

Research Article

Jiawei Zhang*, Weipeng Zhang, Tengji Xu, Lei Xu, Eli A. Doris, Bhavin J. Shastri, Chaoran Huang and Paul R. Prucnal

Online training and pruning of multi-wavelength photonic neural networks

<https://doi.org/10.1515/nanoph-2025-0296>

Received June 23, 2025; accepted October 28, 2025;

published online November 10, 2025

Abstract: CMOS-compatible photonic integrated circuits (PICs) are emerging as a promising platform in artificial intelligence (AI) computing. Owing to the compact footprint of microring resonators (MRRs) and the enhanced interconnect efficiency enabled by wavelength division multiplexing (WDM), MRR-based photonic neural networks (PNNs) are particularly promising for large-scale integration. However, the scalability and energy efficiency of such systems are fundamentally limited by the MRR resonance wavelength variations induced by fabrication process variations (FPVs) and environmental fluctuations. Existing solutions use post-fabrication approaches or thermo-optic tuning, incurring high control power and additional process complexity. In this work, we introduce an online training and pruning method that addresses this challenge, adapting to FPV-induced and thermally induced shifts in MRR resonance wavelength. By incorporating a power-aware pruning term into the conventional loss function, our approach simultaneously optimizes the PNN accuracy and the total power

consumption for MRR tuning. In proof-of-concept on-chip experiments on the Iris dataset, our system PNNs can adaptively train to maintain above 90 % classification accuracy in a wide temperature range of 26–40 °C while achieving a 44.7 % reduction in tuning power via pruning. Additionally, our approach reduces the power consumption by orders-of-magnitude on larger datasets. By addressing chip-to-chip variation and minimizing power requirements, our approach significantly improves the scalability and energy efficiency of MRR-based integrated analog photonic processors, paving the way for large-scale PICs to enable versatile applications including neural networks, photonic switching, LiDAR, and radio-frequency beamforming.

Keywords: microring resonators; photonic neural networks; resonance variations; online training; pruning

1 Introduction

Large neural networks (NNs) have demonstrated exceptional performance in edge computing [1], natural language processing [2], and autonomous systems [3]. CMOS-compatible silicon photonic integrated circuits (PICs) are emerging as a promising platform in artificial intelligence (AI) computing [4], [5], offering significant advantages including low latency, high bandwidth, and fully programmability [5]–[7]. Integrated photonic neural networks (PNNs) generally fall into two major categories: coherent PNNs based on interferometric meshes (e.g., Mach–Zehnder interferometers [MZIs]) [4], [6] and wavelength-division multiplexed (WDM) PNNs based on wavelength-selective filters (e.g., microring resonators [MRRs]) [8], [9]. Owing to the compact footprint of MRRs and the enhanced interconnect efficiency enabled by WDM, MRR-based PNNs can be implemented with significantly less chip area than their coherent equivalents [10], and are promising for large-scale integration using CMOS-compatible silicon photonic foundry processes [11]–[13]. In addition to NN inference [5], [14], such MRR-based integrated analog photonic processors have also found important applications in photonic switching [15],

Jiawei Zhang, Weipeng Zhang and Tengji Xu contributed equally to this work.

***Corresponding author: Jiawei Zhang**, Department of Electrical and Computer Engineering, Princeton University, Princeton, 08544, NJ, USA, E-mail: jiawei.zhang@princeton.edu, <https://orcid.org/0000-0001-8682-1763>

Weipeng Zhang, Lei Xu, Eli A. Doris and Paul R. Prucnal, Department of Electrical and Computer Engineering, Princeton University, Princeton, 08544, NJ, USA, E-mail: weipengz@princeton.edu (W. Zhang), leixu@princeton.edu (L. Xu), edoris@princeton.edu (E.A. Doris), prucnal@princeton.edu (P.R. Prucnal)

Tengji Xu and Chaoran Huang, Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China, E-mail: tengjixu@link.cuhk.edu.hk (T. Xu), crhuang@ee.cuhk.edu.hk (C. Huang)

Bhavin J. Shastri, Department of Physics, Engineering Physics and Astronomy, Queen's University, Kingston, K7L 3N6, ON, Canada, E-mail: shastri@ieee.org, <https://orcid.org/0000-0001-5040-8248>

LiDAR [16], [17], RF beamforming [18], [19], and data interconnects [20], [21].

However, a key challenge in realizing large-scale MRR-based analog photonic processors is the functional variation of MRRs caused by unavoidable fabrication process variations (FPVs) and dynamic environmental fluctuations (e.g., thermal crosstalk and polarization drifts), which can induce significant random shifts in the MRR resonance wavelengths. This shift can be expressed as [22]:

$$\delta\lambda_0 = \frac{\delta^{(\text{env})}n_{\text{eff}} \cdot \lambda_0}{n_g}, \quad (1)$$

where λ_0 is the MRR resonance wavelength, $\delta^{(\text{env})}n_{\text{eff}}$ is the effective index shift due to environmental changes, and n_g is the group index accounting for waveguide dispersion. The effective index shift can be further expressed as [23]

$$\delta^{(\text{env})}n_{\text{eff}} = c \int \Delta\epsilon \mathbf{E}_v \cdot \mathbf{E}_v^* dx dy, \quad (2)$$

where c is the speed of light, $\Delta\epsilon(x, y)$ denotes a local change in the dielectric constant, and $\mathbf{E}_v(x, y)$ is the normalized modal electric field vector of the waveguide mode. All the static and dynamic variations – including sub-wavelength FPVs in geometric parameters, ambient temperature disturbances, and thermal crosstalk – can alter $\Delta\epsilon(x, y)$, resulting in a resonance shift $\delta\lambda_0$ comparable to the free spectral range (FSR) of MRRs [24]. For example, in a recent study [25], measurements of 371 identically designed racetrack-shaped resonators, revealed resonance shifts ranging from 1.76 nm (median) to 6 nm (maximum), as a result of inherent silicon thickness variations (± 5 nm fluctuations in 220 nm layers) across wafers and fabrication batches. It is also shown that fluctuations in ambient temperature can lead to a drift in the MRR resonance wavelength of tens of pm, resulting in a degradation of the accuracy of an MRR-based photonic neural network (PNN) to 67 % from 99 % for a two-layer MNIST classification [26]. While previous work has explored various post-fabrication strategies to counteract these variations – including germanium (Ge) ion implantation [27], [28], integration of phase change materials (PCMs) [29]–[32], and deposition of photochromic materials [33], [34] – these techniques are only effective at correcting static FPVs, and they require additional post-fabrication processing complexity and precise control over the materials involved. Alternatively, thermo-optic tuning remains widely used method due to its broad tuning range, but it is power-intensive; consuming 28 mW/FSR using an embedded N-doped heater [35], [36]. Consequently, the scalability

and energy efficiency of MRR-based analog photonic processors are fundamentally constrained, hindering their applications to large-scale networks such as large AI models and WDM transceivers.

To address fundamental limitation, we propose online training and pruning in MRR-based PNNs that adapt to FPV-induced and thermally induced shifts in MRR resonance wavelength. Based on perturbation-based gradient descent algorithm, we develop an online training framework that maps the trainable NN parameters to MRR-based PNN chips without the need of look-up tables (LUTs). We further incorporate a power-aware pruning term into the conventional loss function, which simultaneously optimizes the PNN accuracy and the total power consumption for MRR tuning. In proof-of-concept on-chip experiments, we demonstrate online training with an iterative feedback system with a PIC performing fast analog matrix-vector multiplications (MVMs), combined with a central processing unit (CPU) digitally computing gradients at a slower timescale. Using a 3×2 PNN on the Iris dataset, our system PNNs can adaptively train to maintain above 90 % classification accuracy across static (FPV) or dynamic (thermal drifts) variations, while achieving a 44.7 % reduction in tuning power via pruning. Additionally, simulations with larger and deeper convolutional neural networks (CNNs) on standard datasets – including MNIST [37], CIFAR-10 [38], and Fashion-MNIST [39] – validate the scalability of our method, showing orders-of-magnitude reductions in power consumption. By addressing chip-to-chip variation and minimizing power requirements, our approach significantly improves the scalability and energy efficiency of MRR-based integrated analog photonic processors, paving the way for large-scale PICs to enable versatile applications including NNs, photonic switching, LiDAR, and RF beamforming.

2 Results

2.1 Concept and principle

2.1.1 Offline training

While PNNs can operate with picosecond latency for real-time applications [40]–[42], they rely on slower digital computers (i.e. CPU or GPU) for training, a process called offline training [40], [43]. As shown in Figure 1a, in this approach, NN parameters (e.g., neuron weights and biases) are optimized in software using backpropagation (BP) based gradient descent algorithm:

$$\mathbf{w}^{k+1} \leftarrow \mathbf{w}^k - \alpha \nabla_{\mathbf{w}^k} \mathcal{L}, \quad (3)$$

where $\mathbf{w}^k$ denotes the weight matrix in the k th training epoch, α is the learning rate, $\mathcal{L}$ is the loss function. The parameters are then mapped onto the physical hardware (PNN chips) using highly accurate calibration models (also called look-up table (LUT)). For the MRR in the i th row, the j th column of an $M \times N$ MRR weight bank, it writes as

$$\mathbf{I}_{ij} = f_{ij}(\mathbf{w}_{ij}) = f_j(\mathbf{w}_{ij}). \quad (4)$$

Here $\mathbf{w}_{ij}$ denotes the weight to be executed on the i , j th MRR, and $\mathbf{I}_{ij}$ is the required tuning current. The LUT, denoted as a mapping function f_{ij} , is determined by the relative position of the j th laser wavelength and the MRR resonant state. Ideally, the MRRs operating at the same laser wavelength are expected to exhibit identical resonant wavelengths, simplifying the LUT to f_j .

However, in practice, the MRRs in the same column may exhibit significantly different resonant wavelengths due to static and dynamic variations. As derived in Supplementary Note 1, an LUT for the i , j th MRR – accounting for FPVs, ambient temperature variations, input optical power, and the self-heating effect – is approximately given by

$$\mathbf{I}_{ij} = \tilde{f}_{ij}(\mathbf{w}_{ij}, \sigma_{ij}, T, P_{\text{opt}}), \quad (5)$$

where σ_{ij} denotes the deviation of resonant wavelength due to FPVs, T is the time-varying ambient temperature, P_{opt} is the input optical power. Based on Eq. (5), it remains complicated to generate an LUT that accounts for all the non-idealities across large-scale MRR weight banks. Any inaccuracies in LUTs will directly introduce arbitrary deviations in the mapped weight parameters, leading to significant performance degradation across NN system benchmarks such as classification tasks.

In an MRR-based weight bank using thermo-optic tuning to counteract these non-idealities, the required electrical power for actively programming weights of the i , j th MRR can be written as [24]:

$$P_{ij}^{\text{weight}} = \mathbf{I}_{ij}^2 R = P_{ij}^{\text{lock}} + P_{ij}^{\text{conf}}, \quad (6)$$

where R is the resistance of the metal heater. The power breaks down into a static weight locking power that locks the MRR resonance to the desired wavelength, and a configuration power to program the weight value [24]. Therefore, the total power for the weight configuration consumed by an $M \times N$ MRR weight bank is

$$\mathbf{P}^{\text{weight}} = \sum_{i,j} P_{ij}^{\text{weight}}. \quad (7)$$

For simplicity, we denote $\mathbf{P}^{\text{weight}}$ as $\mathbf{P}$ in the following sections except Section 2.3.

2.1.2 Online training

Online training, also known as “*in situ*” or “chip-in-the-loop” training, was originally proposed to alleviate the intensive use required of CPU/GPUs for training non Von-Neumann architecture based computing hardware [44], where the training occurs on the same hardware used for inference. It was also recognized that online training can mitigate hardware nonidealities, and promise for iterative weight updates in real-time [45]. The state-of-the-art online training algorithms for PNNs [6], [45]–[48] can be classified into two major categories: gradient-based and gradient-free algorithms. Gradient-free algorithms, such as genetic algorithms [47], are straightforward to implement in practice but often face inherent challenges with convergence and scalability. In contrast, gradient-based algorithms are typically more efficient and, therefore, more widely adopted. The most commonly used gradient-based training algorithm for software-based NNs, backpropagation, analytically computes gradients by back-propagating errors using the chain rule [37]. While significant progress has been made in experimentally realizing BP on photonic hardware [48]–[50], the process typically requires global optical power monitoring and evaluation of nonlinear activation function gradients in software. This approach introduces additional system complexity and latency overhead. Other gradient-based algorithms, such as direct feedback alignment [45], replace the chain rule in back-propagation with a random weight matrix, but their validation has been limited to small, shallow NNs.

In our approach, the online training of our MRR-based PNNs is implemented by a perturbation-based gradient descent algorithm, which estimates gradients only based on forward inference running on PNN chips. Here, the NN parameters to be optimized are set to be MRR tuning currents instead of neuron weights:

$$\mathbf{I}^{k+1} \leftarrow \mathbf{I}^k - \alpha \nabla_{\mathbf{I}^k} \mathcal{L}. \quad (8)$$

The gradient of loss function with respect to MRR tuning currents ($\mathbf{I}^k$) is approximately given by perturbation measurement:

$$\nabla_{\mathbf{I}^k} \mathcal{L} \approx \frac{\Delta \mathcal{L}}{\Delta \mathbf{I}^k}, \quad (9)$$

where $\Delta \mathbf{I}^k$ is the perturbation rate of the MRR tuning currents in the k th training epoch, and $\Delta \mathcal{L}$ is the measured change of loss function induced by the perturbation. In practice, $\Delta \mathbf{I}^k$ should be sufficiently small to preserve the validity of approximation in Eq. (9), while remaining large enough to produce measurable $\Delta \mathcal{L}$ above experimental noise.

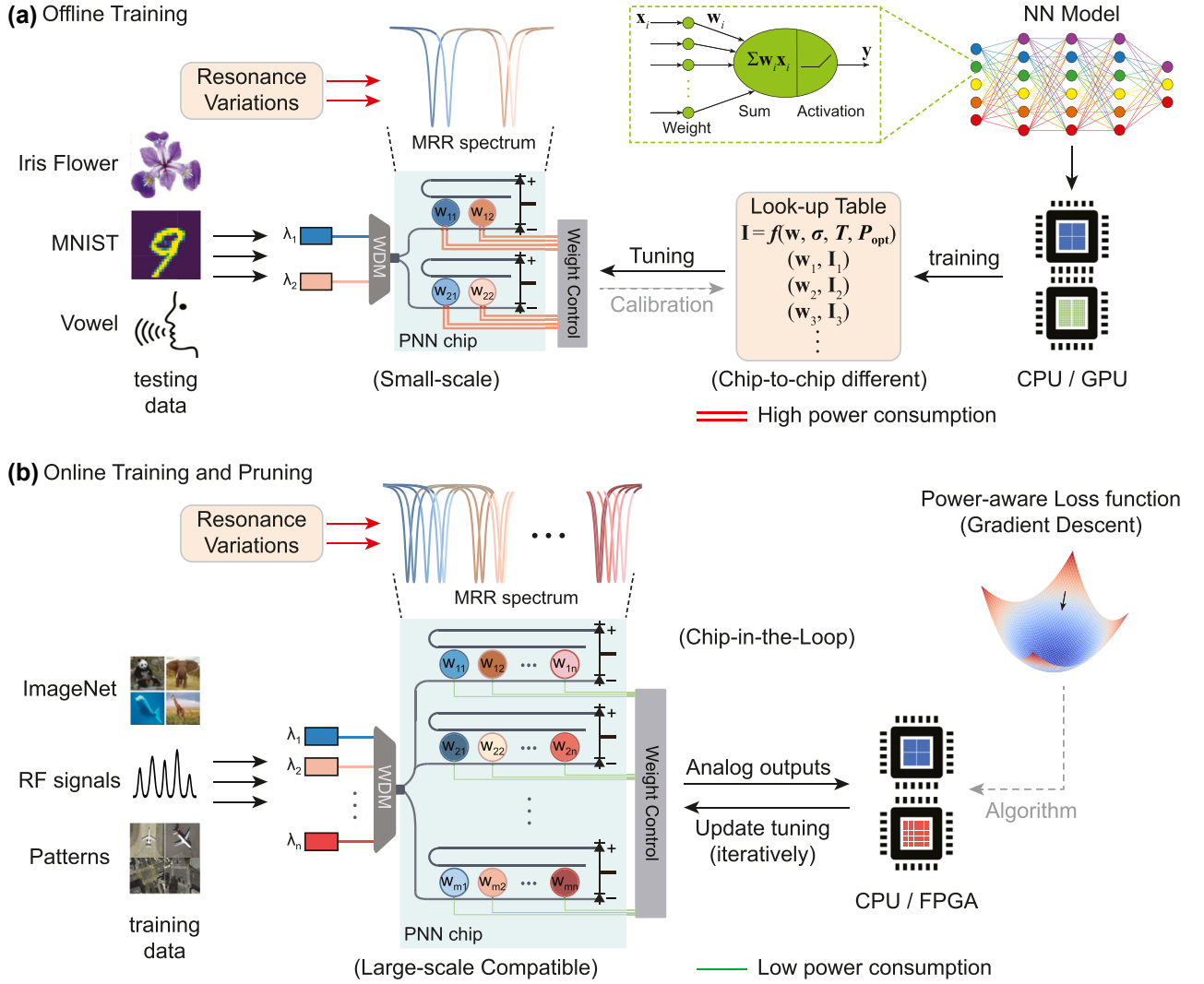


Figure 1: Comparison of offline training with online training and pruning approach. (a) Conventional offline training method supporting only small-scale PNN chips. In this approach, NN parameters are calculated in software and mapped onto the PNN chips. The resonance variations of MRRs necessitate complicated look-up tables, and lead to higher power consumption for MRR control. (b) Online training and pruning method compatible for large-scale PNN chips. The training of PNNs occurs on the same chip used for inference, accounting for any chip-to-chip variations and environmental fluctuations. Our approach simultaneously optimizes the PNN accuracy and the total power consumption for tuning all the MRRs.

2.1.3 Pruning

In software-based NNs, pruning is a model compression technique that removes redundant weight parameters (by setting their values to zero) whilst maintaining accuracy. While the implementation of pruning has also been investigated in PNNs, prior approaches either lack robustness against nonidealities [51], [52] or require extensive offline training [26]. As shown in Figure 1b, our proposed approach accounts for the total MRR tuning power in the “chip-in-the-loop” process, by incorporating a power-aware “pruning” term into the conventional loss function:

$$\tilde{\mathcal{L}} = \mathcal{L} + \gamma \mathbf{P}, \quad (10)$$

where $\tilde{\mathcal{L}}$ is the modified loss function given by the sum of the conventional loss function $\mathcal{L}$ and the power-aware pruning term $\gamma \mathbf{P}$, γ is an empirically determined hyperparameter that defines pruning strength. According to Eqs. (6) and (7), $\tilde{\mathcal{L}}$ can also be expressed as

$$\tilde{\mathcal{L}} = \mathcal{L} + \gamma \sum_{i,j} P_{ij} = \mathcal{L}(\mathbf{I}) + \gamma \sum_{i,j} \mathbf{I}_{ij}^2 R. \quad (11)$$

The selection of the hyperparameter γ is critical in this framework, as it explicitly parameterizes the trade-off between the NN accuracy and energy efficiency. Depending

on the demand of the end users, minimizing this modified loss function can enable concurrent optimization of both the NN accuracy and the total power consumption. For example, in a power-constrained environment, the training problem becomes

$$\min \mathcal{L} + \gamma \mathbf{P} \quad \text{s.t. } \mathbf{P} \leq \mathbf{P}_{\max} \quad (12)$$

with a relatively larger value of γ , where $\mathbf{P}_{\max}$ denotes the maximum power available for weight configurations. Oppositely, in scenarios prioritizing high accuracies, it becomes

$$\min \mathcal{L} + \gamma \mathbf{P} \quad \text{s.t. } \mathcal{L} \leq \mathcal{L}_{\max} \quad (13)$$

with a relatively smaller value of γ , where $\mathcal{L}_{\max}$ denotes the upper bound of the conventional loss function permitted under a given accuracy constraint.

2.2 Experiment

2.2.1 Setup

Recently, integrated PNNs have been extensively investigated and experimentally validated across various on-chip scales, achieving low latencies of hundreds of picoseconds at diverse applications [6], [41], [53]. Notable demonstrations include a 3-layer PNN with 9 neurons for image classification [53], a 3-layer 6×6 PNN for vowel classification task [6], and a single-layer 4×2 PNN for fiber nonlinearity compensation [40]. In our proof-of-concept experiment, a 3×2 MRR weight bank is used to demonstrate our proposed online training and pruning approach and the associated energy savings.

Our experimental setup is illustrated in Figure 2a. First, three channels of input data (denoted as $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$) are generated from a high sampling rate signal generator (Keysight M8196A), and modulated onto laser 1, 2, and 3 (Pure-Photonics, PPCL500) respectively via Mach-Zehnder modulators (MZMs). The lights from these three lasers, each at a distinct wavelength, are combined using a wavelength-division multiplexer (MUX) and then split equally between two MRR weight banks. Each MRR weight bank consists of three MRRs, which are designed with slightly different radii (8 μm , 8.012 μm , 8.024 μm). The MRRs at the same column (i.e., MRR1 and MRR4, MRR2 and MRR5, MRR3 and MRR6) are expected to share the same corresponding resonance wavelengths, aligned with 200 GHz spaced ITU grids (1,546.92 nm, 1,548.51 nm, 1,550.12 nm, respectively). However, as shown in Figure 2b, the measured weight bank spectrum indicates that the resonance wavelengths of all six MRRs deviate from the designed values due to FPVs and temperature changes. To counteract FPVs, the wavelengths of three lasers are further tuned (1,547.24 nm, 1,548.60 nm, and 1,550.40 nm)

to align with the deviated resonance wavelengths. Moreover, all MRRs are thermally tuned via embedded N-doped heaters to actively program and configure weights (denoted as $\mathbf{w}_{11}, \mathbf{w}_{12}, \mathbf{w}_{13}, \mathbf{w}_{21}, \mathbf{w}_{22}, \mathbf{w}_{23}$), allowing for individual weighting of input analog data in the three wavelength channels. The thermal tuning characteristics of six MRRs at 30 °C, are illustrated in Figure 2c.

Our proof-of-concept experiment is validated on a standard Iris dataset, where three species of Iris flowers (Setosa, Versicolour, Virginica) are classified using only three out of four input features (sepal length, petal length, and petal width), as shown in Figure 2d. This three-class classification problem is converted into a two-step binary classification (Figure 2e) mapped to the on-chip PNN, which utilizes eight trainable parameters. The parameters include the tuning currents of six MRRs – $\mathbf{I}_{11}, \mathbf{I}_{12}, \mathbf{I}_{13}, \mathbf{I}_{21}, \mathbf{I}_{22}, \mathbf{I}_{23}$ (corresponding to six weights) – and two biases, b_1 and b_2 . The 150 samples are split into 120 for training and 30 for testing.

The analog optical signals from the drop and through port are then captured and differentiated by two balanced photodetectors (BPDs), which gives the electrical output of the weighted summations, $\sum_i \mathbf{w}_{1i} \mathbf{x}_i$ and $\sum_i \mathbf{w}_{2i} \mathbf{x}_i$. The output is further read by an oscilloscope and demodulated in a CPU, which evaluates the power-aware loss function (Eq. (10)) and calculates gradients based on perturbation (Eq. (9)). Finally, the CPU updates the MRR tuning currents for the next training epoch (Eq. (8)) and commands the MRR driver, which is equipped on a customized printed circuit board (PCB) (Supplementary Note 2). As shown in Figure 2f, the training of our PNN occurs with the photonic chip in the loop, iteratively optimizing the loss function and reducing the total power consumption.

2.2.2 Demonstration on 3×2 PNN

PNN training is conducted under three conditions: conventional offline training, online training without pruning, and online training with pruning. Each training configuration is repeated 10 times to ensure consistency. The perturbation rate of MRR tuning currents ($\Delta \mathbf{I}^k$) is set to 0.05 mA. To evaluate the gradients with respect to the MRR tuning currents ($\nabla_{\mathbf{I}^k} \mathcal{L}$), each training epoch batches the training data seven times through the PNN: six times perturbing each tuning current ($\mathbf{I}_{ij}^k + \Delta \mathbf{I}_{ij}^k$) and one time to measure $\tilde{\mathcal{L}}$ with non-perturbed currents $\mathbf{I}_{ij}^k$. The plots in Figure 3a show the average of both categorical cross-entropy loss ($\mathcal{L}$) and power-aware loss ($\tilde{\mathcal{L}}$) versus training epoch. It is observed that the PNN can quickly converge to the optimal weights within 25 epochs using perturbation-based online training.

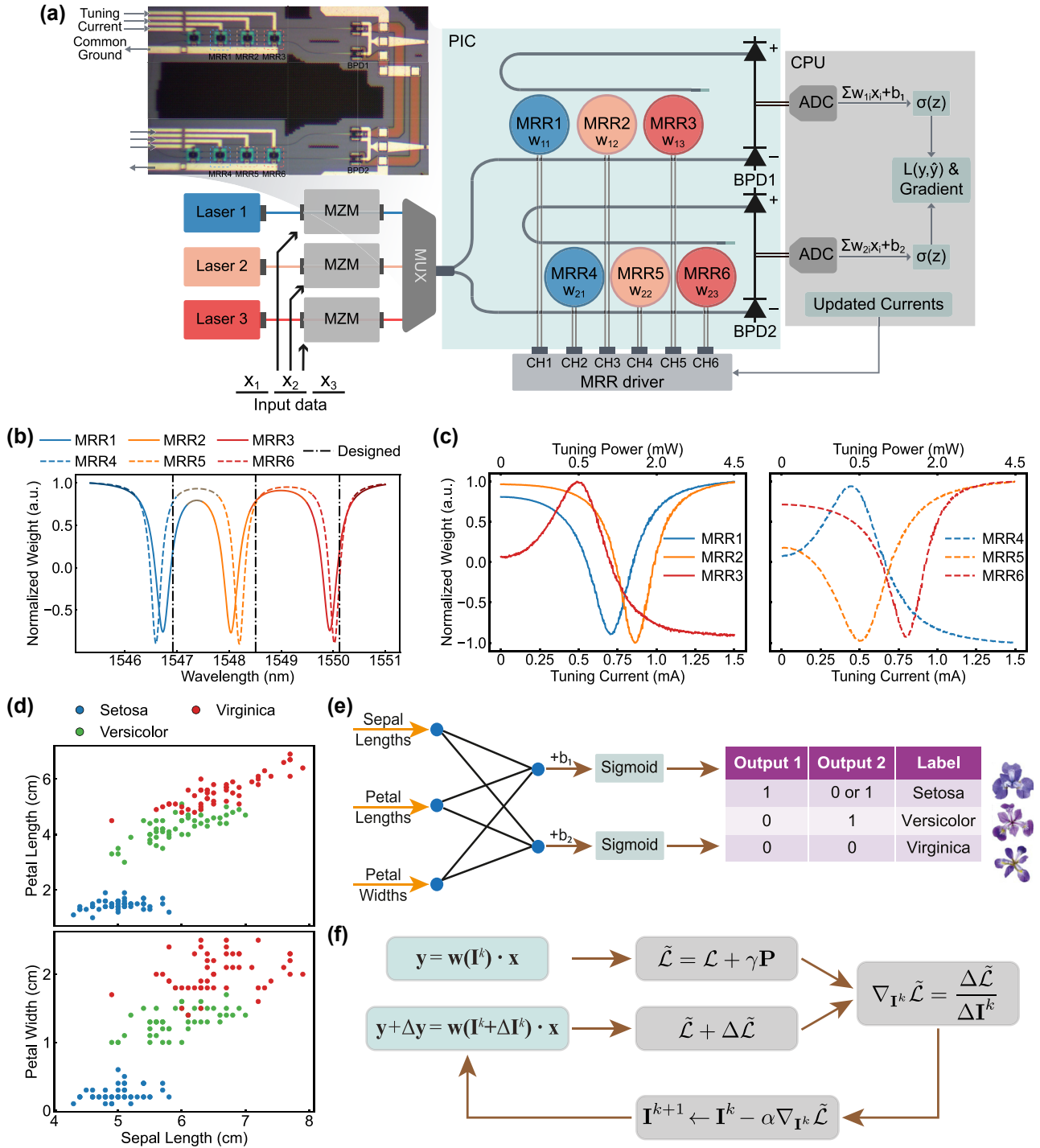


Figure 2: System setup of online training and pruning. (a) Schematic of our experimental setup. MZM, Mach-Zehnder modulator. MUX, wavelength multiplexer. PIC, photonic integrated circuit. MRR, microring resonator. BPD, balanced photodetector. ADC, analog-to-digital converter. CPU, central processing unit. $\sigma(z)$ denotes the nonlinear activation function performed on software. The inset shows the micrograph of the two MRR weight banks and BPDs. (b) Normalized weight spectrum of two MRR weight banks at 30 °C. The MRRs at the same colors (i.e., MRR1 and MRR4, MRR2 and MRR5, MRR3 and MRR6) are designed to share the same resonance wavelengths, aligned with 200 GHz spaced ITU grids (1,546.92 nm, 1,548.51 nm, 1,550.12 nm). (c) Tuning characteristics of six MRRs at 30 °C. (d) Scatterplot of Iris flower dataset. (e) Schematic of a simple 3×2 neural network for Iris classification. (f) Online training and pruning procedure. At each iteration, the PIC performs matrix-vector multiplication without and with perturbations, while the CPU evaluates the gradients of power-aware loss function, and updates the MRR tuning currents for the next iteration.

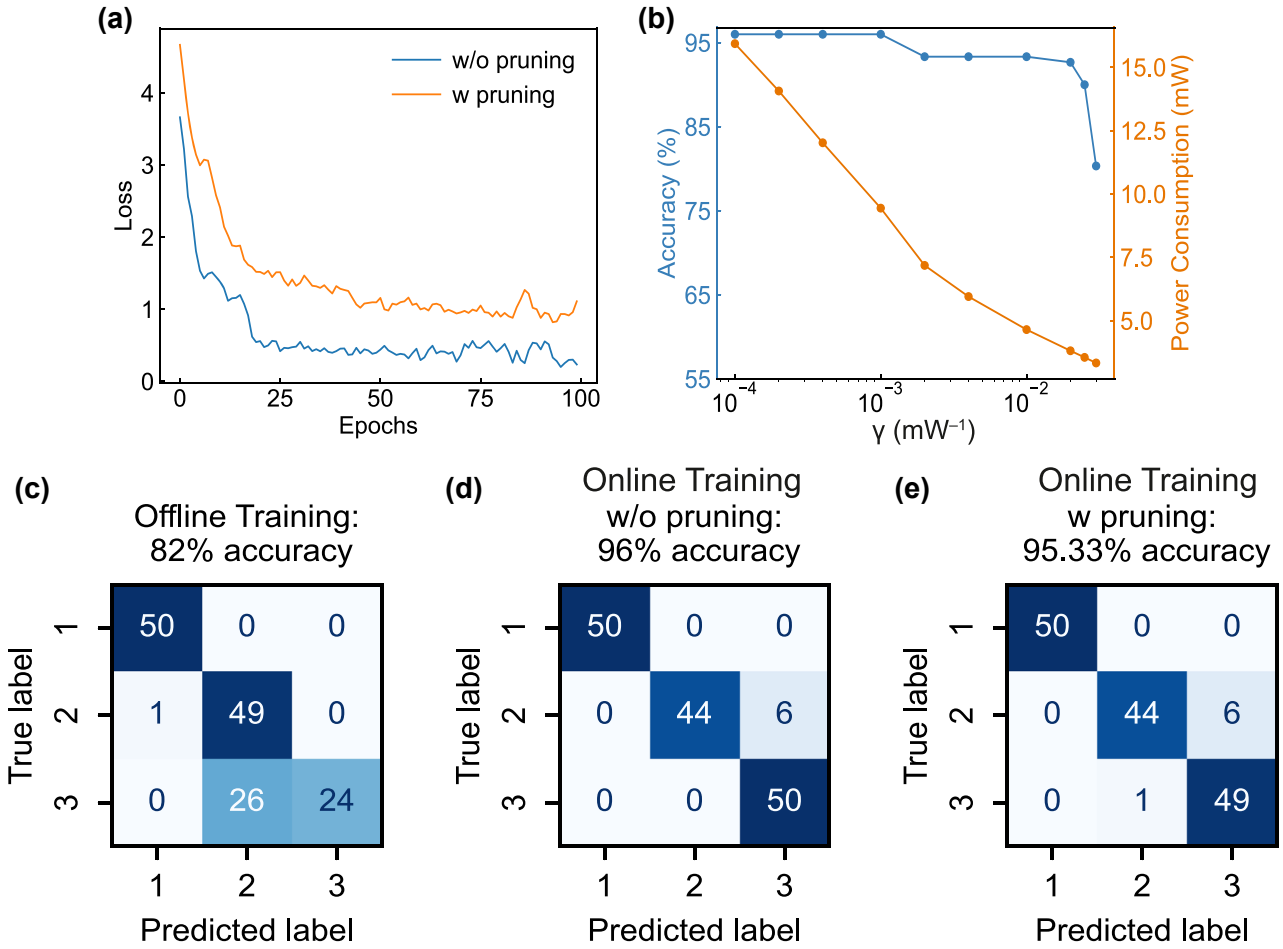


Figure 3: Online training results without, or with pruning method. (a) Experimental results of online training losses without, or with the pruning method. (b) Simulated result indicating the tradeoff between the prediction accuracy and power efficiency. (c–e) Confusion matrices for the 150 samples, obtained by the conventional offline training, online training without, or with the pruning method, respectively.

We also simulate the trade-off between the NN accuracy and the associated energy savings (Figure 3b). The simulation indicates that the trade-off is optimized at $\gamma_{\text{opt}} = 0.01 \text{ mW}^{-1}$, where significant reduction of overall tuning power (exceeding 70 %) can be achieved before the accuracy drops off. As predicted in Section 2.1.3, in scenarios prioritizing high accuracies where $\gamma < \gamma_{\text{opt}}$, the loss function $\tilde{\mathcal{L}}$ is dominated by the conventional loss function $\mathcal{L}$, such that the PNN maintains a high accuracy of >93 %. Contrarily, in scenarios constraining overall power consumption where $\gamma > \gamma_{\text{opt}}$, the power-aware loss function $\tilde{\mathcal{L}}$ is dominated by the pruning term $\gamma \mathbf{P}$, leading to a larger amount of energy savings but may degrade the PNN accuracy. As shown by the confusion matrix in Figure 3c, the offline training experiment only produces 82 % accuracy on the 150 samples. In the accuracy-prioritized online training experiment (without pruning, γ set to 0) (Figure 3d), the overall classification of the Iris flower task is improved to 96 %, with a total MRR

tuning power of 9.54 mW. In the online training experiment constraining power consumption (with pruning, γ set to 0.0075 mW^{-1}) (Figure 3e), we observe a 44.7 % reduction (from 9.54 mW to 5.28 mW) of total tuning power while the PNN maintains a classification accuracy of 95.33 %.

Furthermore, to demonstrate the adaptability of our online training method to temperature drifts, additional experiments are conducted at multiple different temperatures ranging from 26 °C to 40 °C. Before online training, the MRR tuning currents are randomly initialized within the tuning range of 0–1.5 mA (only weights $\mathbf{w}_{11}$ and $\mathbf{w}_{21}$ are plotted, Figure 4a), producing untrained classification outputs with only 33.33 % accuracy. After online training for 100 epochs (Figure 4b), the PNN can be trained to the optimal weights, regardless of the changes of MRR tuning characteristics due to temperature drifts. As a result, high classification accuracies is consistently achieved at both 30 °C (99.33 %, upper row of Figure 4a and b) and 34 °C (96.67 %, lower row of Figure 4a and b).

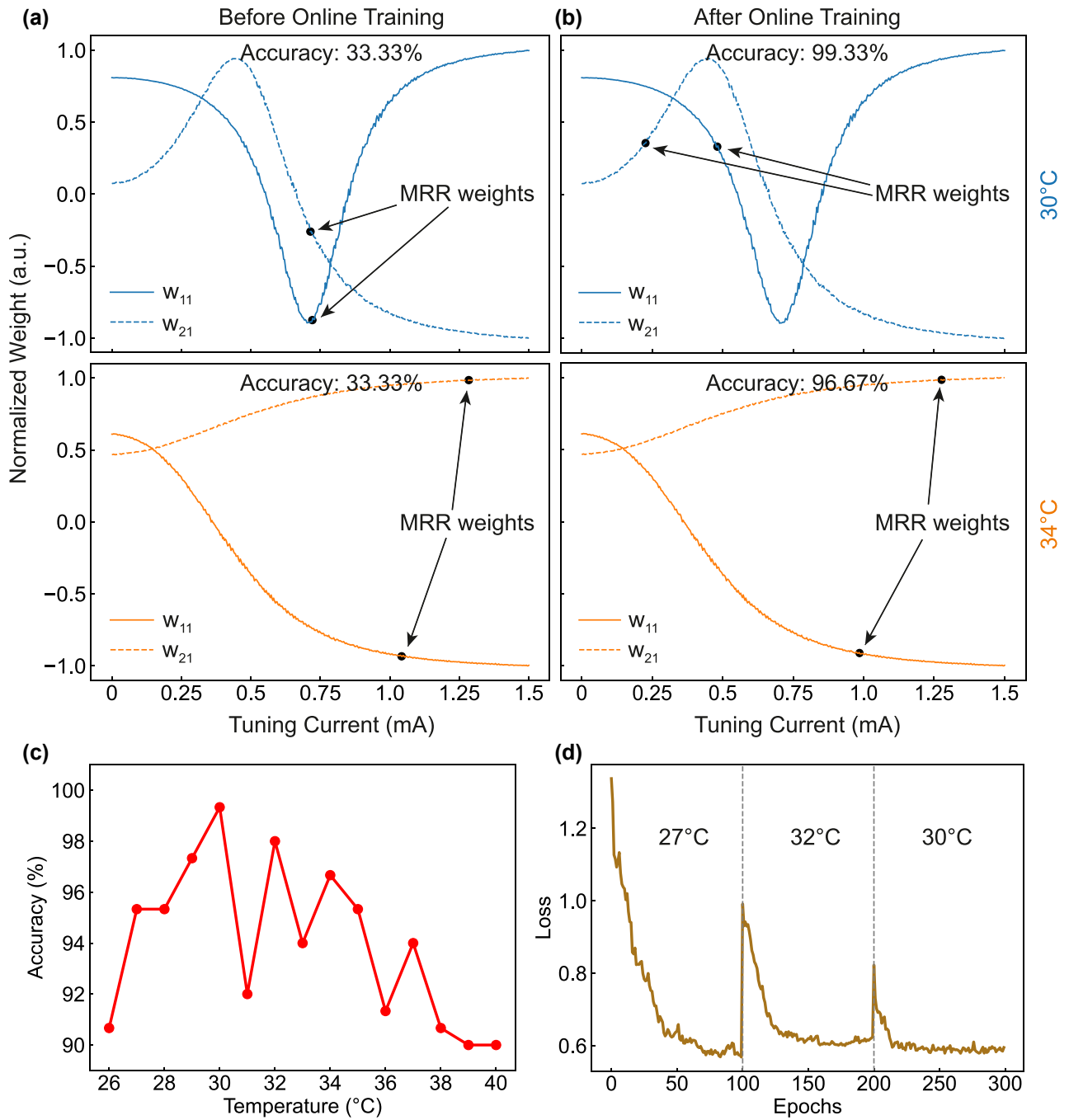


Figure 4: Adaptive PNN training at different temperatures. (a, b) Illustrate two of the six MRR tuning currents (weights I_{11} and I_{21}) before and after online training, respectively, at 30 °C (the upper plots) and 34 °C (the lower plots). (c) Experimental obtained Iris classification accuracies at different temperatures. (d) Online training experiment performed with thermal disturbances. The temperature is changed from 27 °C to 32 °C at the 100th epoch, and from 32 °C to 30 °C at the 200th epoch.

lower row of Figure 4a and b). As shown in Figure 4c, the classification accuracies maintain above 90 % even if the external temperature changes from 26 °C to 40 °C. We also evaluate the resilience of our PNN training to intentional thermal disturbances. In Figure 4d, it is shown that the PNN can quickly recover its classification performance within

about 25 epochs even if the external temperature is disturbed from 27 °C to 32 °C (at the 100th epoch), and from 32 °C to 30 °C (at the 200th epoch). These proof-of-concept experimental results further confirm that our online training and pruning method can handle any PNN hardware non-idealities, including chip-to-chip FPGAs, temperature

disturbances, and even nonlinearities induced by input optical power.

2.3 Energy savings in large-scale NNs via pruning

In addition to the MRR tuning power for configuring weights, the overall power consumed by a multiwavelength PNN also includes the power needed for laser pumping and O/E/O data conversions (e.g., modulators, photodetectors, and analog-to-digital converters [ADCs]). For a multiwavelength PNN with N neurons and N^2 MRRs, the overall power can be expressed by [24]:

$$\mathbf{P}^{\text{total}} = N^2 \times \mathbf{P}^{\text{weight}} + N \times \mathbf{P}^{\text{laser}} + N \times BW \times \mathbf{E}^{\text{OEO}}, \quad (14)$$

where $\mathbf{P}^{\text{laser}}$ is the power needed for laser pumps, BW represents the bandwidth of the signal modulated on the optical carriers, and $\mathbf{E}^{\text{OEO}} = \mathbf{E}^{\text{mod}} + \mathbf{E}^{\text{det}} + \mathbf{E}^{\text{ADC}}$ is the energy of O/E/O data conversion, associated with modulation, detection, and ADCs. The MRR tuning power $\mathbf{P}^{\text{weight}}$ is shown to scale quadratically with the number of neurons, whereas the power use of laser pumps and data conversions scales linearly. While various technologies have been demonstrated to address different power contributors [33], [54], [55], we specifically look into MRR tuning power dominated scenarios and the energy savings in larger-scale multiwavelength PNNs via online pruning.

We extend the simulations to larger, deeper feedforward and CNN architectures for various classification tasks with standard datasets, including scikit-learn Moons, scikit-learn Wine dataset, MNIST, CIFAR-10, and Fashion-MNIST. The FVPs in the MRRs are simulated by inducing Gaussian-distributed variations (with a standard deviation of σ) to their resonance wavelengths (Supplementary Note 4) [24]. We quantify the impact of online pruning on inference power consumption, defined as the control power for all the weighting MRRs. Consistent with the theoretical framework in Section 2.1.3, the accuracy-power trade-off for large-scale NNs is governed by the pruning strength γ , as exemplified by the Iris dataset results in Figure 3b. As shown in Figure 5, the baseline unpruned PNNs (blue line) exhibit a linear increase in total MRR tuning power with the PNN scale (number of MRRs), prioritizing accuracy under unlimited power budgets ($\gamma = 0$). In contrast, for power-constrained systems tolerating specific classification errors, our method reduces power consumption by orders of magnitude. For instance, in the simulation using the scikit-learn Wine dataset (featuring a three-layer feedforward NN with 2,976 weights), our method achieves total MRR tuning

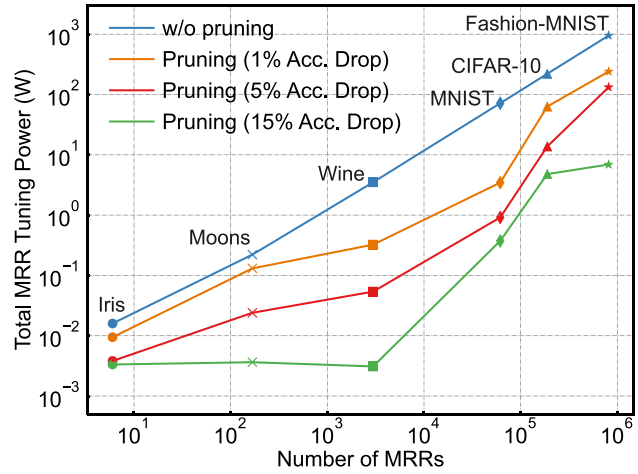


Figure 5: Simulated results on reducing overall power consumption enabled by online pruning across different size MRR-based PNNs.

power reductions of 10.8 $\times$, 65.2 $\times$, and 1,133.4 $\times$ under validation accuracy drops of 1 %, 5 %, and 15 %, respectively. It is also observed that the total MRR tuning power with 15 % accuracy drop remains almost constant across small-scale datasets (the number of MRRs is below 10^4), primarily resulting from the redundancies of fully-connected feedforward NNs compared to CNNs (Supplementary Note 4.3). The simulation results support that the scale of MRR-based PNNs that can be trained with our online training and pruning method can be up to one million MRR weights, far exceeding the current fabrication capability and available channels in 100 GHz spaced ITU grids, which is up to 72 different wavelengths. Nevertheless, our findings significantly improve the scalability and energy efficiency of MRR-based PNNs, particularly paving the way for their operations in power-constrained scenarios such as edge computing, wearable devices, and autonomous driving systems.

3 Discussion

With the emergence of new AI computing paradigms, such as brain-inspired neuromorphic computing and quantum computing, online training has been proposed that revolutionizes the definitions of unconventional hardware computing architecture and AI training algorithms. In this work, we focus on online training and pruning for integrated silicon photonic neural networks, which among the new computing paradigms appear attractive for their full-programmability and CMOS-compatibility. Our proposed approach provides a methodology that breaks one fundamental limiting factor on scalability due to chip-to-chip

variations, and simultaneously optimizes overall power consumption in thermally controlled MRR-based PICs.

Recent advances in photonic AI hardware systems demonstrate features that go beyond the performance of traditional digital electronics. In diffractive free-space optical systems, PNNs at the million-neuron scale have been realized [56], with efficient online training for compensating alignment errors and device non-idealities [57], [58]. In coherent PNNs based on MZIs, a single-chip, fully-integrated PNN with six neurons and three cascaded layers achieves end-to-end (410 ps) processing with online training improving the data throughput [6]. Additionally, a system-on-chip microwave photonic processor using MRR weight banks is developed to solve dynamic radio-frequency (RF) interference [41], showing real-time adaptability enabled by online learning and weight adjustments. We envision that the online training and pruning method presented in this study is also generalizable to other PNNs being actively investigated. For example, the generalized loss function ($\mathcal{L}$) writes as

$$\mathcal{L} = \mathcal{L} + \sum_i \gamma_i \Theta_i, \quad (15)$$

where $\mathcal{L}$ is given by the sum of the conventional loss function and all the parameter-aware pruning term $\gamma_i \Theta_i$, Θ_i denotes any hardware parameter related terms to be optimized, such as power consumption, DC electrical control currents/voltages, and modulator biases.

Taking advantage of low-latency photonic processing, online PNN training can benefit applications that necessitate real-time adaptability, including RF interference cancellation [41], fiber nonlinearity compensation [40], and edge computing [1]. In our experiment on Iris dataset, the total average time for running each training epoch is about 13.1 s, including all the latencies induced by the inter-device communications, MRR weight actuating, and digital calculations (the breakdown latency analysis is presented in Supplementary Note 5). The system training latency is currently dominated by the slow single-thread CPU processing and the communication time between equipments (for example, it takes an average of 0.726 s to download the analog waveforms from the oscilloscope to the CPU). This can be significantly reduced to millisecond level by speeding up with high-speed analog-to-digital converters and high-parallelism FPGA processors [41]. Further improvements of the system training latency include the exploration of more efficient training algorithms [6], [59] to reduce the perturbation runs. The exploitation of using faster MRR modulators [60], [61] for rapid weight updates can also potentially reduce the system latency from millisecond to microsecond level, despite that the proposed pruning technique remains critical for reducing the power consumption caused by thermal tuning.

A comprehensive comparative analysis of our work with other literatures on the online training of PNNs, together with the positioning of our pruning method in different tuning mechanisms used in integrated silicon photonics, can be found in Supplementary Note 6.

4 Conclusions

To summarize, we have proposed and demonstrated an online training and pruning method on multi-wavelength PNNs with MRR weight banks that addresses the fundamental issue on scalability and energy efficiency due to MRR resonance variations. We experimentally validate our training framework with an iterative feedback system, and show that superior performances of PNNs can be attained without any software-based pre-training involved. By incorporating the power-aware pruning term into the conventional loss function, our approach significantly optimizes overall power consumption in thermally controlled MRR-based PNNs. This study serves as a fundamental methodology for addressing the chip-to-chip variations in PICs, and represents a significant milestone towards building large-scale, energy-efficient MRR-based integrated analog photonic processors for versatile applications including NNs, LiDAR, RF beamforming, and data interconnects.

Acknowledgments: This research was supported by the Office of Naval Research (ONR) (N00014-22-1-2527 P.R.P.), NEC Laboratories America (Princeton E-filiates Partnership administered by the Andlinger Center for Energy and the Environment), Princeton's Eric and Wendy Schmidt Transformative Technology Fund, and NJ Health Foundation Award. The devices were fabricated at the Advanced Micro Foundry (AMF) in Singapore through the support of CMC Microsystems. BJS acknowledges support from the Natural Sciences and Engineering Research Council of Canada (NSERC).

Research funding: Office of Naval Research (ONR) (N00014-22-1-2527 P.R.P.).

Author contributions: JZ, WZ, and TX conceived the ideas. JZ performed the simulation, and designed the experiment with support from WZ. JZ and WZ developed the experimental photonic setup, including the co-integration of DAC control board and the associated control software. JZ conducted the experimental measurements and analyzed the results with support from WZ and EAD. JZ wrote the manuscript with support from LX, EAD, BJS, and CH. PRP supervised the research and contributed to the vision and execution of the experiment. All the authors contributed to the manuscript.

All authors have accepted responsibility for the entire content of this manuscript and consented to its submission to the journal, reviewed all the results and approved the final version of the manuscript.

Conflict of interest: JZ, WZ, LX, EAD, and PRP are inventors on a provisional patent application related to this work filed by Princeton University (63/836,043).

Data availability: All data and codes used in this study are available from the corresponding author upon reasonable request.

References

- [1] J. Chen and X. Ran, “Deep learning with edge computing: a review,” *Proc. IEEE*, vol. 107, no. 8, pp. 1655–1674, 2019.
- [2] L. Ouyang *et al.*, “Training language models to follow instructions with human feedback,” in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 27730–27744.
- [3] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [4] Y. Shen *et al.*, “Deep learning with coherent nanophotonic circuits,” *Nat. Photonics*, vol. 11, no. 7, pp. 441–446, 2017.
- [5] B. J. Shastri *et al.*, “Photonics for artificial intelligence and neuromorphic computing,” *Nat. Photonics*, vol. 15, no. 2, pp. 102–114, 2021.
- [6] S. Bandyopadhyay *et al.*, “Single-chip photonic deep neural network with forward-only training,” *Nat. Photonics*, vol. 18, no. 12, pp. 1335–1343, 2024.
- [7] Y. Chen *et al.*, “All-analog photoelectronic chip for high-speed vision tasks,” *Nature*, vol. 623, no. 7985, pp. 48–57, 2023.
- [8] A. N. Tait *et al.*, “Neuromorphic photonic networks using silicon photonic weight banks,” *Sci. Rep.*, vol. 7, no. 1, p. 7430, 2017.
- [9] J. Feldmann *et al.*, “Parallel convolutional processing using an integrated photonic tensor core,” *Nature*, vol. 589, no. 7840, pp. 52–58, 2021.
- [10] J. C. Lederman *et al.*, “Low-latency passive thermal desensitization of a silicon micro-ring resonator with self-heating,” *APL Photonics*, vol. 9, no. 7, p. 076117, 2024.
- [11] S. Y. Siew *et al.*, “Review of silicon photonics technology and platform development,” *J. Lightwave Technol.*, vol. 39, no. 13, pp. 4374–4389, 2021.
- [12] K. Giewont *et al.*, “300-mm monolithic silicon photonics foundry technology,” *IEEE J. Sel. Top. Quantum Electron.*, vol. 25, no. 5, p. 8200611, 2019.
- [13] L. Chrostowski *et al.*, “Silicon photonic circuit design using rapid prototyping foundry process design kits,” *IEEE J. Sel. Top. Quantum Electron.*, vol. 25, no. 5, p. 8201326, 2019.
- [14] C. Huang *et al.*, “Prospects and applications of photonic neural networks,” *Adv. Phys. X*, vol. 7, no. 1, p. 1981155, 2022.
- [15] H. Jayatilaka, H. Shoman, L. Chrostowski, and S. Shekhar, “Photoconductive heaters enable control of large-scale silicon photonic ring resonator circuits,” *Optica*, vol. 6, no. 1, pp. 84–91, 2019.
- [16] H. Larocque *et al.*, “Beam steering with ultracompact and low-power silicon resonator phase shifters,” *Opt. Express*, vol. 27, no. 24, pp. 34639–34654, 2019.
- [17] L. Huang *et al.*, “Integrated light sources based on micro-ring resonators for chip-based LiDAR,” *Laser Photonics Rev.*, vol. 19, no. 2, p. 2400343, 2025.
- [18] M. Nichols, H. Morison, A. Eshaghi, B. Shastri, and L. Lampe, “Photonics-based fully-connected hybrid beamforming: towards scaling up massive MIMO at higher frequency bands,” *Res. Sq.*, 2024. <https://doi.org/10.21203/rs.3.rs-4627859/v1>.
- [19] Y. Liu *et al.*, “Ultra-low-loss silicon nitride optical beamforming network for wideband wireless applications,” *IEEE J. Sel. Top. Quantum Electron.*, vol. 24, no. 4, p. 8300410, 2018.
- [20] A. Rizzo *et al.*, “Massively scalable Kerr comb-driven silicon photonic link,” *Nat. Photonics*, vol. 17, no. 9, pp. 781–790, 2023.
- [21] S. Daudlin, *et al.*, “Three-dimensional photonic integration for ultra-low-energy, high-bandwidth interchip data links,” *Nat. Photonics*, vol. 19, no. 5, pp. 502–509, 2025.
- [22] W. Bogaerts *et al.*, “Silicon microring resonators,” *Laser Photonics Rev.*, vol. 6, no. 1, pp. 47–73, 2012.
- [23] A. Densmore *et al.*, “A silicon-on-insulator photonic wire based evanescent field sensor,” *IEEE Photonics Technol. Lett.*, vol. 18, no. 23, pp. 2520–2522, 2006.
- [24] A. N. Tait, “Quantifying power in silicon photonic neural networks,” *Phys. Rev. Appl.*, vol. 17, no. 5, p. 054029, 2022.
- [25] L. Chrostowski, X. Wang, J. Flueckiger, Y. Wu, Y. Wang, and S. T. Fard, “Impact of fabrication non-uniformity on chip-scale silicon photonic integrated circuits,” in *Optical Fiber Communication Conference*, 2014, pp. 2–37.
- [26] T. Xu *et al.*, “Control-free and efficient integrated photonic neural networks via hardware-aware training and pruning,” *Optica*, vol. 11, no. 8, pp. 1039–1049, 2024.
- [27] M. M. Milosevic *et al.*, “Ion implantation in silicon for trimming the operating wavelength of ring resonators,” *IEEE J. Sel. Top. Quantum Electron.*, vol. 24, no. 4, p. 8200107, 2018.
- [28] B. Chen *et al.*, “Real-time monitoring and gradient feedback enable accurate trimming of ion-implanted silicon photonic devices,” *Opt. Express*, vol. 26, no. 19, pp. 24953–24963, 2018.
- [29] C. Ríos *et al.*, “Integrated all-photonic non-volatile multi-level memory,” *Nat. Photonics*, vol. 9, no. 11, pp. 725–732, 2015.
- [30] Z. Fang *et al.*, “Ultra-low-energy programmable non-volatile silicon photonics based on phase-change materials with graphene heaters,” *Nat. Nanotechnol.*, vol. 17, no. 8, pp. 842–848, 2022.
- [31] J. Zheng *et al.*, “GST-on-silicon hybrid nanophotonic integrated circuits: a non-volatile quasi-continuously reprogrammable platform,” *Opt. Mater. Express*, vol. 8, no. 6, pp. 1551–1561, 2018.
- [32] X. Yang *et al.*, “Non-volatile optical switch element enabled by low-loss phase change material,” *Adv. Funct. Mater.*, vol. 33, no. 42, p. 2304601, 2023.
- [33] S. Bilodeau *et al.*, “All-optical organic photochemical integrated nanophotonic memory: low-loss, continuously tunable, non-volatile,” *Optica*, vol. 11, no. 9, pp. 1242–1249, 2024.
- [34] L. Xu *et al.*, “Building scalable silicon microring resonator-based neuromorphic photonic circuits using post-fabrication processing with photochromic material,” *Adv. Opt. Mater.*, vol. 13, no. 11, p. 2402706, 2025.
- [35] H. Jayatilaka *et al.*, “Wavelength tuning and stabilization of microring-based filters using silicon in-resonator photoconductive heaters,” *Opt. Express*, vol. 23, no. 19, pp. 25084–25097, 2015.
- [36] C. Huang *et al.*, “Demonstration of scalable microring weight bank control for large-scale photonic integrated circuits,” *APL Photonics*, vol. 5, no. 4, p. 040803, 2020.

- [37] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [38] A. Krizhevsky and G. Hinton, “Learning multiple layers of features from tiny images,” *Tech. Rep.*, 2009. Available at: <https://www.cs.utoronto.ca/~tnqx223C/kriz/learning-features-2009-TR.pdf>.
- [39] H. Xiao, K. Rasul, and R. Vollgraf, “Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms,” arXiv preprint arXiv:1708.07747, 2017.
- [40] C. Huang *et al.*, “A silicon photonic–electronic neural network for fibre nonlinearity compensation,” *Nat. Electron.*, vol. 4, no. 11, pp. 837–844, 2021.
- [41] W. Zhang *et al.*, “A system-on-chip microwave photonic processor solves dynamic RF interference in real time with picosecond latency,” *Light: Sci. Appl.*, vol. 13, no. 1, p. 14, 2024.
- [42] X. Zhou, D. Yi, D. W. U. Chan, and H. K. Tsang, “Silicon photonics for high-speed communications and photonic signal processing,” *npj Nanophoton.*, vol. 1, no. 1, p. 27, 2024.
- [43] X. Lin *et al.*, “All-optical machine learning using diffractive deep neural networks,” *Science*, vol. 361, no. 6406, pp. 1004–1008, 2018.
- [44] S. M. Buckley, A. N. Tait, A. N. McCaughan, and B. J. Shastri, “Photonic online learning: a perspective,” *Nanophotonics*, vol. 12, no. 5, pp. 833–845, 2023.
- [45] M. J. Filipovich *et al.*, “Silicon photonic architecture for training deep neural networks with direct feedback alignment,” *Optica*, vol. 9, no. 12, pp. 1323–1332, 2022.
- [46] T. W. Hughes, M. Minkov, Y. Shi, and S. Fan, “Training of photonic neural networks through in situ backpropagation and gradient measurement,” *Optica*, vol. 5, no. 7, pp. 864–871, 2018.
- [47] H. Zhang *et al.*, “Efficient on-chip training of optical neural networks using genetic algorithm,” *ACS Photonics*, vol. 8, no. 6, pp. 1662–1672, 2021.
- [48] S. Pai *et al.*, “Experimentally realized in situ backpropagation for deep learning in photonic neural networks,” *Science*, vol. 380, no. 6643, pp. 398–404, 2023.
- [49] M. Hermans, M. Burm, T. Van Vaerenbergh, J. Dambre, and P. Bienstman, “Trainable hardware for dynamical computing using error backpropagation through physical media,” *Nat. Commun.*, vol. 6, no. 1, p. 6729, 2015.
- [50] Z. Gu, Z. Huang, Y. Gao, and X. Liu, “Training optronic convolutional neural networks on an optical system through backpropagation algorithms,” *Opt. Express*, vol. 30, no. 11, pp. 19416–19440, 2022.
- [51] S. Banerjee, M. Nikdast, S. Pasricha, and K. Chakrabarty, “Pruning coherent integrated photonic neural networks,” *IEEE J. Sel. Top. Quantum Electron.*, vol. 29, no. 2, p. 6101013, 2023.
- [52] J. Gu, C. Feng, Z. Zhao, Z. Ying, R. T. Chen, and D. Z. Pan, “Efficient on-chip learning for optical neural networks through power-aware sparse zeroth-order optimization,” *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 9, pp. 7583–7591, 2021.
- [53] F. Ashtiani, A. J. Geers, and F. Aflatouni, “An on-chip photonic deep neural network for image classification,” *Nature*, vol. 606, no. 7914, pp. 501–506, 2022.
- [54] E. Timurdogan, C. M. Sorace-Agaskar, J. Sun, E. Shah Hosseini, A. Biberman, and M. R. Watts, “An ultralow power athermal silicon modulator,” *Nat. Commun.*, vol. 5, no. 1, pp. 1–11, 2014.
- [55] Y. Xiang, H. Cao, C. Liu, J. Guo, and D. Dai, “High-speed waveguide Ge/Si avalanche photodiode with a gain-bandwidth product of 615 GHz,” *Optica*, vol. 9, no. 7, pp. 762–769, 2022.
- [56] Z. Xu, T. Zhou, M. Ma, C. Deng, Q. Dai, and L. Fang, “Large-scale photonic chiplet Taichi empowers 160-TOPS/W artificial general intelligence,” *Science*, vol. 384, no. 6692, pp. 202–209, 2024.
- [57] Z. Xue, T. Zhou, Z. Xu, S. Yu, Q. Dai, and L. Fang, “Fully forward mode training for optical neural networks,” *Nature*, vol. 632, no. 8024, pp. 280–286, 2024.
- [58] T. Zhou *et al.*, “Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit,” *Nat. Photonics*, vol. 15, no. 5, pp. 367–373, 2021.
- [59] A. N. McCaughan, B. G. Oripov, N. Ganesh, S. W. Nam, A. Dienstfrey, and S. M. Buckley, “Multiplexed gradient descent: fast online training of modern datasets on hardware neural networks without backpropagation,” *APL Mach. Learn.*, vol. 1, no. 2, p. 026118, 2023.
- [60] D. W. Chan, X. Wu, Z. Zhang, C. Lu, A. P. T. Lau, and H. K. Tsang, “Ultra-wide free-spectral-range silicon microring modulator for high capacity WDM,” *J. Lightwave Technol.*, vol. 40, no. 24, pp. 7848–7855, 2022.
- [61] Y. Yuan *et al.*, “A 5 × 200 Gbps microring modulator silicon chip empowered by two-segment Z-shape junctions,” *Nat. Commun.*, vol. 15, no. 1, p. 918, 2024.

Supplementary Material: This article contains supplementary material (<https://doi.org/10.1515/nanoph-2025-0296>).