burstein, J., Doran, C. & Solorio, T.) 4171–4186 (Association for Computational Linguistics, 2019).
22. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
23. Jouppi, N. P. et al. in *Proc. 44th Annual International Symposium on Computer Architecture (ISCA '17)* 1–12 (ACM, 2017).
24. Peng, B., Hua, S., Su, Z., Xu, Y. & Shen, Y. in *Proc. 2022 IEEE Photonics Conference (IPC)* (IEEE, 2022).
25. Youngblood, N. Coherent photonic crossbar arrays for large-scale matrix-matrix multiplication. *IEEE J. Sel. Top. Quantum Electron.* **29**, 1–11 (2023).
26. Zhang, H. et al. An optical neural chip for implementing complex-valued neural network. *Nat. Commun.* **12**, 457 (2021).
27. Wetzstein, G. et al. Inference in artificial intelligence with deep optics and photonics. *Nature* **588**, 39–47 (2020).
28. Demirkiran, C. et al. An electro-photonic system for accelerating deep neural networks. *ACM J. Emerg. Technol. Comput. Syst.* **19**, 1–31 (2023).
29. Pintus, P. et al. Integrated non-reciprocal magneto-optics with ultra-high endurance for photonic in-memory computing. *Nat. Photonics* **19**, 54–62 (2025).
30. Shastri, B. J. et al. Photonics for artificial intelligence and neuromorphic computing. *Nat. Photonics* **15**, 102–114 (2021).
31. Xu, X. et al. 11 TOPS photonic convolutional accelerator for optical neural networks. *Nature* **589**, 44–51 (2021).
32. Jacob, B. et al. in *Proc. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2704–2713 (IEEE, 2018).
33. Courbariaux, M., Bengio, Y. & David, J.-P. Training deep neural networks with low precision multiplications. Preprint at https://arxiv.org/abs/1412.7024 (2015).
34. Kirtas, M. et al. Mixed-precision quantization-aware training for photonic neural networks. *Neural Comput. Appl.* **35**, 21361–21379 (2023).
35. Basumallik, A. et al. Adaptive block floating-point for analog deep learning hardware. Preprint at https://arxiv.org/abs/2205.06287 (2022).
36. Giewont, K. et al. 300-mm monolithic silicon photonics foundry technology. *IEEE J. Sel. Top. Quantum Electron.* **25**, 1–11 (2019).
37. Ghafarian, H. et al. A 9-bit, 45 mW, 0.05 mm^2 source-series-terminated DAC driver with echo canceller in 22-nm CMOS for in-vehicle communication. *IEEE Solid-State Circuits Lett.* **4**, 10–13 (2021).
38. Yu, K. et al. in *Proc. 2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers* https://doi.org/10.1109/isscc.2015.7063098 (IEEE, 2015).
39. McCreary, J. L. & Gray, P. R. All-MOS charge redistribution analog-to-digital conversion techniques. I. *IEEE J. Solid-State Circuits* **10**, 371–379 (1975).
40. Jang, M. et al. Design techniques for energy-efficient analog-to-digital converters. *IEEE Open J. Solid-State Circuits Soc.* **3**, 145–161 (2023).
41. Ramkaj, A. T. et al. A 5-GS/s 158.6-mW 9.4-ENOB passive-sampling time-interleaved three-stage pipelined-SAR ADC with analog-digital corrections in 28-nm CMOS. *IEEE J. Solid-State Circuits* **55**, 1553–1564 (2020).
42. Lagos, J. et al. A 10.1-ENOB, 6.2-fJ/conv.-step, 500-MS/s, ringamp-based pipelined-SAR ADC with background calibration and dynamic reference regulation in 16-nm CMOS. *IEEE J. Solid-State Circuits* **57**, 1112–1124 (2022).
43. de Lima, T. F. et al. Noise analysis of photonic modulator neurons. *IEEE J. Sel. Top. Quantum Electron.* **26**, 1–9 (2020).
44. Karpathy, A. nanoGPT. *GitHub* https://github.com/karpathy/nanoGPT (2023).
45. Wang, C. et al. Integrated lithium niobate electro-optic modulators operating at CMOS-compatible voltages. *Nature* **562**, 101–104 (2018).
46. Abel, S. et al. Large Pockels effect in micro- and nanostructured barium titanate integrated on silicon. *Nat. Mater.* **18**, 42–47 (2019).
47. Youngblood, N., Chen, C., Koester, S. J. & Li, M. Waveguide-integrated black phosphorus photodetector with high responsivity and low dark current. *Nat. Photonics* **9**, 247–252 (2015).
48. Parkhi, O. M., Vedaldi, A., Zisserman, A. & Jawahar, C. V. in *Proc. 2012 IEEE Conference on Computer Vision and Pattern Recognition* 3498–3505 (IEEE, 2012).

Methods

Chip assembly

The assembly process begins with attaching the DCI to the top of the interposer using a lead-free mass reflow process. The PTC is then attached to the bottom of the interposer by means of lead-free mass reflow. The interposer uses stacked vias and densely pitched, plated through holes on the top and bottom layers to enable direct communication between the DCI and PTC dies while minimizing channel losses. Passive components and ball grid array elements are subsequently attached to the interposer, followed by the attachment of a 65 × 80-mm organic substrate, which serves as the foundation for the subassembly. Heat spreaders are integrated with the DCI and PTC chips, as well as the interposer, to ensure efficient thermal management.

Next, mechanical epoxy is dispersed in the strain relief area of the substrate and index-matching epoxy is dispensed into the V-groove and the undercut of the PTC's silicon substrate coupler. Fibre array units align with the V-grooves using an active alignment process to deliver light to the PTC dies. After alignment, the epoxy is cured using ultraviolet light and thermal treatment. A final electrical test and an insertion loss measurement validate the assembly. Once completed, the entire package is mounted onto a PCIe card, finalizing the integration.

PTC fabrication

PTCs were fabricated using GlobalFoundries' CMS90WG silicon photonics platform on 300-mm silicon-on-insulator (SOI) wafers. The process flow begins with the preparation of the SOI substrate, in which a thin silicon device layer (160 nm thick) is used for photonic device fabrication. Initial steps involve defining silicon waveguides and photonic components through immersion lithography, followed by reactive ion etching to shape the waveguides. Subsequent annealing ensures smooth sidewalls, reducing optical propagation losses.

Standard CMOS steps, including gate stack formation, source-drain doping and spacer creation, take place, along with photonic device formation. The CMOS fabrication uses a 90-nm-transistor process for co-integrated electronic components. The photonic devices undergo doping and ion implantation to create active components, such as modulators and thermal phase shifters. The next step involves the integration of germanium photodiodes, which are formed by selective epitaxial growth within pre-etched trenches in the silicon device layer. This ensures seamless alignment with the optical waveguides.

The backend-of-line processing stages integrate the photonic and electronic components. Multilayer copper interconnects are deposited and patterned to form electrical connections between the CMOS and photonic devices. Dielectric layers are added to isolate these interconnects and reduce parasitic capacitance. Finally, fibre-alignment features, such as V-grooves, are etched into the wafer to facilitate optical coupling, and the wafers are diced into individual chips. This end-to-end process enables monolithic integration of photonic and electronic components in the PTCs, ensuring compactness and high performance.

Laser performance

The laser source is a high-power distributed feedback (DFB) laser operating in the 1,310-nm wavelength band, specifically designed to operate on the CWDM4 wavelength grid. Our work uses eight lasers on a PCB card, each delivering an optical output power of 200 mW (in fibre) at operating temperatures of 45 °C. The electrical power consumption, including the thermoelectric cooler, is 1.25 W for each laser. The distributed feedback laser linewidth is 0.15 nm, demonstrating single-mode operation with a side-mode suppression ratio specification exceeding 30 dB minimum and a wavelength tuning coefficient of 0.1 nm per °C. Each laser uses a single laser die packaged in a 14-pin butterfly package.

Photonic device performance

The V-groove coupling loss is 1.56 ± 0.05 dB. The on-chip photodiodes are 8 µm long, with a responsivity of 1.08 ± 0.02 A W^{-1}, a dark current of 12.88 ± 2.13 nA and a 3-dB bandwidth of 10.6 ± 1.0 GHz. The p–n, depletion-based vector modulator 3-dB bandwidth is 3.8 ± 0.7 GHz, with an insertion loss of 1.1 dB, a $V\pi L$ of 0.92 V cm and a footprint of 91 × 770 µm. The silicon waveguide loss for the 440-nm-wide ridge waveguides is 1.35 ± 0.1 dB cm^{-1} and 0.68 ± 0.05 dB cm^{-1} for the 2-µm-wide silicon nitride waveguides. The linear loss in the photonic path is approximately 4.6 dB, with a nonlinear loss of 2.2 dB owing to the silicon sections of the chip around the vector modulator. The linear loss breakdown is as follows: silicon waveguides 1.4 dB, y-junctions 0.4 dB, bends 0.2 dB, p–n-phase junction 1.1 dB, thermal phase shifter 0.1 dB, nitride waveguides 0.7 dB, nitride y-junctions 0.3 dB and silicon nitride bends 0.4 dB.

Data availability

The datasets presented in this study and analysis programs are available at https://github.com/lightmatter-ai/upaia-paper-2025.

Acknowledgements We would like to thank K. C. Buckenmaier, M. Gould, C. Ramey, B. Dobbie, S. McKenzie, O. Yildirim, J. Talmage and M. Todd for their early contributions to the development of the photonic processor. We would also like to thank C. McCarter, N. Dronen, M. Forsythe, T. Lazovich, L. Levkova, D. Walter and D. Widemann for the development and implementation of the ABFP format. Also, we thank C. Chan, P. Clark, S. Cyphers, L. Huang, E. Hein, A. Hussein, S. Iyer, T. Kenney, S. Lines, A. Romano, T. Sarvey and Y. Sanders for their early contributions to the development of the software framework.

Author contributions S.R.A., R.Ba., N.B., R.Br., J.Co., C.C., P.C., J.Ca., K.D., C.D., J.E., B.G., E.G., S.G., R.H., R.J., B.J., A.K., A.Me., E.R., S.S., N.S., J.S., M.T., A.W., J.Z., D.B. and N.C.H. contributed to the design and development of the photonic processor hardware. M.B., A.B., A.O., M.C., P.H., A.Ma., N.M., L.N., S.P., R.Pa., R.Pe., K.W., G.W. and H.J.L. contributed to the design and development of the software stack for the photonic processor. All authors contributed to the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-025-08854-x.
Correspondence and requests for materials should be addressed to Ayon Basumallik, Darius Bunandar or Nicholas C. Harris.
Peer review information *Nature* thanks Anthony Rizzo and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.