Experimentally realized *in situ* backpropagation for deep learning in photonic neural networks

Sunil Pai,[1,2,*] Zhanghao Sun,[1] Tyler W. Hughes,[3,4] Taewon Park,[1] Ben Bartlett,[3,2] Ian A. D. Williamson,[1,5] Momchil Minkov,[1,4] Maziyar Milanizadeh,[7] Nathnael Abebe,[1,6] Francesco Morichetti,[7] Andrea Melloni,[7] Shanhui Fan,[1] Olav Solgaard,[1], and David A.B. Miller[1]

[1]Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA
[2]now at PsiQuantum, Palo Alto, CA, USA
[3]Department of Applied Physics, Stanford University, Stanford, CA 94305, USA
[4]now at Flexcompute Inc., Belmont, MA, USA
[5]now at X Development LLC, Mountain View, CA USA
[6]now at Google, Mountain View, CA USA
[7]Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy
[*]To whom correspondence should be addressed; E-mail: spai@psiquantum.com.

Integrated photonic neural networks provide a promising platform for energy-efficient, high-throughput machine learning with extensive scientific and commercial applications. Photonic neural networks efficiently transform optically-encoded inputs using Mach-Zehnder interferometer mesh networks interleaved with nonlinearities. We experimentally trained a three-layer, four-port silicon photonic neural network with programmable phase shifters and optical power monitoring to solve classification tasks using "*in situ* backpropagation," a photonic analogue of the most popular method to train conventional neural networks. We measured backpropagated gradients for phase shifter voltages by interfering forward- and backward-propagating light and simulated *in situ*

backpropagation of 64-port photonic neural networks trained on MNIST image recognition given errors. All experiments performed comparably to digital simulations (>94% test accuracy), and energy scaling analysis indicated a route to scalable machine learning.

Introduction Neural networks (NNs) are ubiquitous computing models loosely inspired by the structure of a biological brain. Such models are trained on input data to implement complex signal processing or "inference" (*1, 2*), powering various modern technologies ranging from language translation to self-driving cars. The required energy for training and inference to power these technologies has recently been estimated to double every 5 to 6 months (*3*), and thus necessitates an energy-efficient hardware implementation for NNs.

To address this problem, programmable photonic neural networks (PNNs) have been proposed as a promising, scalable, and mass-manufacturable integrated photonic hardware solution (*4*). A popular implementation of PNNs consists of silicon photonic meshes, $N \times N$ networks of Mach-Zehnder interferometers (MZIs) and programmable phase shifters (*5–7*), which optically accelerate the most expensive operation in a PNN: unitary matrix-vector multiplication (MVM). The MVM $\mathbf{y} = U\mathbf{x}$ is implemented by simply sending an input mode vector $\mathbf{x}$ (optical phases and modes in N input waveguides) through the network implementing U to yield output modes $\mathbf{y}$ (*4, 6, 8*). This fundamental mathematical operation, based on optical scattering theory, additionally enables various analog signal processing applications beyond machine learning (*4, 9*) such as telecommunications (*8*), quantum computing (*10, 11*), and sensing (*12*).

Recently, "hybrid" PNNs, which alternately cascade photonic meshes and digital nonlinear activation functions (*9, 13*), have proven to be a low-latency and energy-efficient solution for NN inference in circuit sizes of up to 64×64 (*14*). Compared to fully analog PNNs with optical nonlinear activations (*15, 16*), hybrid PNNs get around the critical problem of photonic loss

and offer more versatility than multilayer PNNs for between-layer logical operations that do not favor optics. Such features may be present in a number of state-of-the-art machine learning architectures such as recurrent neural networks (*17*) and transformers (*18, 19*). When fully optimized, the energy efficiency of hybrid PNNs has been estimated to be one to two orders of magnitude higher than state-of-the-art digital electronic application specific integrated circuits (ASICs) in AI (*20*). However, despite the success in PNN-based inference, on-chip training of PNNs has not been demonstrated due to significantly higher experimental complexity compared to the inference procedure.

In this paper, we experimentally demonstrated photonic implementation of backpropagation, the most widely used method of training NNs (*1, 2*). (A minimal bulk optical demonstration has been previously explored (*21*).) Backpropagation is generally performed by propagating error signals backwards through the NNs to determine programmable parameter gradients via the chain rule. In our multilayer PNN device, we performed *in situ* training on a foundry-manufactured silicon photonic integrated circuit by sending light-encoded errors backwards through the PNN and measuring optical interference with the original forward-going "inference" signal (*22*). Once trained, our chip achieved similar accuracy to digital simulations, adding new capabilities beyond existing inference or *in silico* learning demonstrations (*4, 23, 24*). We further designed and experimentally validated an analog (electro-optic) phase shifter update protocol, a key improvement over past proposals requiring more energy-intensive and quantization error-prone "digital subtraction" (*22*). Finally, we systematically analyzed energy and latency advantages of *in situ* backpropagation and its scalability to larger (64×64) PNN systems. Our findings ultimately pave the way for energy-efficient optoelectronic training of neural networks and optical systems more broadly.

Photonic neural networks We built a hybrid PNN by alternating sequences of analog MVM operations $U^{(\ell)}(\vec{\eta}^{(\ell)})$ (implemented on a custom designed silicon photonic triangular mesh (*6*)) and digital nonlinear transformations $f^{(\ell)}$ (implemented using autodifferentiation software (*25, 26*)) as shown in Fig 1A-D, where layer $\ell \leq L$ (total of L layers). The PNN was parameterized by programmable phase shifts $\vec{\eta} \in [0, 2\pi)^D$, where D represents number of PNN phase shifters. Mathematically, the following "inference" function sequence transformed input $\mathbf{x} = \mathbf{x}^{(1)}$, proceeding in a "feedforward" manner to the output $\hat{\mathbf{z}} := \mathbf{x}^{(L+1)}$ (fig. 1A,B,D):

$$\mathbf{y}^{(\ell)} = U^{(\ell)}\mathbf{x}^{(\ell)} \tag{1}$$

$$\mathbf{x}^{(\ell+1)} = f^{(\ell)}(\mathbf{y}^{(\ell)}) \tag{2}$$

The "cost function" is defined as $\mathcal{L}(\mathbf{x}, \mathbf{z}) = c(\hat{\mathbf{z}}(\mathbf{x}), \mathbf{z})$, where c represents the error between $\hat{\mathbf{z}}$ and ground truth label $\mathbf{z}$. Backpropagation updates parameters $\vec{\eta}$ based on D-dimensional gradient $\partial\mathcal{L}/\partial\vec{\eta}$ evaluated for "training example" $(\mathbf{x}, \mathbf{z})$ (or averaged over a batch of examples).

Each MZI was parametrized by thermo-optic phase shifts controlled by a source module unit (fig. 2A,B). Phase shifts were placed at the input (ϕ, voltage V_ϕ) and internal (θ, voltage V_θ) arms of all MZIs to control propagation pattern of light enabling arbitrary unitary matrix multiplication. We embedded an arbitrary 4×4 unitary matrix multiply in a 6×6 triangular network of MZIs. This configuration incorporated two 1×5 photonic meshes on either end of the 4×4 "matrix unit" capable of sending any input vector $\mathbf{x}$ and measuring any output vector $\mathbf{y}$ from Eqs. 1 and 2. These calibrated "generator" and "analyzer" optical I/O circuits (figs. 1E and 2B) require calibrated voltage mappings $\theta(V_\theta), \phi(V_\phi)$ (*4, 27–29*) (fig. S4).

Backpropagation demonstration Our core result (fig. 1E) was experimental realization of backpropagation on a photonic triangular mesh MVM chip using a custom optical rig (fig. S3) (*22*). Our backpropagation-enabled architecture differs in three ways from a typical PNN photonic mesh (*4*):

1. We enabled "bidirectional light propagation," the ability to send and measure light propagating left-to-right or right-to-left through the circuit (as depicted in fig. 1E).

2. We implemented "global monitoring" to measure optical power p_η propagating through any phase shift η in the circuit using 3% grating taps (shown in the inset of fig. 1E and fig. 2A,B). In our proof-of-concept setup, we used an IR camera mounted on an automated stage to image these taps throughout the chip (fig. S3E).

3. We implemented both amplitude and phase detection (improving on past approaches (*30*)) using a self-configuring programmable matrix unit layer (*27, 31*) on both generator and analyzer subcircuits (fig. 1E and fig. 2B), which by symmetry worked for sending and measuring light that propagated forward or backward through the mesh.

These improvements on an already versatile hardware platform enabled backpropagation entirely using physical measurements of field intensities to obtain loss gradients. As shown in fig. 1E, backpropagation (*22*) required global optical monitoring. Furthermore, bidirectional optical I/O was required to switch between forward- and backward-propagating signals to experimentally realize *in situ* backpropagation. Equipped with these additional elements, our protocol can be implemented on any feedforward photonic circuit (*32*) with the requisite analyzer and generator circuitry (fig. S1-2).

Here we give a quick summary of the procedure explained further in Supplementary Text. The "forward inference" signal $\mathbf{x}^{(\ell)}$ and "backward adjoint" signal $\mathbf{x}_{\mathrm{adj}}^{(\ell)}$ are sent forward and backward respectively through the mesh that implements $U^{(\ell)}$. The "sum" vector $\mathbf{x}^{(\ell)} - i(\mathbf{x}_{\mathrm{adj}}^{(\ell)})^*$ is sent forward and subtracting the forward and backward measurements from it digitally yields the gradient (*22*), a reverse-mode differentation process we call an "optical vector-Jacobian product (VJP)."

Analog update Going beyond an experimental implementation of the theoretical proposal of Ref. (*22*), we additionally explored a more energy-efficient fully analog gradient measurement update for the final step avoiding a digital subtraction update. Instead of global monitoring the first two steps and the final "sum" step, we toggled an adjoint phase $\zeta(t)$, a square wave modulation with period T that periodically toggles between "sum" and "difference" settings $\zeta = 0$ and π corresponding to signal inputs $\mathbf{x}_{\pm}^{(\ell)} = \mathbf{x}^{(\ell)} \mp i(\mathbf{x}_{\text{adj}}^{(\ell)})^*$. The gradient is $\partial\mathcal{L}/\partial\eta = (p_{\eta,+} - p_{\eta,-})/4$, or half the "signed amplitude" of the AC (mean-subtracted) signal (Supplementary Text, fig. S6E). The sum and difference inputs $\mathbf{x}_{\pm}^{(\ell)}$ were computed digitally (off-chip), requiring $\mathcal{O}(N)$ operations to compute per input. The sum and difference inputs were directly programmed at the generator to compute phase gradients, subtracted in the analog domain to update phase shift voltages. One option to efficiently achieve a periodic ζ toggle is to use the summing architecture in fig. 2C which sums $\mathbf{x}^{(\ell)}$ and $i(\mathbf{x}_{\text{adj}}^{(\ell)})^*$ interferometrically with a fast modulator implementing ζ. In an optimized scheme, we would physically measure the gradient and update the phase shift voltage in the analog domain using a photodiode, differential amplifier (implementing an analog subtraction), and a "sample-and-hold" update circuit using only a single toggle (fig. S6B,C). A simple experimental demonstration of the gradient measurement of this circuit for a single phase shifter demonstrated the logic of this electronic feedback scheme, which was extended to "batch updates" incorporating data from multiple training examples ultimately required for our approach to scale (fig. S7). This approach avoided a costly digital-analog and analog-digital conversion and additional digital memory complexity required to program N^2 elements, enabling a truly analog backpropagation scheme.

The local feedback just described updates each phase shifter η using the measured gradient:

$$\frac{\partial\mathcal{L}}{\partial\eta} = \mathcal{I}(x_\eta x_{\eta,\text{adj}}) = \frac{|x_{\eta,+}|^2 - |x_\eta|^2 - |x_{\eta,\text{adj}}|^2}{2} = \frac{p_{\eta,+} - p_\eta - p_{\eta,\text{adj}}}{2} = \frac{p_{\eta,+} - p_{\eta,-}}{4}, \quad (3)$$

where $x_{\eta,+} = x_\eta - ix_{\eta,\text{adj}}^*$ and the last equation indicates the equivalence of "digital subtraction,"

(figs. 1E and 3) and our proposed "analog subtraction" scheme (figs. 2C-D, 4 and S6-7). Pseudocode and the complete backpropagation protocol are discussed in the Supplementary Text. Note that digital and analog gradient updates can both be implemented in parallel across all PNN layers.

We experimentally estimated the accuracy of the analog gradient measurement for a matrix optimization problem (*7*) by digital processing of the optical power measurements (fig. 2D). We programmed a sequence of inputs into the generator unit of our chip and recorded the square wave response oscillating between $p_{\eta,+}$ and $p_{\eta,-}$ and separately subtracted the two measurements to find the gradient with respect to η.

We implemented *in situ* backpropagation in a single photonic mesh layer optimizing the cost function defined for output port i via $\mathcal{L}_r = 1 - |\widehat{\mathbf{u}}_r^T \mathbf{u}_r^*|^2$ or a "batch" cost function $\mathcal{L} = \sum_{r=1}^{4} \mathcal{L}_r$. Here, $\mathbf{u}_r$ is row r of U, a target matrix that we chose to be the four-point discrete Fourier transform (DFT), and $\widehat{\mathbf{u}}_r$ is row r of $\widehat{U}$, the implemented matrix on the device. For our gradient measurement step, we sent in the derivative $\mathbf{y}_{\mathrm{adj}} = \partial \mathcal{L}_r / \partial \mathbf{y} = -2(\widehat{\mathbf{u}}_r^T \mathbf{u}_r^*)^* \mathbf{e}_r$ to measure an adjoint field $\mathbf{x}_{\mathrm{adj}}$, where $\mathbf{e}_r$ is the r^{th} standard basis vector (1 at position m, 0 everywhere else).

We evaluated gradient direction error as $1 - \mathbf{g} \cdot \hat{\mathbf{g}}$ comparing normalized measured ($\hat{\mathbf{g}}$) and predicted gradients $\mathbf{g} = \partial \mathcal{L} / \partial \vec{\eta} \cdot \|\partial \mathcal{L} / \partial \vec{\eta}\|^{-1}$. Both digital and analog gradients were less accurate near convergence (fig. 2F) with the errors empirically increasing quadratically as the inverse of fidelity error (Methods, fig. 2F). The analog batch gradient (trained by summing all four gradients together to give $\partial \mathcal{L} / \partial \eta$) validated the photonic portion of the batch scheme in fig. S6B (with the electronic portion being separately demonstrated in fig. S7). All gradient errors, regardless of implementation, scaled similarly with convergence distance; uncalibrated thermal crosstalk likely resulted in gradient measurement errors comparable to systematic power errors at the taps. Digital subtraction (fig. 1E) encountered different losses and coupling efficiencies in bidirectional gratings, whereas analog gradient measurements involved subtraction of only

forward-going fields at forward gratings, likely resulting in superior performance (fig. 2F). Finally, error in the full analog subtraction scheme was independent of batch size for the gradient calculation, and no significant deviation due to electronic jitter or signal distortion was observed (fig. S7D).

Photonic neural network training To test overall on-chip training, we assessed accuracy of *in situ* backpropagation to train multi-layer PNNs using digital subtraction protocol in Ref. (*22*) (fig. 3A). We trained our chip to implement $L = 3$ layers with $N = 4$ ports to assign labelled noisy synthetic data, generated using Scikit-Learn (*33*), in 2D space to a 0 or 1 label based on the point's spatial location (fig. 1A and 3E,H). We performed a 80%:20% train-test split (200 train points, 50 test points) to avoid overfitting.

To implement classification, our PNN assigned a probability to each point being assigned a 0 or 1 based on the following model:

$$\hat{\mathbf{z}}(\mathbf{x}) = \mathrm{softmax2}(|U^{(3)}|U^{(2)}|U^{(1)}\mathbf{x}|||), \tag{4}$$

where softmax2 is the standard softmax (normalized sigmoid) function applied to two quantities: the total power in outputs 1, 2 and total power in ports 3, 4. The input data $\mathbf{x}$ was engineered such that any 2D point had the same total input power as a four port vector (Methods). Each point was classified red or blue (0 or 1 respectively) based on whether output of eq. 4 obeyed the condition $z_0 > z_1$ for each input (fig. 3), which we optimized using a cross entropy cost function (Methods).

Our chip performed data input, output and matrix operations for all PNN layers. At each layer output, we digitally performed a square-root operation on output power to implement absolute value nonlinearities (off-chip via JAX and Haiku (*25, 26*)) and recorded output phases for the backward pass of *in situ* backpropagation. Ideally, PNNs are controlled by separate photonic meshes of MZIs for each linear layer to achieve low power consumption. However, to save on

footprint we reprogrammed the same chip to perform successive linear layers since basic operating principles remain the same. We used the Adam gradient update (*34*) with a learning rate of 0.01 and performed digital simulations at each step to fully compare measured and predicted performance. Before on-chip training experiments, we performed forward inference with digitally pre-trained neural network weights to verify accurate calibration. We achieved 90% and 98% device test set accuracy for ring and moons datasets respectively (fig. S5I,J). Since our photonic and digital implementation agreed closely in inference accuracy, we performed network training on-chip while conducting evaluations off-chip for convenience.

During training of the circle dataset, predicted and measured powers for grating tap-to-camera monitor measurements showed excellent agreement across all waveguide segments required for accurate gradient computation. The training curves in fig. 3C indicate that stochastic gradient descent was a highly noisy training process for both predicted and measured curves due to the noisy synthetic dataset about the boundary and our choice of single-example training. These large swings appeared roughly correlated between the simulated and measured training curves (fig. 3E), and we successfully achieved 96% train and 93% test model accuracy (fig. 3D). We then trained the moons dataset, applying same procedure to achieve 87% train and 94% test model accuracy (fig. 3F, green vs red). When using the predicted phase for phase measurement, we reduced gradient error by roughly an order of magnitude on average resulting in 95% train and 98% test model accuracy which agreed with digital training (fig. 3F-H). This improvement underscores the importance of accurate phase measurement for improved training efficiency.

Simulations and scalability Given that our experimental results for $N = 4$ PNNs showed evidence of hardware error impacting training, we assessed the scalability for $N = 64$ PNNs in the presence of error to better understand the relative contributions at scale. We implemented a PNN simulation framework in Simphox (*35*) using JAX and Haiku (*25, 26*) to simulate an

in situ backpropagation training given a grid search of systematic and noise errors (Methods). After 100 epochs using $M = 600$ batch size, we achieved a maximum test accuracy of roughly 97.2% in the ideal case and a performance degradation to roughly 95% on average (fig. 4B,C). Phase and amplitude errors arising from photodetector noise and phase shift quantization and calibration errors affected convergence in error the most. This suggests in-situ backpropagation is relatively robust to noise and hardware errors at scale, which are difficult to totally eliminate in current analog computing systems.

We also considered the energy and latency tradeoff with accuracy for the optimized analog gradient update scheme, adding more details about feasible gradient update circuit designs using current state-of-the-art electronics co-integrated with active photonic components (*36–39*). Collectively, our simulation results (fig. 4) and energy calculation contours (fig. S8, supported by tables S1-6) indicated minimal performance degradation for MNIST training simultaneously with $3\times$ improvement in energy efficiency assuming 100 fJ floating point operations for equivalent digital models (*40*) and tap noise factor of $s_{\mathrm{tap}} < 0.01$ in the regime where optical power begins to dominate the energy consumption. Errors may be further reduced by improving avalanche photodiode sensitivity, reducing optical component loss, or increasing overall input optical power, a key factor in the energy-error tradeoff (tables S1 and S5). Note these photodiode noise error considerations must be considered for inference tasks and photonic matrix multiplication tasks more broadly (*16, 41*).

Discussion and outlook In this paper, we demonstrated practically useful photonic machine learning hardware by physically measuring gradients calculated via interferometric measurements of *in situ* backpropagation (fig. 1). We concluded that gradient accuracy plays an important role in reaching optimal results during training (fig. 2). Optical I/O and calibration errors and photodetector noise at the global monitoring taps caused errors in gradient accuracy,

resulting in poorer convergence. By correcting for phase error, the resulting accurate gradients yielded training curves highly correlated to digital predictions (fig. 3). From these results, we determined that optical I/O calibration accuracy is vital; even though individual updates were ideally faster to compute, higher error resulted in effectively longer training times that mitigated this benefit. To better understand this tradeoff, we explored an optimized regime of our system, which considered co-integration of CMOS electronics with photonics (fig. S8, tables S1-6), and found that in the regime of photonic advantage (e.g., $N = 64$ at sufficiently large batch sizes), we could successfully train MNIST close to digital equivalents (fig. 4).

Our demonstration (fig. 3) and energy calculations (fig. S8) suggest that *in situ* backpropagation is the most efficient approach for training hybrid PNNs. Our hybrid training approach optically accelerated the most computationally intensive $\mathcal{O}(N^2)$ operations. All other $\mathcal{O}(N)$ computations, such as nonlinearities and their derivatives, were implemented digitally; this is reasonable because $\mathcal{O}(N)$ time is already needed to modulate and measure optical inputs and outputs for the overall network. In some cases, such as in data center machine learning and neural network accelerators (e.g., GPUs) with optical interconnects (where data is already optically encoded), our *in situ* backpropagation scheme opens up opportunities to improve energy efficiency of model training. Many digital schemes already are exploring reduction of the communication bottleneck in digital chips in the race to address the energy doubling AI problem (*3*). As optics is the ideal modularity for low-latency and low-energy signal communication, mixed-signal schemes are converging on some of these hybrid schemes that benefit from our proposal (*40, 42*).

Other techniques, e.g. population-based methods (*43*), direct feedback alignment (*44, 45*), and perturbative approaches (*16*), have some advantages but are ultimately less efficient for training neural networks versus backpropagation. This is especially true for hybrid PNNs, which unlike "receiverless" fully analog PNNs (*16*), require optoelectronic (i.e. digital-analog

and analog-digital) conversions for every update. In contrast to perturbative approaches, *in situ* backpropagation calculates gradients in a modular framework compatible with complex logic typically required in larger scale artificial intelligence applications.

Although this paper focuses on applications to hybrid PNNs, it is worthwhile noting that our backpropagation scheme can in principle be made to be compatible with all-optical or receiverless implementations (*15, 16*). Our alternate configuration enables all-optical inference and hybrid *in situ* backpropagation training (provided accurate nonlinearity models, fig. S8E) without incurring exponential loss in the number of layers L, which can limit gradient accuracy (fig. 4). Previous proposals suffer this exponential loss scaling because they propagate the same optical modes through all layers in a feedforward configuration (*15, 16*), whereas our proposal splits input light equally across the layers (linear scaling in number of layers). Furthermore, it is possible to switch between all-analog PNN inference and hybrid PNN training modes of operation (fig. S8F) or perform an all-analog backpropagation through a choice of nonlinearity (*46*); studying the scaling and errors of this scheme (given the need to calibrate nonlinearity models for training) are left to a future work.

Ultimately, however, such all-optical schemes suffer from limited versatility (data cannot be moved around at will in a scalable manner), still require input and output electronics (saving energy proportional to the number of mesh layers L), and the tradeoff between error sensitivity and energy consumption of the nonlinearity. A single photonic mesh circuit arbitrarily interfering large number of modes requires large-depth circuits pushing the limit of the footprint-loss tradeoff of an all-optical scheme. Instead, a hybrid PNN receives and re-injects optical power at various points (at least for all nonlinearities or lossy elements) to "rescue" lossy modes and scale PNNs to larger and, more importantly, deeper and more complex models that only use optics when convenient.

Based on these versatility principles, large scale hybrid PNN models have already achieved

high ResNet-50 image classification accuracy using existing commercially viable photonic circuits (*14*). Our proposal indicates a route to train models on such devices that few other training methods can efficiently produce. *In situ* backpropagation is also uniquely positioned to train "optical transformers" that leverage hybrid PNNs for computation; inference has been recently experimentally demonstrated in a free space computing element, a timely application for natural language and video processing applications (*19*). Due to the complex movement of data in many of these neural networks, hybrid PNNs offer extreme flexibility to tackle a number of complex problems and are not tied to any specific connectivity between layers, as is the case for all-optical solutions.

Given the generality of photonic meshes and wide applicability of gradient measurements, our protocol can be applied beyond the photonic machine learning to more traditional photonics problems including processing of free-space signals (*12, 47*) and, more generally, efficient photonic mesh calibration (*7, 32*). In fact, the analog gradient update experiment in fig. 2 specifically targets solving this problem of mesh calibration (*6*), because those gradients can be used to optimize the implemented matrix on the device. This is useful for linear optical elements with no obvious calibration scheme; for instance, our approach can be useful for training robust arbitrary unitary elements consisting of large multi-waveguide coupling elements (*48*) and more generally any reconfigurable optical device with active phase elements. More generally in the photonics space, our demonstration can be thought of as an experimental analogue of "inverse design" of photonic devices at the circuit level. Inverse design implements reverse-mode autodifferentiation with respect to material relative permittivity by interfering adjoint and forward fields, the basis of the original proof of *in situ* backpropagation (*22*) since phases are trivially related to material relative permittivity changes. This suggests an even broader application domain for our technique to arbitrary linear optical devices, including free-space and recirculating designs (*46, 49*).

Our results ultimately have wide-ranging implications for bridging the fields of photonics and machine learning. Backpropagation is the most efficient and widely used neural network training algorithm for machine learning, and our demonstration of this popular technique as a physical implementation presents promising capabilities of hybrid PNNs to reduce carbon footprint and counter the exponentially increasing costs of AI computation.

References and Notes

1. S. Linnainmaa, Taylor expansion of the accumulated rounding error, *BIT Numer Math* **16**, 146–160 (1976).

2. D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning representations by back-propagating errors, *Nature* **323**, 533–536 (1986).

3. J. Sevilla, L. Heim, A. Ho, T. Besiroglu, M. Hobbhahn, P. Villalobos, Compute Trends Across Three Eras of Machine Learning, *Proceedings of the International Joint Conference on Neural Networks* (2022).

4. Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, M. Soljačić, Deep learning with coherent nanophotonic circuits, *Nature Photonics* **11**, 441–446 (2017).

5. M. Reck, A. Zeilinger, H. J. Bernstein, P. Bertani, Experimental realization of any discrete unitary operator, *Physical Review Letters* **73**, 58–61 (1994).

6. D. A. B. Miller, Self-configuring universal linear optical component [Invited], *Photonics Research* **1**, 1 (2013).

7. S. Pai, B. Bartlett, O. Solgaard, D. A. B. Miller, Matrix Optimization on Universal Unitary Photonic Devices, *Physical Review Applied* **11**, 064044 (2019).

8. A. Annoni, E. Guglielmi, M. Carminati, G. Ferrari, M. Sampietro, D. A. Miller, A. Melloni, F. Morichetti, Unscrambling light - Automatically undoing strong mixing between modes, *Light: Science and Applications* **6** (2017).

9. W. Bogaerts, D. Pérez, J. Capmany, D. A. Miller, J. Poon, D. Englund, F. Morichetti, A. Melloni, Programmable photonic circuits, *Nature* **586**, 207–216 (2020).

10. J. Carolan, C. Harrold, C. Sparrow, E. Martín-López, N. J. Russell, J. W. Silverstone, P. J. Shadbolt, N. Matsuda, M. Oguma, M. Itoh, G. D. Marshall, M. G. Thompson, J. C. Matthews, T. Hashimoto, J. L. O'Brien, A. Laing, Universal linear optics, *Science* **349**, 711-716 (2015).

11. B. Bartlett, S. Fan, Universal programmable photonic architecture for quantum information processing, *Physical Review A* **101**, 042319 (2020).

12. M. Milanizadeh, S. M. Seyedin Navadeh, F. Zanetto, V. Grimaldi, C. De Vita, C. Klitis, M. Sorel, G. Ferrari, D. A. Miller, A. Melloni, F. Morichetti, Separating arbitrary free-space beams with an integrated photonic processor, *Light: Science & Applications 2022 11:1* **11**, 1–12 (2022).

13. N. C. Harris, J. Carolan, D. Bunandar, M. Prabhu, M. Hochberg, T. Baehr-Jones, M. L. Fanto, A. M. Smith, C. C. Tison, P. M. Alsing, D. Englund, Linear programmable nanophotonic processors, *Optica* **5**, 1623 (2018).

14. C. Ramey, Silicon Photonics for Artificial Intelligence Acceleration (Lightmatter), *IEEE Hot Chips 32 Symposium (HCS)* pp. 1–26 (2020).

15. I. A. D. Williamson, T. W. Hughes, M. Minkov, B. Bartlett, S. Pai, S. Fan, Reprogrammable Electro-Optic Nonlinear Activation Functions for Optical Neural Networks, *IEEE Journal of Selected Topics in Quantum Electronics* **26**, 1–12 (2020).

16. S. Bandyopadhyay, A. Sludds, S. Krastanov, R. Hamerly, N. Harris, D. Bunandar, M. Streshinsky, M. Hochberg, D. Englund, Single chip photonic deep neural network with accelerated training, *arXiv preprint* (2022).

17. L. Jing, Y. Shen, T. Dubcek, J. Peurifoy, S. Skirlo, Y. LeCun, M. Tegmark, M. Soljačić, *Proceedings of Machine Learning Research* (2017), pp. 1733–1741.

18. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, *Advances in neural information processing systems* (2017), pp. 5998–6008.

19. M. G. Anderson, S.-Y. Ma, T. Wang, L. G. Wright, P. L. McMahon, Optical Transformers, *arXiv preprint* (2023).

20. M. A. Nahmias, T. F. De Lima, A. N. Tait, H. T. Peng, B. J. Shastri, P. R. Prucnal, Photonic Multiply-Accumulate Operations for Neural Networks, *IEEE Journal of Selected Topics in Quantum Electronics* **26** (2020).

21. A. A. Cruz-Cabrera, M. Yang, G. Cui, E. C. Behrman, J. E. Steck, S. R. Skinner, Reinforcement and backpropagation training for an optical neural network using self-lensing effects, *IEEE Transactions on Neural Networks* **11**, 1450–1457 (2000).

22. T. W. Hughes, M. Minkov, Y. Shi, S. Fan, Training of photonic neural networks through in situ backpropagation and gradient measurement, *Optica* **5**, 864 (2018).

23. L. G. Wright, T. Onodera, M. M. Stein, T. Wang, D. T. Schachter, Z. Hu, P. L. McMahon, Deep physical neural networks trained with backpropagation, *Nature* **601**, 549–555 (2022).

24. J. Spall, X. Guo, X. Guo, X. Guo, A. I. Lvovsky, A. I. Lvovsky, A. I. Lvovsky, Hybrid training of optical neural networks, *Optica* **9**, 803–811 (2022).

25. J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, Q. Zhang, JAX: composable transformations of Python+NumPy programs [Computer software], https://github.com/google/jax (2022).

26. Tom Hennigan, Trevor Cai, Tamara Norman, Igor Babuschkin, Haiku: Sonnet for JAX, [Computer software] https://github.com/deepmind/dm-haiku (2020).

27. D. A. B. Miller, Analyzing and generating multimode optical fields using self-configuring networks, *Optica* **7**, 794 (2020).

28. M. Prabhu, C. Roques-Carmes, Y. Shen, N. Harris, L. Jing, J. Carolan, R. Hamerly, T. Baehr-Jones, M. Hochberg, V. Čeperić, J. D. Joannopoulos, D. R. Englund, M. Soljačić, Accelerating recurrent Ising machines in photonic integrated circuits, *Optica* **7**, 551 (2020).

29. D. A. B. Miller, Perfect optics with imperfect components, *Optica* **2**, 747 (2015).

30. H. Zhang, M. Gu, X. D. Jiang, J. Thompson, H. Cai, S. Paesani, R. Santagati, A. Laing, Y. Zhang, M. H. Yung, Y. Z. Shi, F. K. Muhammad, G. Q. Lo, X. S. Luo, B. Dong, D. L. Kwong, L. C. Kwek, A. Q. Liu, An optical neural chip for implementing complex-valued neural network, *Nature Communications* **12**, 1–11 (2021).

31. D. A. B. Miller, Setting up meshes of interferometers – reversed local light interference method, *Optics Express* **25**, 29233 (2017).

32. S. Pai, I. A. Williamson, T. W. Hughes, M. Minkov, O. Solgaard, S. Fan, D. A. Miller, Parallel Programming of an Arbitrary Feedforward Photonic Network, *IEEE Journal of Selected Topics in Quantum Electronics* **26** (2020).

33. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

34. D. P. Kingma, J. L. Ba, Adam: A Method for Stochastic Optimization, *International Conference on Learning Representations* (2015).

35. S. Pai, simphox: Another inverse design library, [Computer software] https://github.com/fancompute/simphox (2022).

36. J. K. Perin, M. Sharif, J. M. Kahn, Sensitivity Improvement in 100 Gb/s-per-Wavelength Links Using Semiconductor Optical Amplifiers or Avalanche Photodiodes, *Journal of Lightwave Technology* **34**, 5542–5553 (2016).

37. M. H. Taghavi, L. Belostotski, J. W. Haslett, P. Ahmadi, 10-Gb/s 0.13-μm CMOS Inductorless Modified-RGC Transimpedance Amplifier, *IEEE Transactions on Circuits and Systems I: Regular Papers* **62**, 1971–1980 (2015).

38. B. Sedighi, M. Khafaji, J. C. Scheytt, *2011 6th European Microwave Integrated Circuit Conference* (2011), pp. 192–195.

39. S. Buhr, J. Pliva, T. Schirmer, M. M. Khafaji, F. Ellinger, *Proceedings of the 17th European Microwave Integrated Circuits Conference* (Institute of Electrical and Electronics Engineers (IEEE), 2022), pp. 260–265.

40. D. A. Miller, Attojoule Optoelectronics for Low-Energy Information Processing and Communications, *Journal of Lightwave Technology* **35**, 346–396 (2017).

41. S. Pai, T. Park, M. Ball, B. Penkovsky, M. Milanizadeh, M. Dubrovsky, N. Abebe, F. Morichetti, A. Melloni, S. Fan, O. Solgaard, D. A. B. Miller, Experimental evaluation of digitally-verifiable photonic computing for blockchain and cryptocurrency, *arXiv preprint* (2022).

42. B. Murmann, Mixed-Signal Computing for Deep Neural Network Inference, *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* **29**, 3–13 (2021).

43. H. Zhang, J. Thompson, M. Gu, X. D. Jiang, H. Cai, P. Y. Liu, Y. Shi, Y. Zhang, M. F. Karim, G. Q. Lo, X. Luo, B. Dong, L. C. Kwek, A. Q. Liu, Efficient On-Chip Training of Optical Neural Networks Using Genetic Algorithm, *ACS Photonics* **8**, 1662–1672 (2021).

44. A. Nøkland, Direct Feedback Alignment Provides Learning in Deep Neural Networks, *Advances in Neural Information Processing Systems* pp. 1045–1053 (2016).

45. M. J. Filipovich, M. J. Filipovich, Z. Guo, M. Al-Qadasi, B. A. Marquez, H. D. Morison, V. J. Sorger, P. R. Prucnal, S. Shekhar, B. J. Shastri, B. J. Shastri, B. J. Shastri, Silicon photonic architecture for training deep neural networks with direct feedback alignment, *Optica* **9**, 1323–1332 (2022).

46. X. Guo, T. D. Barrett, Z. M. Wang, A. I. Lvovsky, Backpropagation through nonlinear units for the all-optical training of neural networks, *Photonics Research, Vol. 9, Issue 3, pp. B71-B80* **9**, B71-B80 (2021).

47. D. A. B. Miller, Self-aligning universal beam coupler, *Optics Express* **21**, 6360 (2013).

48. R. Tang, R. Tanomura, T. Tanemura, Y. Nakano, Ten-Port Unitary Optical Processor on a Silicon Photonic Chip, *ACS Photonics* **8**, 2074–2080 (2021).

49. D. Pérez, I. Gasulla, L. Crudgington, D. J. Thomson, A. Z. Khokhar, K. Li, W. Cao, G. Z. Mashanovich, J. Capmany, Multipurpose silicon photonics signal processor core, *Nature Communications* (2017).

50. S. Pai, N. Abebe, dphox: photonic layout and device design [Computer software], https://github.com/solgaardlab/dphox (2022).

51. S. Pai, Z. Sun, T. Park, phox: Base repository for simulation and control of photonic devices [Computer software], https://github.com/solgaardlab/phox/ (2022).

52. L. Deng, The MNIST database of handwritten digit images for machine learning research, *IEEE Signal Processing Magazine* **29**, 141–142 (2012).

53. W. R. Clements, P. C. Humphreys, B. J. Metcalf, W. S. Kolthammer, I. A. Walsmley, Optimal design for universal multiport interferometers, *Optica* **3**, 1460 (2016).

54. L. Biewald, Experiment tracking with weights and biases (2020). Software available from wandb.com.

55. S. Pai, Mnist onn: Weights and biases (2023). Software available from wandb.com.

56. S. Pai, Z. Sun, solgaardlab/photonicbackprop: Data and code for the paper "Experimentally realized in situ backpropagation for deep learning in energy-efficient nanophotonic neural networks", *Zenodo* (2022).

57. A. Dembo, T. Kailath, Model-Free Distributed Learning, *IEEE Transactions on Neural Networks* **1**, 58–70 (1990).

58. G. Cauwenberghs, A Fast Stochastic Error-Descent Algorithm for Supervised Learning and Optimization, *Advances in Neural Information Processing Systems 5* (1992).

59. J. Alspector, R. Meir, B. Yuhas, A. Jayakumar, D. Lippe, *Advances in Neural Information Processing Systems* (1992), pp. 836–844.

60. J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, Li Fei-Fei, *IEEE Conference on Computer Vision and Pattern Recognition* (Institute of Electrical and Electronics Engineers (IEEE), 2010), pp. 248–255.

61. S. Pai, S. Fan, O. Solgaard, D. A. Miller, Scalable and self-correcting photonic computation using balanced photonic binary tree cascades, *arXiv preprint* (2022).

62. J. Mower, N. C. Harris, G. R. Steinbrecher, Y. Lahini, D. Englund, High-fidelity quantum state evolution in imperfect photonic integrated circuits, *Physical Review A* **92**, 032322 (2015).

63. J. M. García, D. Pozo, S. Celma, M. T. Sanz, J. P. Alegre, CMOS tunable TIA for 1.25 Gbit/s optical gigabit Ethernet, *Electronics Letters* **43** (2007).

64. C. Wang, M. Zhang, B. Stern, M. Lipson, M. Lončar, Nanophotonic lithium niobate electro-optic modulators, *Optics Express* **26**, 1547 (2018).

65. C. Xiong, W. H. Pernice, J. H. Ngai, J. W. Reiner, D. Kumah, F. J. Walker, C. H. Ahn, H. X. Tang, Active silicon integrated nanophotonics: Ferroelectric BaTiO3 devices, *Nano Letters* **14**, 1419–1425 (2014).

66. R. Hamerly, L. Bernstein, A. Sludds, M. Soljačić, D. Englund, Large-Scale Optical Neural Networks Based on Photoelectric Multiplication, *Physical Review X* **9**, 021032 (2019).

67. M. Horowitz, 1.1 Computing's energy problem (and what we can do about it), *Digest of Technical Papers - IEEE International Solid-State Circuits Conference* **57**, 10–14 (2014).

68. P. O. Leisher, J. Thomas, J. Campbell, I. Gonzalez, D. Renner, L. Johansson, M. Mashanovitch, *2016 International Semiconductor Laser Conference (ISLC)* (2016), pp. 1–2.

69. N. C. Harris, Y. Ma, J. Mower, T. Baehr-Jones, D. Englund, M. Hochberg, C. Galland, Efficient, compact and low loss thermo-optic phase shifter in silicon, *Optics Express* **22**, 10487 (2014).

70. C. Errando-Herranz, A. Y. Takabayashi, P. Edinger, H. Sattari, K. B. Gylfason, N. Quack, MEMS for Photonic Integrated Circuits, *IEEE Journal of Selected Topics in Quantum Electronics* **26** (2020).

71. M. Wuttig, H. Bhaskaran, T. Taubner, Phase-change materials for non-volatile photonic applications, *Nature Photonics* **11**, 465–476 (2017).

72. P. Edinger, A. Y. Takabayashi, C. Errando-Herranz, U. Khan, H. Sattari, P. Verheyen, W. Bogaerts, N. Quack, K. B. Gylfason, Silicon photonic microelectromechanical phase shifters for scalable programmable photonics, *Optics Letters* **46**, 5671 (2021).

73. Y. Kang, H. D. Liu, M. Morse, M. J. Paniccia, M. Zadka, S. Litski, G. Sarid, A. Pauchard, Y. H. Kuo, H. W. Chen, W. S. Zaoui, J. E. Bowers, A. Beling, D. C. McIntosh, X. Zheng, J. C. Campbell, Monolithic germanium/silicon avalanche photodiodes with 340 GHz gain–bandwidth product, *Nature Photonics* **3**, 59–63 (2008).

74. A. Tsuchiya, A. Hiratsuka, K. Tanaka, H. Fukuyama, N. Miura, H. Nosaka, H. Onodera, A 45 Gb/s, 98 fJ/bit, 0.02 mm2 Transimpedance Amplifier with Peaking-Dedicated Inductor in 65-nm CMOS, *International System on Chip Conference* pp. 150–154 (2019).

75. **Acknowledgements:** We would like to acknowledge Advanced MicroFoundries (AMF) in Singapore for their help in fabricating and characterizing the photonic circuit for our demon-

stration and Silitronics for their help in packaging our chip for our demonstration. Thanks also to Payton Broaddus for helping with wafer dicing, Simon Lorenzo for help in fiber splicing the fiber switch for bidirectional operation, Joseph Kahn for guidance on avalanche photodetector noise estimates, Nagaraja Pai for advice on electronics, scalability and electrical and thermal control packaging, Ronald Quan for help in building our all-analog gradient measurement electronics, and finally Carsten Langrock and Karel Urbanek for their help in building our movable optical breadboard. **Funding:** We would also like to acknowledge funding from Air Force Office of Scientific Research (AFOSR) grants FA9550-17-1-0002 in collaboration with UT Austin and FA9550-18-1-0186 through which we share a close collaboration with UC Davis under Dr. Ben Yoo. **Contributions:** SP ran all experiments with input from ZS, TH, TP, BB, IW, NA, MM, OS, SF, DM. SP, TF, MM, and IW conceptualized the experimental protocol. SP, NA, FM, MM, and AM contributed to the design of the photonic mesh. SP and ZS wrote code to control the photonic integrated circuit active elements and camera detection and electronic circuit for analog gradient measurement. TP designed the custom PCB with input from SP. SP wrote the manuscript with input from all coauthors. All coauthors contributed to discussions of the protocol and results. **Competing interests:** SP, ZS, TH, IW, MM, SF, OS, DM have filed a patent for the analog backpropagation update protocol discussed in this work with Prov. Appl. No.: 63/323743. DM would like to disclose two related patents on the SVD architecture: US Patent #10,877,287 and #10,534,189. The authors declare no other conflicts of interest. **Data and materials:** Materials and methods are available as Supplementary Materials at the Science website. All other software and data for running the simulations and experiments are available through Zenodo (*56*) and Github through the Phox framework, including our experimental code via Phox (*51*), simulation code via Simphox (*35*), and circuit design code via Dphox (*50*).

Supplementary Materials:

Materials and Methods

Supplementary Text

Figs. S1-S8

Tables S1-S6

Refs. 44-66

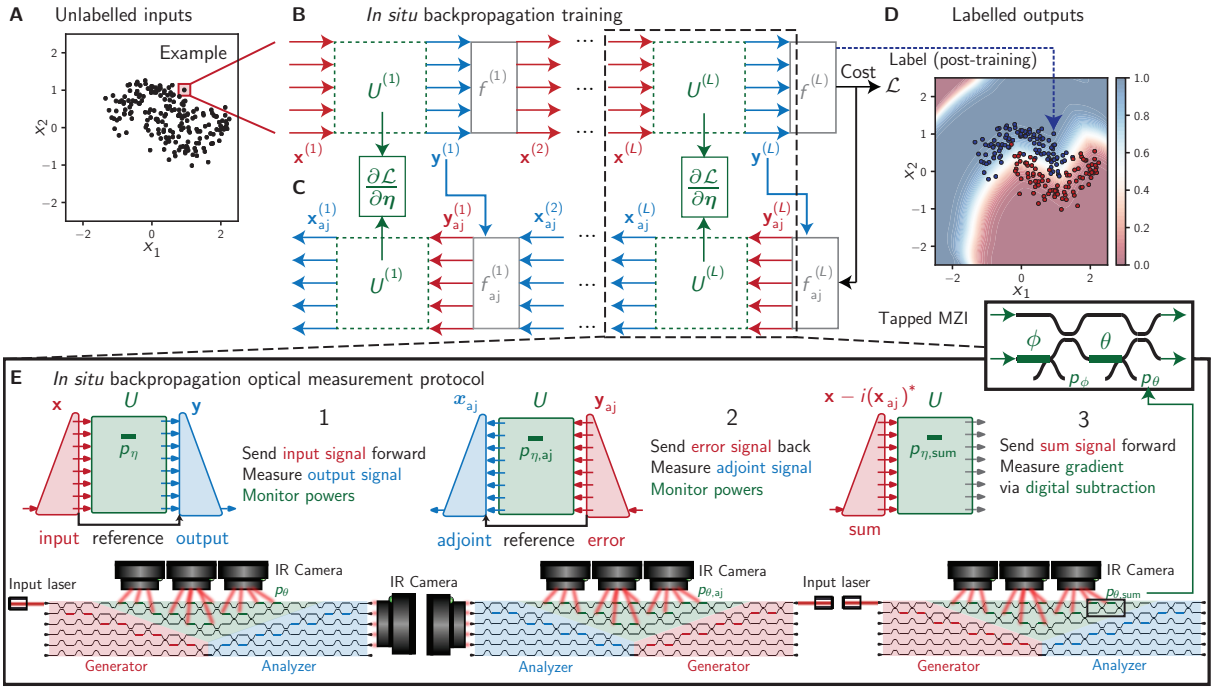

Figure 1: ***In situ* backpropagation concept**: (A) Example machine learning problem: an unlabelled 2D set of points that are formatted to be input into a PNN. (B) *In situ* backpropagation training of an L-layer PNN for the forward direction and (C) the backward direction showing the dependence of gradient updates for phase shifts on backpropagated errors. (D) An inference task implemented on the actual chip results in good agreement between the chip-labelled points and the ideal implemented ring classification boundary (resulting from the ideal model) and a 90% classification accuracy. (E) We show how our proposed scheme performs the three steps of *in situ* (analog) backpropagation, using a 6×6 mesh implementing coherent 4×4 bidirectional unitary matrix-vector products using a reference arm. We depict the (1) forward (2) backward (3) sum steps of *in situ* backpropagation. Arbitrary input setting and complete amplitude and phase output measurement are enabled in both directions using the reciprocity and symmetries of the triangular architecture. All powers throughout the mesh are monitored by an IR camera using the tapped MZI shown in the inset for each step, allowing for digital subtraction to compute the gradient (*22*). These power measurements performed at phase shifts are indicated by green horizontal bars.

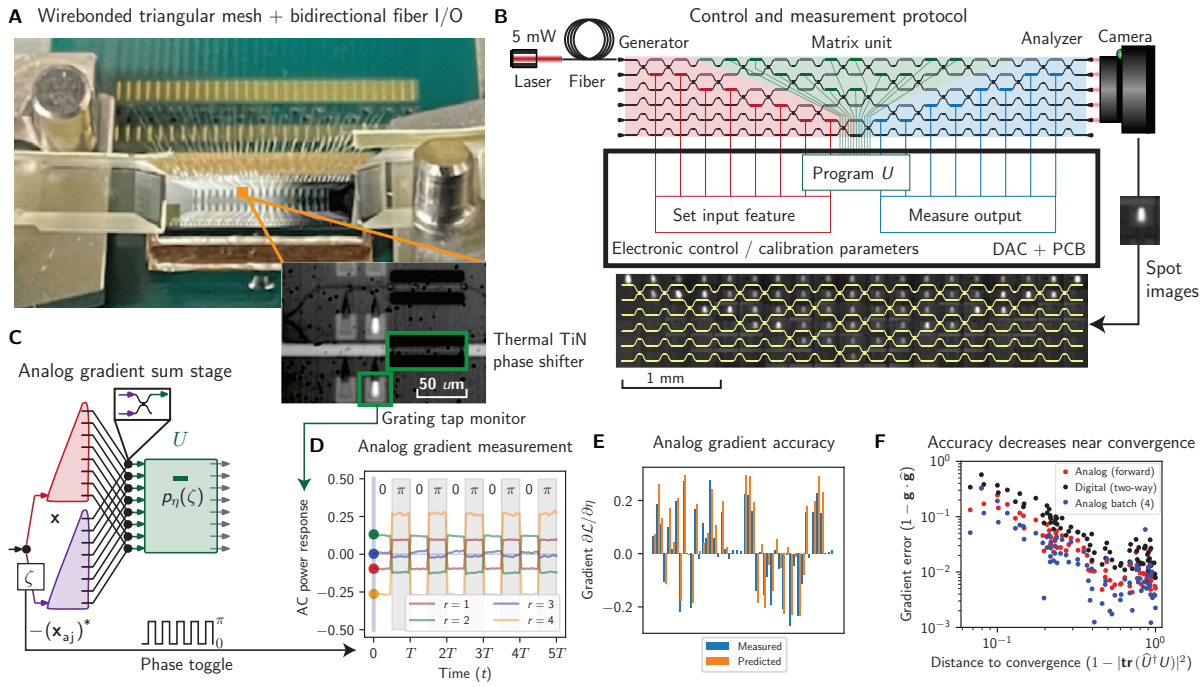

Figure 2: **Analog gradient experiment and simulation**: (A) Photonic chip was wirebonded to a custom PCB with fiber array for laser input/output and a camera overhead for imaging the chip. Zooming in reveals the core control-and-measurement unit of the chip, enabling power measurement using 3% grating tap monitors and a thermal TiN phase shifter nearby. (B) A calibrated control unit was used for input generation and output coherent detection. The IR camera over the chip imaged all grating tap monitors necessary for backpropagation. (C) Analog gradient update might optionally be implemented by introducing a summing interference circuit (not implemented on the chip in (B)) between the input and adjoint fields. (D) We toggled the adjoint phase between $\zeta = 0$ and π to evaluate the analog gradient measurement $\partial\mathcal{L}_i/\partial\eta$ for $i = 1$ to 4. (E) Gradients measured using the toggle scheme yielded approximately correct gradients when the implemented mesh was perturbed from the optimal (target) unitary given 1 rad phase standard deviation. (F) Measured normalized gradient error decreased with distance of the device implementation $\widehat{U}(\vec{\eta})$ from the optimal $U = \mathrm{DFT}(4)$, and analog batch and single-example gradients outperformed digital gradients.

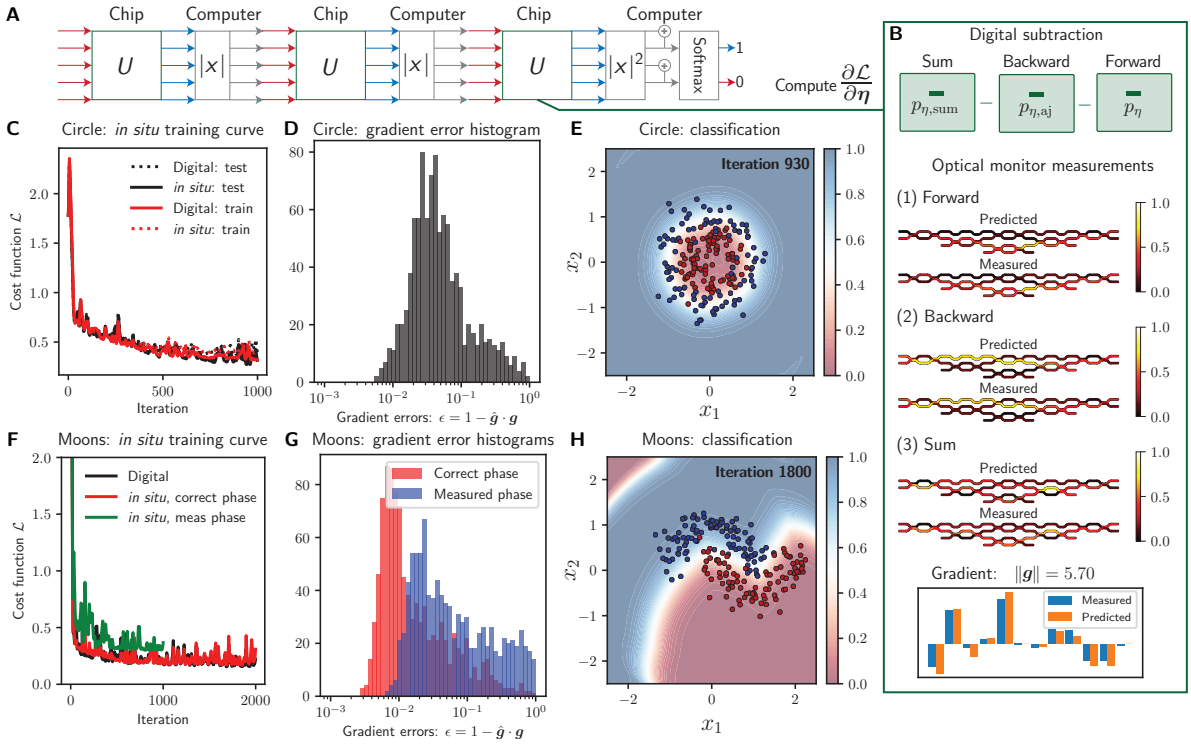

Figure 3: ***In situ* backpropagation experiment and simulation**: We performed *in situ* backpropagation training (*34*) for two classification tasks solvable by (A) a three layer hybrid PNN consisting of absolute value nonlinearities and a softmax (effectively sigmoid) decision layer. (B) Three-step digital subtraction gradient update given monitored waveguide powers and the measured gradient output. (C) For the circle dataset, the digital and *in situ* backpropagation training curves show excellent agreement resulting in (D) model accuracy of 96% test and 93% model (depicted here for iteration 930, showing the true labels and the learned classification model outcomes) and (E) histogram of low gradient error. (F) For the moons dataset, our phase measurements were sufficiently inaccurate to impact training leading to a lower model train accuracy of 87% (green). Using ground truth phase (red), we arrived at (G) sufficiently high model test accuracy 98% (train 95%) and (H) histogram of gradient errors improving considerably by roughly an order of magnitude using the correct phase measurement.

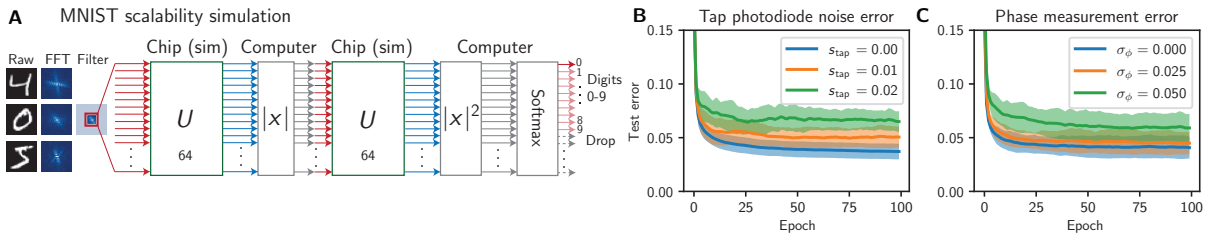

Figure 4: ***In situ* backpropagation simulation:** (A) We simulated a two-layer PNN on MNIST data using a previously explored PNN benchmark (*32*). (B, C) We aggregated marginal training curve statistics (shaded regions indicate standard deviation error range about the mean) over a grid search of 72 tap noise, loss, and I/O amplitude and phase errors. (B) tap noise factor s_{tap} (2.7% increase for $s_{\text{tap}} = 0.02$ from $3.7 \pm 0.7\%$ average error) and (C) phase error σ_ϕ (1.9% increase for $\sigma_\phi = 0.05$ from $4 \pm 1\%$ average error) affected training most.