nature photonics

Article

https://doi.org/10.1038/s41566-023-01205-0

Lithography-free reconfigurable integrated photonic processor

Received: 14 November 2022

Accepted: 29 March 2023

Published online: 27 April 2023

Check for updates

Tianwei Wu [1], Marco Menarini[2], Zihe Gao [1] & Liang Feng [1,2]

Integrated photonics, because of its intrinsic high speed, large bandwidth and unlimited parallelism, is critical in the drive to ease the increasing data traffic. Its technological enabler is high-precision lithography, which allows for the fabrication of high-resolution photonic structures. Here, in complete contrast to the state of the art, where photonic functions are predefined by lithographically modulating the real index, we report a lithography-free paradigm for an integrated photonic processor, targeting dynamic control of spatial-temporal modulations of the imaginary index on an active semiconductor platform, without the need for lithography. We demonstrate an imaginary-index-driven methodology to tailor optical-gain distributions to rationally execute prescribed optical responses and configure desired photonic functionality to route and switch optical signals. Leveraging its real-time reconfigurability, we realize photonic neural networks with extraordinary flexibility, performing in situ training of vowel recognition with high accuracy. The programmability and multifunctionality intrinsically arising from the lithography-free characteristics can lead to a new paradigm for integrated photonic signal processing to conduct and reconfigure complex computation algorithms, accelerating the information-processing speed to achieve long-term performance requirements.

Photonics forms the backbone of today's information infrastructure, processing large datasets at unprecedented speed and with minimal energy consumption by exploiting the intrinsic parallelism, elevated frequency rates and large bandwidths that inherently come with working in the optical domain[1–6]. When targeting in situ signal control, information processing or general photonic computational operations, programmability and multifunctionality are critical factors as photonic integrated circuits evolve into a new era[7–13]. Recently, programmable photonic networks with functions of reconfigurable switching and routing have become possible through the heterogeneous integration of a range of materials[14–16] (for example, phase-change materials) and structures[17,18] (such as microelectromechanical systems (MEMS)) with tunable optical properties on semiconductor photonic chips. However, with existing integrated photonics platforms, control of optical signals is implemented by cascading discrete devices, where each device has a single functionality and a distinct morphology that is predefined by high-precision lithography of multilayered structures and is specific for its task. Strategic node connections between individual devices (such as coupled waveguides, splitters, filters and phase shifters) must be included to realize on-chip networks[19–21]. When scaling up, the complexity of the architecture inevitably grows exponentially as the number of connecting nodes and the number of single devices both increase nonlinearly with the size of the chip[22]. As a consequence, extremely complex architectures are inevitable for the realization of fully reconfigurable, high-performance integrated photonic processors that are able to handle data-intensive tasks, such as in situ training of modern artificial intelligence. Additionally, it remains a challenge to precisely control nano-lithographic features during the manufacture of very large-scale integrated photonics[23,24]. Any lithographic imperfection may cause a defect that degrades or even completely deteriorates the designed performance.

[1]Department of Materials Science and Engineering, University of Pennsylvania, Philadelphia, PA, USA. [2]Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA, USA. e-mail: fenglia@seas.upenn.edu

https://doi.org/10.1038/s41566-023-01205-0

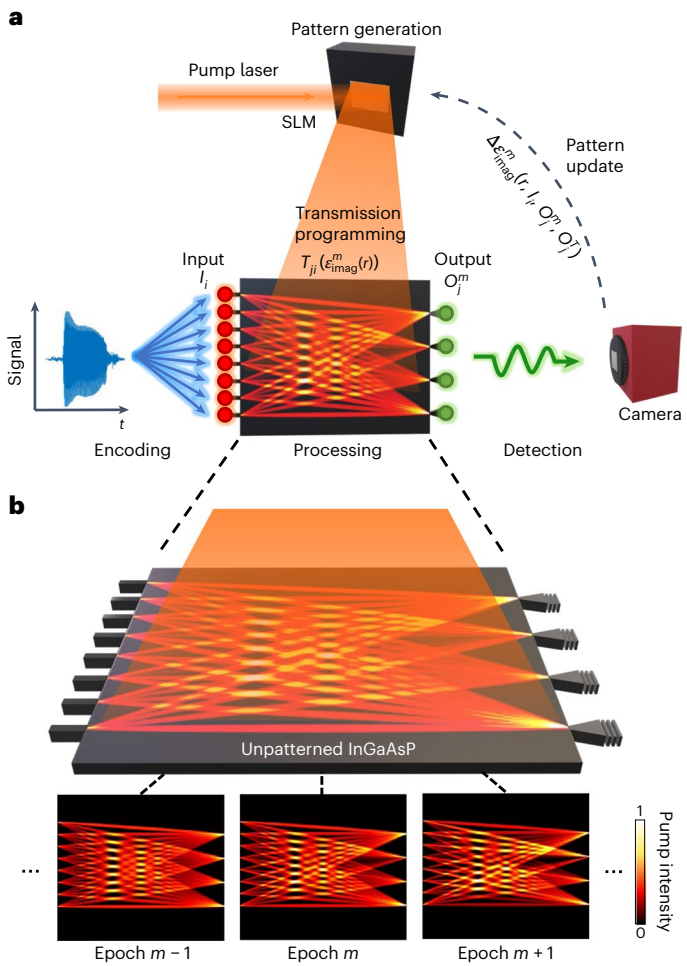


Fig. 1 | Lithography-free integrated photonic processor for on-chip signal processing and network training. a, Conceptual illustration of the imaginary-index-driven processor with real-time feedback, together with signal encoding and detection modules. With the signal encoded as the intensity of input light in different input channels (I_i), the matrix operation based on the imaginary index, $T_{ji}\left(\varepsilon^m_{\mathrm{imag}}(r)\right)$, in training epoch m, is fully programmed by an external pumping pattern generated by a SLM. The pattern as a function of $\varepsilon^m_{\mathrm{imag}}\left(r, I_i, O^m_j, O^{\mathrm{T}}_j\right)$ is real-time optimized to perform an in situ training of a photonic neural network, based on the measured powers versus the targets in different output channels (O^m_j versus O^{T}_j). **b**, The information-processing area of the lithography-free imaginary-index-driven processor is a layer of unpatterned InGaAsP. Its networking connectivity and computational function can be dynamically reconfigured by the spatial-temporal control of pumping patterns during the training process. The bottom panels display a sequence of pumping patterns updated after each training epoch.

In this Article we demonstrate a completely new, lithography-free paradigm for a reconfigurable integrated photonic processor that, as a result of its lithography-free nature, creates exceptional field programmability and functionality by fully eliminating the requirement for connecting nodes. Our work delivers a brand new and ultra-flexible integrated photonic paradigm for reconfigurable networking and computing, with great potential to process large, non-local datasets with high throughputs. Figure 1 illustrates the concept of the lithography-free integrated photonic processor, which has a central unpatterned area where dynamic control and spatial patterning of optical gain on an active III–V semiconductor platform[25–27] provide an arbitrarily field-programmable photonic network. The absence of any predefined features on this unpatterned wafer comprising InGaAsP multiple quantum wells provides the convenience of reconfigurability, where optical coding of patterned pumping light defines the gain–loss distribution and thus spatially modulates the imaginary index, in place of of real-index modulation by lithographically defined features. Note that intrinsic material losses associated with unpumped areas correspond to the imaginary index being negative, whereas optical gain arising from active pumping turns the imaginary index positive, with the modulation strength ($-0.1 \le \varepsilon_{\mathrm{imag}} \le 0.03$) being precisely controlled by the intensity of the pumping light[28,29] (Supplementary Sections 1 and 2). The algorithm-optimized spatial imaginary-index distribution forms an on-chip imaginary-index-driven photonic network that directly connects inputs with outputs, performing optical information processing according to the desired matrix-vector multiplication (that is, $O_j = \sum_i T_{ji} I_i$, where T_{ji} denotes the power transmission from input port i to output port j), where the signals are encoded by the light power in each input and output channel. The advantage of this imaginary-index-driven network is its intrinsic reconfigurability associated with convenient pattern generation and real-time transformation by optical coding using a spatial light modulator (SLM). In this scenario, the measurement results for the output light power are monitored in a real-time manner, and feedback from the detection is delivered to the SLM to update the pumping pattern either for self error correction or in situ training (Fig. 1b). Although there are typically a very large number of variables to be carefully designed and tuned, layer by layer, in a large-scale network architecture, a promising feature of our lithography-free, reconfigurable integrated photonic processor is that the information needed for pattern optimization comprises just the measured power from each input (I_i) and output (O^m_j) port for epoch m, together with the predefined target output (O^{T}_j). This unique feature of global input–output connections substantially simplifies the algorithm needed for pattern reconfiguration, thereby enabling the acceleration of simulation-free, real-time reconfigurable computing for in situ training.

To efficiently generate and optimize the spatial imaginary-index map for a specific functionality, we developed novel imaginary-index-driven inverse design algorithms (assuming the real index remains the same in the modulation region): an offline algorithm and its derived online algorithm, both following gradient-descent methods. In both algorithms, a general loss function is defined for the target function, and the algorithms minimize the loss function by estimating the gradient over the variables, which is the spatial imaginary-index profile. Figure 2a illustrates the two algorithms in flowchart form for the realization of an arbitrary power transmission matrix between the input and output ports. Here, with an imaginary-index-driven photonic processor (that is, the central information processing area without any lithographically patterned features) connected with four input (I_1–I_4) and four output (O_1–O_4) waveguides, we choose a 4 × 4 triangular matrix T as a proof of concept:

$$T = \begin{pmatrix} 0.25 & 0.33 & 0.5 & 1 \\ 0.25 & 0.33 & 0.5 & 0 \\ 0.25 & 0.33 & 0 & 0 \\ 0.25 & 0 & 0 & 0 \end{pmatrix} \quad (1)$$

The offline algorithm (Methods) is an inverse design[30,31] algorithm based on an electromagnetic field simulation and the adjoint method[32,33]. For N input channels, N parallel solvers are used to solve two-dimensional Maxwell equations under excitation of each individual waveguide (marked by red in Fig. 2a, excitation I_3 as an example) by application of the finite-difference frequency-domain (FDFD) method[34]. According to the solved field in solver i, a target-defined loss function L_i is calculated, which evaluates the deviation from the target function. An adjoint field corresponding to the loss function is subsequently simulated, similarly to the adjoint method for real-index inverse design. The map of the negative gradient to the

https://doi.org/10.1038/s41566-023-01205-0

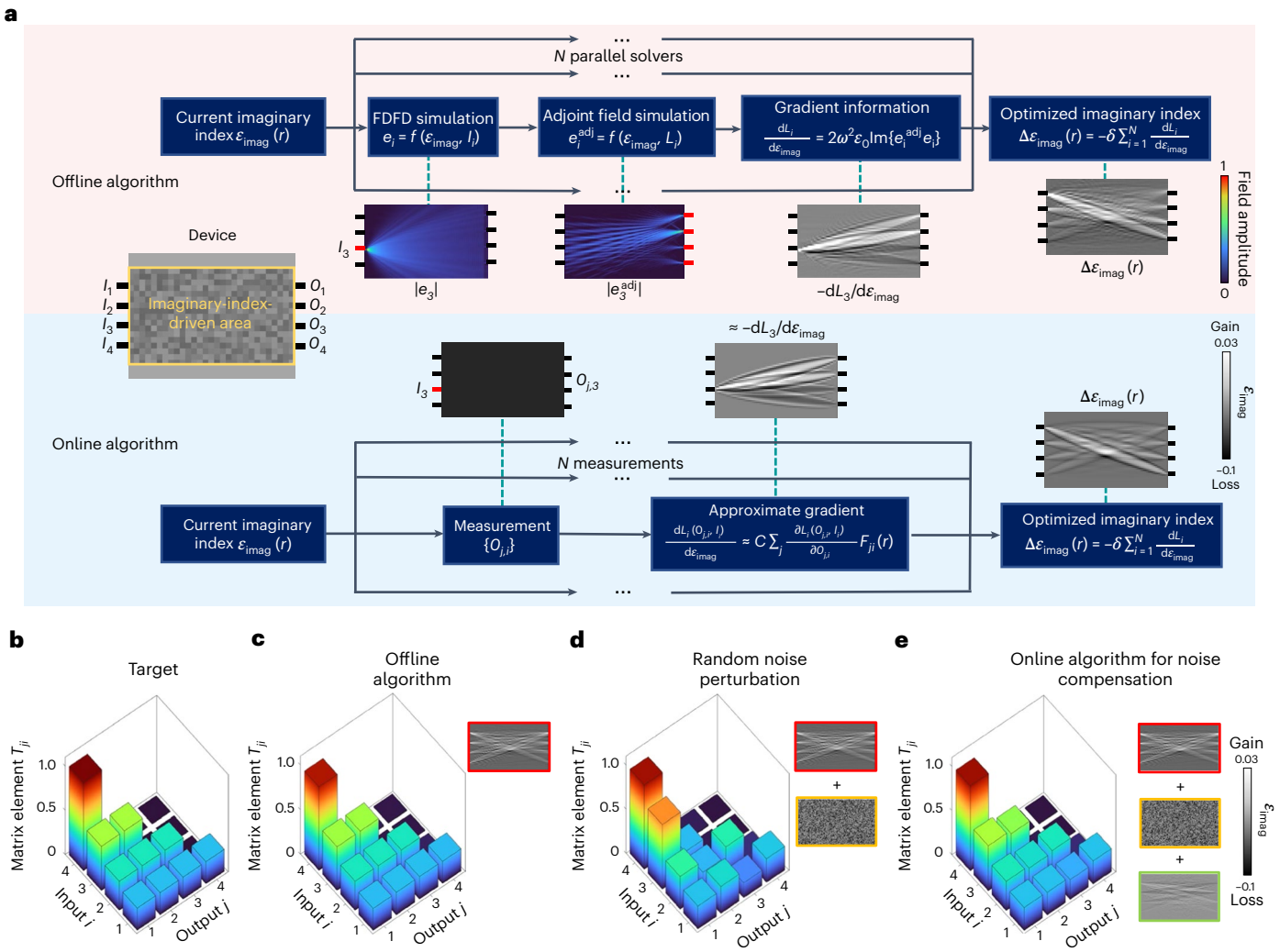


Fig. 2 | Imaginary-index-driven inverse design algorithms. a, Illustration of the two algorithms used to generate the imaginary-index maps to execute the target matrix operation. The offline algorithm (top) is simulation-based. The flowchart shows the procedure in one optimization iteration targeting a specific power transmission matrix between the input and output ports. Electromagnetic fields are simulated in the parallel FDFD solvers excited by each input channel (e_3 is shown here for excitation I_3). Next, the corresponding adjoint fields launched from the outputs are simulated (e_3^{adj}). The gradient information ($-dL_3/\varepsilon_{imag}$) is extracted from the product of the simulated field distribution and its adjoint field. Finally, the change in the imaginary index, $\Delta\varepsilon_{imag}(r)$ (shown on the right), is achieved based on the global gradient from all parallel solvers and a step constant δ. The online algorithm (bottom) is measurement-based. N measurements are performed rather than the computation-expensive large-scale electromagnetic simulations. The power in all the output channels is measured with the input channels excited one by one. An approximate gradient ($-dL_3/\varepsilon_{imag}$) is extracted by using only the measurement results and predefined spatial maps $\{F_{ji}(r)\}$. This greatly resembles the precise gradient extracted by the offline algorithm. The change in the imaginary index, $\Delta\varepsilon_{imag}(r)$ (right), is again achieved based on the global gradient. **b**–**e**, Simulation results, providing an example of a robust arbitrary power transmission matrix programmed by the imaginary-index-driven inverse design algorithms. **b**, Target transmission matrix elements. **c**, Transmission matrix and the spatial imaginary index calculated by the offline algorithm only (inset outlined in red). **d**, The transmission matrix is perturbed by a random perturbation of the imaginary index (inset outlined in yellow). **e**, The online algorithm is applied to compensate the random perturbation for the revival of the target matrix. The imaginary-index spatial map (inset outlined in green) corresponds to the change of the imaginary index optimized by the online algorithm.

imaginary index ($-dL_3/\varepsilon_{imag}$) is extracted from the results of these two simulations, that is, the spatial amplitude distribution of the simulated field e_3 and adjoint field e_3^{adj}. The final imaginary-index map $\Delta\varepsilon_{imag}(r)$, which conducts the transmission matrix T in the imaginary-index-driven area, is generated according to the global gradient by combining the gradient information from all N parallel solvers. Note that the algorithm does not limit the coherency of light between different input channels, but for the convenience of our experimental demonstrations we focus on the case where the signals from different input channels are incoherent and do not have a stable phase relation. In this case, both the signal and the transmission matrix are strictly positive real-valued.

The offline algorithm is precise and efficient in the realization of an arbitrary transmission matrix, but the device is offline during the whole process, so any mismatch between the simulation and the actual device may deteriorate the device performance, especially when the scale of the device becomes large. To bring imaginary-index-driven computing into reality, an online algorithm must be realized in which the actual device is online for real-time measurements during the entire optimization process. Although the offline algorithm requires time-consuming simulations and is thus not suitable for real-time optimizations, its generated imaginary-index map guides the development of the online algorithm (Methods), and the field profile connecting input i with

output j in the imaginary-index-driven area can be described as a series of analytical spatial maps

$$F_{ji}(r) = \begin{cases} \cos\left[k_{\text{eff}} R_{ji}(r)\right], & \text{if } R_{ji}(r) \le R_0 \\ 0, & \text{else} \end{cases} \quad (2)$$

where k_{eff} is the effective wavevector and $R_{ji}(r)$ is defined as $R_{ji}(r) = |r - r_i| + |r - r_j| - |r_i - r_j|$, where r_i and r_j are the positions of the corresponding input and output ports, respectively, and R_0 controls the spatial range of the map depending on the actual pumping pixel resolution (Supplementary Section 3). A series of spatial maps $\{F_{ji}(r)\}$, alongside the measured power at the input and output channels, can yield an approximate gradient map ($\sim -dL_3/\varepsilon_{\text{imag}}$ with excitation I_3, for example), in excellent agreement with the exact one ($-dL_3/\varepsilon_{\text{imag}}$) from the offline algorithm. Similarly, a global approximate gradient is achieved by the summation of all measurement results from all input channels, producing the target imaginary-index map online. Here, we fully exploit the aforementioned unique feature of global input–output connections in the imaginary-index-driven matrix processing area to demonstrate the simulation-free online algorithm, which enables real-time optimization for reconfigurable photonic computing and dynamic online learning.

To realize a robust performance, a combination of the two algorithms can be applied strategically, as validated by three numerical simulations taking different scenarios into account (Fig. 2c–e). With the target matrix in equation (1) displayed in Fig. 2b, an almost perfect match (Fig. 2c) is achieved using the offline algorithm. However, in practical applications, the result may deviate from the offline simulations because of a slight mismatch of index, an imperfect generation of the pumping pattern, or any random noise. To mimic such errors in a realistic experimental scenario, we introduce additional random perturbations of the imaginary index with a standard deviation of 0.01 in the imaginary-index-driven area, which consequently perturbs all the matrix elements in T, which deviate from the target result (Fig. 2d). The online algorithm is applied to successfully compensate the adverse influence of the perturbation based on only the output power, showing the capability of real-time optimizations to revive the target matrix, despite random noise (Fig. 2e).

To experimentally confirm real-time optimizations of reconfigurable photonic routing, switching and networking using our infrastructure, we next demonstrate the generation of an arbitrary matrix processor by means of dynamically controlled pumping patterns, with the corresponding intensity distribution equivalently translated from the imaginary-index map obtained by the inverse design algorithms. The reconfigurable imaginary-index-driven photonic processor comprises a 300 μm × 240 μm unpatterned area, connected by four input and four output channels and based on the InGaAsP multiple-quantum-wells platform, as shown in Fig. 3a. To take full advantage of the active nature of InGaAsP, integrated microring lasers were fabricated to directly encode information for a convenient signal input. In the experiments, two pumping beams at a wavelength of 1,064 nm were applied as the matrix processor and signal encoder (Methods). The first pumping beam, patterned by an SLM according to the algorithm-optimized imaginary-index map, impinges on the lithography-free, imaginary-index-driven area to define the photonic network and real-time-optimize the power transmission matrix. The top and bottom regions outside the optimization area remain unpumped and hence are dissipative and eliminate boundary reflections (similar to absorbing boundaries for perfect matched layers as used in numerical simulations). The other separate pumping beam is focused on individual microring lasers to excite one signal channel at a time. Note that, although the emissions from the microring lasers all occur at ~1,500 nm, they differ slightly from one another, with the detuning measured in the range of 3 nm. As a result, the input signals from different channels become intrinsically incoherent with respect to one another. On the other end, the normalized output 4 × 1 vector is collected for power transmission, and grating couplers are implemented for efficient detection of the signal output in free space.

With the same target transmission matrix T as given in equation (1) (Fig. 3b), a pumping pattern according to the offline algorithm was first generated and applied in the imaginary-index-driven matrix processing area (Fig. 3c, right). Although the offline simulation yields a nearly perfect transmission matrix with a fidelity $f = \frac{\text{tr}\left(\sqrt{T'\sqrt{M}}\right)}{\sqrt{\text{tr}\left(\sqrt{T'\sqrt{T}}\right)}\sqrt{\text{tr}\left(\sqrt{M'\sqrt{M}}\right)}}$ (where M represents the measured transmission matrix, T' denotes the transpose of T, and the square root is applied to each matrix element) of over 99%, the performance in real experiments does not match the target (Supplementary Section 4), yielding a fidelity of only 93% and thus leaving sufficient space for online optimizations (Fig. 3c). To compensate the deviation, the online algorithm is applied to adjust the pumping pattern according to the measurement results in real time, leading to an increased fidelity of the measured transmission matrix (to 99.2%) after nine iterations (Fig. 3d). With the accumulated optimization of the pumping distribution in the real-time optimization process (Fig. 3e, top panels), the evolution of the transmission matrix shows its gradual convergence to the target (Fig. 3e, bottom panels). The improvement arising from measurement feedback convincingly demonstrates the validity of our online algorithm, which is critical to prevent error cascading in a large-scale network. Because the matrix processor is fully programmed by dynamic control of the pumping pattern and its functionality does not rely on any lithography-patterned structures, the imaginary-index-driven optical coding scheme can be arbitrarily reconfigured and optimized in a real-time manner for reconfigurable computing acceleration.

To exploit the demonstrated dynamic reconfigurability for the acceleration of computing so as to handle data-intensive tasks, we performed in situ machine learning, where the pumping pattern was trained online to reconfigure in real time the network connectivity or weight. A classical four-vowel ('er', 'iy', 'oa' and 'ae') classification task was applied to demonstrate the concept. The dataset[35] consists of the speech of different vowels (from both males and females), divided into a training set and a testing set, each containing 64 audio files. A fully connected neuro-photonic network was executed using an imaginary-index-driven photonic processor with an unpatterned, active area of 500 μm × 324 μm (Fig. 4a). Despite a large amount of redundant information in the audio files, eight prominent features in frequency bands associated with the vowels were selectively extracted to accurately represent the training database, to then be encoded as input signals. In the input layer, any 8 out of 12 microring lasers could be excited, and the strength of each feature was encoded as the power of the corresponding microring laser emission, which was precisely controlled by the pumping intensity. An iterative method was applied to guarantee that the power of the eight microring lasers perfectly matched the features in the dataset, with an average encoding fidelity of 99.9% achieved in experiments (Supplementary Section 5). In the output layer, the four vowels were categorized as output channels 1 to 4 (any 4 out of 12) corresponding to 'er', 'iy', 'oa' and 'ae', respectively. The predicted class was directly indicated by the highest intensity among the outputs.

In contrast to the computer-trained target matrix, dynamic online learning is used to process the training dataset with iterative measurement feedbacks to in situ-identify the most appropriate matrix for the classification task of vowel recognition. Therefore, instead of starting with the pumping pattern from the offline algorithm, the photonic processor is initialized with a symmetric pumping pattern that connects all input and output channels, which is subsequently in situ-trained using the online algorithm. In the mth training epoch, the error $\text{err}_{j,k}^{m}$ between the network prediction (measured outputs) and the ground truth (target scenario) can be calculated at output j for training data k, by which the variations of all the matrix elements

https://doi.org/10.1038/s41566-023-01205-0

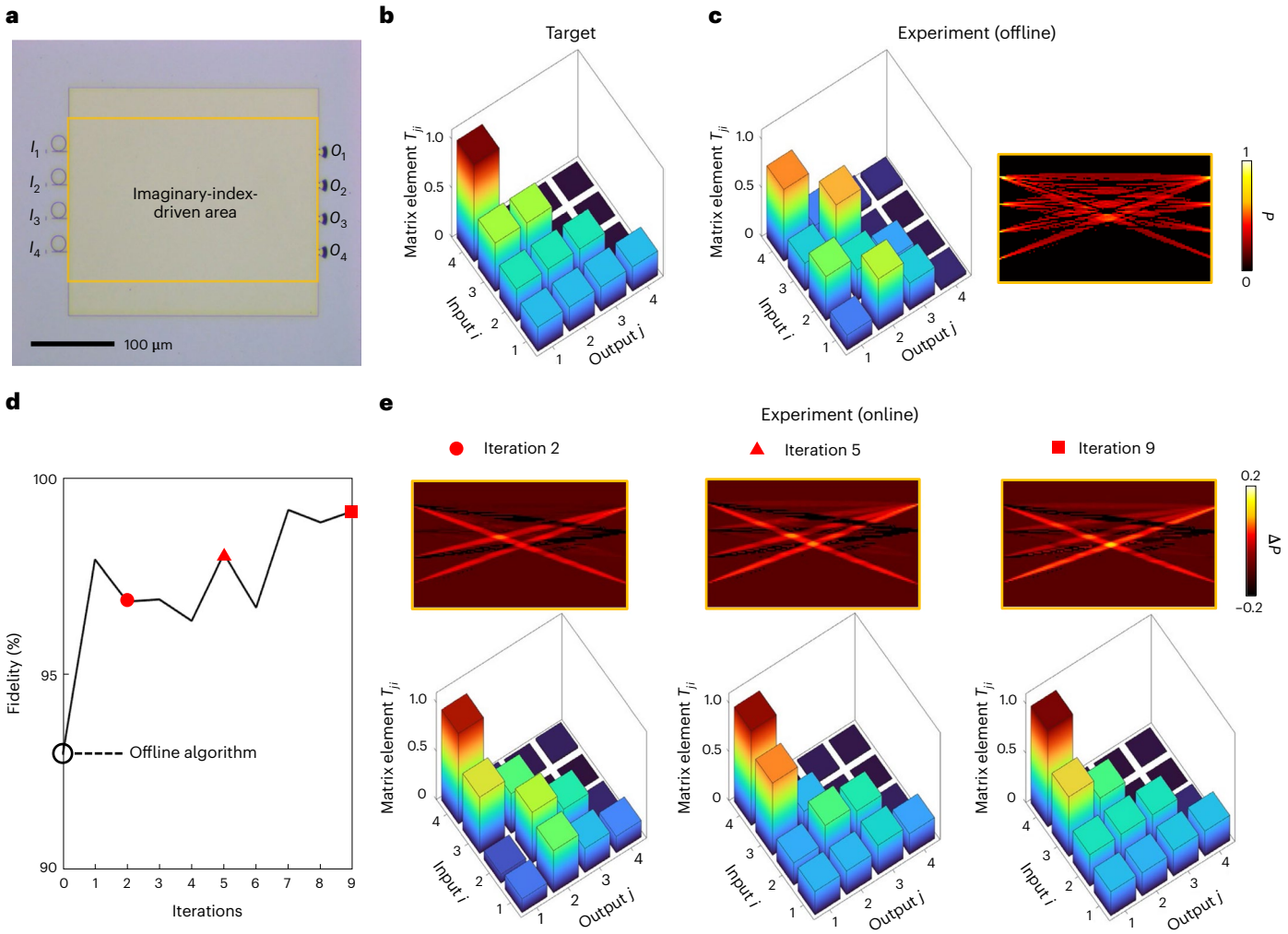


Fig. 3 | Experimental demonstration of an imaginary-index-driven arbitrary matrix processor. a, Optical microscope image of a 4 × 4 device on the InGaAsP platform, which consists of an unpatterned imaginary-index-driven area for signal processing, connected to four microring lasers for input signal encoding ($I_1 - I_4$) and four grating couplers for output signal detection ($O_1 - O_4$). **b**, The target transmission matrix. **c**, The measured transmission matrix (left), and the pumping pattern generated by the offline algorithm (right). **d**, The online algorithm is applied to improve the fidelity of the matrix operation. A fidelity of 99.2% is achieved after nine iterations. **e**, Evolution of the pumping profile change, ΔP (top panels), leads to the real-time optimization of the matrix operation (bottom panels). Although the lithography-free processor can in principle respond as fast as the carrier lifetime of InGaAsP (that is, ~200 ps; ref. 37), real measurements in one iteration (five frames) take ~100 ms, limited by the frame rate of the used infrared camera.

needed for the next epoch can be in situ-updated according to the error backpropagation: $\Delta T_{ji}^m \propto -\frac{\partial L}{\partial T_{ji}^m} = -\sum_k I_{i,k} \mathrm{err}_{j,k}^m$. Here, L is the loss function, defined in a mean-square-error format, and $I_{i,k}$ denotes the measured input power at input i for training data k. Consequently, with the preloaded analytical spatial maps $\{F_{ji}(r)\}$ in equation (2), the updated imaginary-index map can be online obtained in real time from

$$\Delta\varepsilon_{\mathrm{imag}}^m(r) = -\delta\frac{\partial L}{\partial \varepsilon_{\mathrm{imag}}^m(r)} = -\delta\sum_{i,j}\frac{\partial L}{\partial T_{ji}^m}\frac{\partial T_{ji}^m}{\partial \varepsilon_{\mathrm{imag}}^m(r)} = -\delta\sum_{i,j,k} I_{i,k}\mathrm{err}_{j,k}^m F_{ji}(r) \quad (3)$$

which guides the dynamic reconfiguration of the pumping pattern for the next in situ training epoch. Here, δ is a constant learning rate. In this scenario, the optical network is in situ-trained without the physically implemented error backpropagation and its associated complex algorithms (Methods). With dynamic online learning, our device demonstrates high accuracy in vowel recognition (Fig. 4b). After carrying out 65 training epochs to achieve a classification accuracy of 98.4% for the training dataset, our device achieves a high accuracy of 93.8% for the testing dataset, in contrast to the initial accuracy of only 15.6%. More specifically, Fig. 4c,d shows the distribution of measured optical signals in the input and output layers, respectively, showing the high performance of in situ dynamic learning for this classification task. In the input layer, all four classes of data are mixed and overlapped with each other, making the recognition task challenging. In particular, vowels 'oa' (blue) and 'ae' (grey) have a significant overlap in the parameter space of input features, which leads to a small overlap (but distinguishable) between them in the output layer. Meanwhile, vowels 'er' (red) and 'iy' (green) are completely separated in the output layer. The performance of the classification results is quantitatively demonstrated based on the confusion matrix of the testing dataset in Fig. 4e, which defines the percentage of correctly identified vowels along its diagonal and the percentage of incorrectly identified vowels in the off-diagonal terms. The strong diagonal distribution highlights the impressive performance, showing the potential to handle data-intensive computing tasks in real time. The most unique feature

https://doi.org/10.1038/s41566-023-01205-0

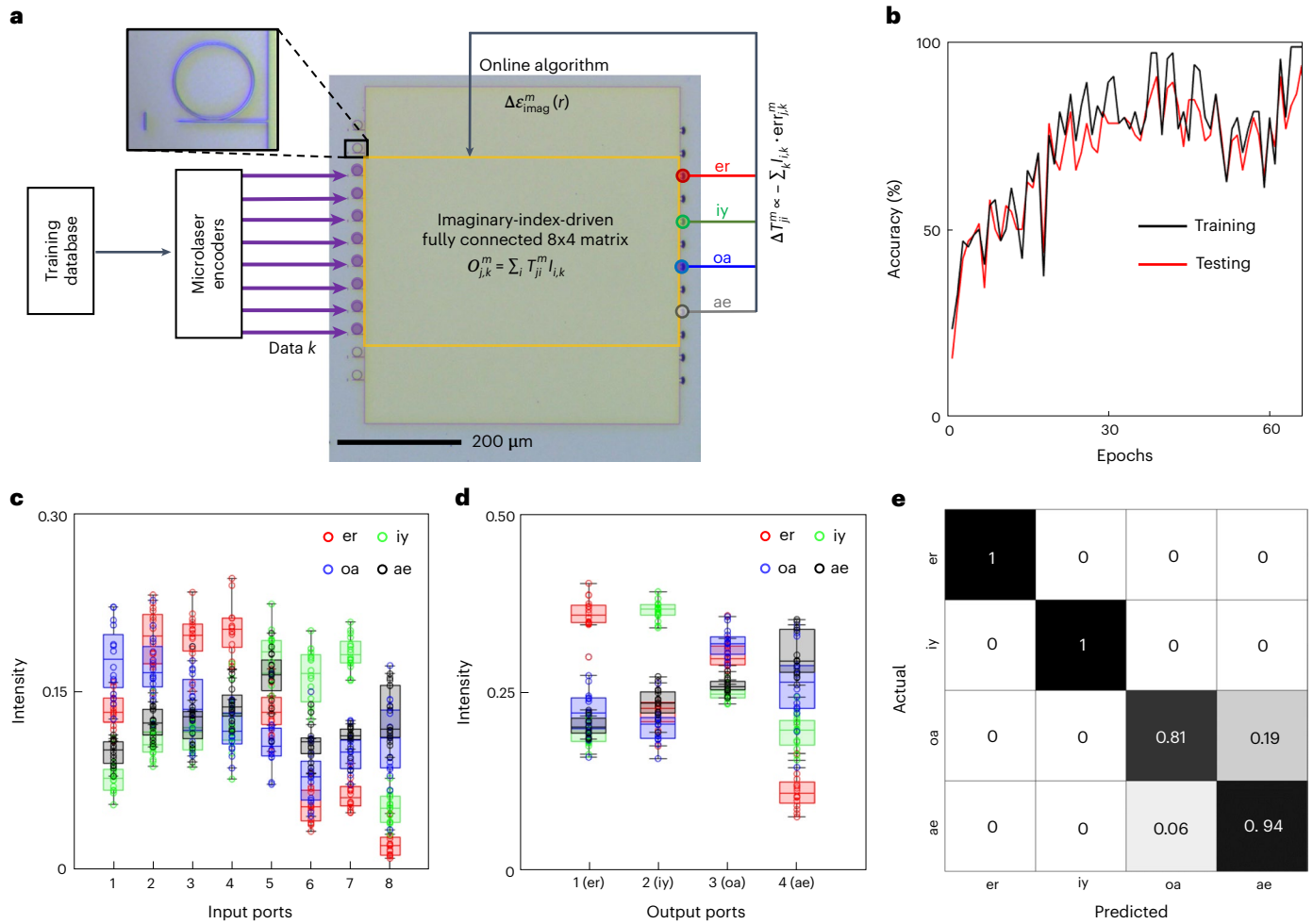


Fig. 4 | In situ training for vowel recognition. a, Optical microscope image of the device and the schematic for in situ training of a four-class vowel recognition task. Eight input channels (purple) and four output channels (red, green, blue and grey, corresponding to the different vowel classes) are used. The features extracted from the raw audio files are applied as the input neurons, which are encoded by the microring lasers and monitored by the camera through the laser output, from the left (see the inset for details). The central yellow box indicates the imaginary-index-driven photonic processor for this in situ training task, where the pumping pattern is updated after each training epoch by the online algorithm. **b**, Evolution of the recognition accuracy of the training (black) and testing (red) data with iterative training epochs. **c**,**d**, Scattering plots of the measured power at the input (**c**) and output (**d**) ports for all 64 testing data after training. The four vowel classes are shown in different colours corresponding to the colour identification in **a**. For each vowel class, its associated interquartile range (indicated by the boxes) is featured in all the input and output ports. The coloured line inside each box denotes the median, and the black whiskers connect the upper (lower) quartile to the non-outlier maximum (minimum), marked by the horizontal black lines. The scatters that are more than 1.5× the interquartile range away from the edges of the boxes are considered to be outliers. **e**, Confusion matrix for the testing data, where the values are normalized in each row.

associated with the in situ learning process, in contrast to any ex situ ones, is that the gradient information that drives the weight update is directly measured and extracted from the real device. Hence, this real-time optimization process, with the device in the loop, can assure high-performance computing in a large-scale network, instead of relying on either perfect fabrication or computationally expensive and complicated modelling.

We have demonstrated a new lithography-free integrated photonic processor, where the lithography-free nature provides the convenience of reconfigurability, demonstrated by optical coding of spatial-temporal modulations of the imaginary index on an active semiconductor platform. Dynamic control of the imaginary-index modulation is used to reconfigure the global photonic network connectivity for in situ machine learning. Note that, although the photonic processor itself does not require any lithographically defined features inside, its connections with other devices for signal input and output (such as microring lasers and grating couplers in our experiment, which could be replaced by lensed fibre systems) may still require elementary-level lithography. Nevertheless, it is worth emphasizing that the optical signals here are fully on-chip-processed in a lithography-free core driven by spatial-temporal control of the imaginary index. Accordingly, in this scenario, the need for high-precision lithography in integrated photonics can be drastically reduced. With the gain spectrum of the active semiconductor over 100 nm (Supplementary Section 6), the imaginary-index-driven photonic processor holds potential for broadband operation. Moreover, beyond the demonstrated reconfigurable computing in the linear regime, the carrier dynamics in the active semiconductor platform could be further explored to create optical nonlinearity[36], for example, with saturable gain or loss where the imaginary-index modulation becomes nonlinear with respect to photon density (Supplementary Section 7). The successful realization of optical nonlinearity in integrated photonics could further enhance neural-photonic computing acceleration for high-throughput, data-intensive applications.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41566-023-01205-0.

References

1. Shastri, B. J. et al. Photonics for artificial intelligence and neuromorphic computing. *Nat. Photon.* **15**, 102–114 (2021).
2. Feldmann, J. et al. Parallel convolutional processing using an integrated photonic tensor core. *Nature* **589**, 52–58 (2021).
3. Cheben, P., Halir, R., Schmid, J. H., Atwater, H. A. & Smith, D. R. Subwavelength integrated photonics. *Nature* **560**, 565–572 (2018).
4. Atabaki, A. H. et al. Integrating photonics with silicon nanoelectronics for the next generation of systems on a chip. *Nature* **556**, 349–354 (2018).
5. Sludds, A. et al. Delocalized photonic deep learning on the internet's edge. *Science* **378**, 270–276 (2022).
6. Piggott, A. Y. et al. Inverse design and demonstration of a compact and broadband on-chip wavelength demultiplexer. *Nat. Photon.* **9**, 374–377 (2015).
7. Bogaerts, W. et al. Programmable photonic circuits. *Nature* **586**, 207–216 (2020).
8. Hughes, T. W., Minkov, M., Shi, Y. & Fan, S. Training of photonic neural networks through in situ backpropagation and gradient measurement. *Optica* **5**, 864–871 (2018).
9. Arrazola, J. M. et al. Quantum circuits with many photons on a programmable nanophotonic chip. *Nature* **591**, 54–60 (2021).
10. Feldmann, J., Youngblood, N., Wright, C. D., Bhaskaran, H. & Pernice, W. H. P. All-optical spiking neurosynaptic networks with self-learning capabilities. *Nature* **569**, 208–214 (2019).
11. Zhang, W. & Yao, J. Photonic integrated field-programmable disk array signal processor. *Nat. Commun.* **11**, 406 (2020).
12. Liu, W. et al. A fully reconfigurable photonic integrated signal processor. *Nat. Photon.* **10**, 190–195 (2016).
13. Zhao, H., Li, B., Li, H. & Li, M. Enabling scalable optical computing in synthetic frequency dimension using integrated cavity acousto-optics. *Nat. Commun.* **13**, 5426 (2022).
14. Zhang, W., Mazzarello, R., Wuttig, M. & Ma, E. Designing crystallization in phase-change materials for universal memory and neuro-inspired computing. *Nat. Rev. Mater.* **4**, 150–168 (2019).
15. Wu, C. et al. Programmable phase-change metasurfaces on waveguides for multimode photonic convolutional neural network. *Nat. Commun.* **12**, 96 (2021).
16. Wuttig, M., Bhaskaran, H. & Taubner, T. Phase-change materials for non-volatile photonic applications. *Nat. Photon.* **11**, 465–476 (2017).
17. Han, S. et al. Large-scale polarization-insensitive silicon photonic MEMS switches. *J. Lightwave Technol.* **36**, 1824–1830 (2018).
18. Seok, T. J., Quack, N., Han, S., Muller, R. S. & Wu, M. C. Large-scale broadband digital silicon photonic switches with vertical adiabatic couplers. *Optica* **3**, 64–70 (2016).
19. Shen, Y. et al. Deep learning with coherent nanophotonic circuits. *Nat. Photon.* **11**, 441–446 (2017).
20. Zhang, H. et al. An optical neural chip for implementing complex-valued neural network. *Nat. Commun.* **12**, 457 (2021).
21. Ashtiani, F., Geers, A. J. & Aflatouni, F. An on-chip photonic deep neural network for image classification. *Nature* **606**, 501–506 (2022).
22. Reck, M., Zeilinger, A., Bernstein, H. J. & Bertani, P. Experimental realization of any discrete unitary operator. *Phys. Rev. Lett.* **73**, 58–61 (1994).
23. Nagarajan, R. et al. Large-scale photonic integrated circuits. *IEEE J. Sel. Top. Quantum Electron.* **11**, 50–65 (2005).
24. Chrostowski, L. et al. Impact of fabrication non-uniformity on chip-scale silicon photonic integrated circuits. In *Proc. Optical Fiber Communication Conference* paper Th2A.37 (OSA, 2014); https://doi.org/10.1364/ofc.2014.th2a.37
25. Zhang, Z. et al. Tunable topological charge vortex microlaser. *Science* **368**, 760–763 (2020).
26. Bahari, B. et al. Photonic quantum Hall effect and multiplexed light sources of large orbital angular momenta. *Nat. Phys.* **17**, 700–703 (2021).
27. Mao, X.-R., Shao, Z.-K., Luan, H.-Y., Wang, S.-L. & Ma, R.-M. Magic-angle lasers in nanostructured moiré superlattice. *Nat. Nanotechnol.* **16**, 1099–1105 (2021).
28. Zhao, H. et al. Non-Hermitian topological light steering. *Science* **365**, 1163–1166 (2019).
29. Qiao, X. et al. Higher-dimensional supersymmetric microlaser arrays. *Science* **372**, 403–408 (2021).
30. Molesky, S. et al. Inverse design in nanophotonics. *Nat. Photon.* **12**, 659–670 (2018).
31. Piggott, A. Y., Petykiewicz, J., Su, L. & Vučković, J. Fabrication-constrained nanophotonic inverse design. *Sci. Rep.* **7**, 1786 (2017).
32. Hughes, T. W., Minkov, M., Williamson, I. A. D. & Fan, S. Adjoint method and inverse design for nonlinear nanophotonic devices. *ACS Photonics* **5**, 4781–4787 (2018).
33. Veronis, G., Dutton, R. W. & Fan, S. Method for sensitivity analysis of photonic crystal devices. *Opt. Lett.* **29**, 2288–2290 (2004).
34. Rumpf, R. C. Simple implementation of arbitrarily shaped total-field/scattered-field regions in finite-difference frequency-domain. *Prog. Electromagn. Res. B* **36**, 221–248 (2012).
35. Hillenbrand, J., Getty, L. A., Clark, M. J. & Wheeler, K. Acoustic characteristics of American English vowels. *J. Acoust. Soc. Am.* **97**, 3099–3111 (1995).
36. Hall, K. L., Lenz, G., Darwish, A. M. & Ippen, E. P. Subpicosecond gain and index nonlinearities in InGaAsP diode lasers. *Opt. Commun.* **111**, 589–612 (1994).
37. Zhang, Z. et al. Ultrafast control of fractional orbital angular momentum of microlaser emissions. *Light Sci. Appl.* **9**, 179 (2020).

Methods

Offline algorithm

For a processor with N input ports, N parallel solvers work simultaneously, and each of them simulates a case with one excited channel. The total loss function is defined as

$$L = \sum_{i=1}^{N} L_i = \sum_{i=1}^{N} \frac{1}{2} \sum_j (O_{j,i} - T_{ji} I_i)^2$$

where $O_{j,i}$ is the power in output port j when only input port i is excited, and T is the target transmission matrix. To find a spatial imaginary-index modulation that gives the target transmission, the gradient information $\partial L/\partial \varepsilon_{\text{imag}}(r)$ is critical. The adjoint method used in real-index inverse design was adapted for our imaginary-index-driven photonic processor for gradient extraction. First, two-dimensional Maxwell equations are solved to obtain the field $e_i(r)$, where the sources are incorporated using the total-field/scattered field (TF/SF) formulation. The fields are then used to calculate the derivative $\partial L/\partial e_i(r)$, which is applied as the excitation source for the adjoint field $e_i^{\text{adj}}(r)$ following Maxwell's equations in the same system:

$$\mu_0^{-1} \nabla \times \nabla \times e_i^{\text{adj}}(r) - \omega^2 \varepsilon_0 \varepsilon_{\text{r}}(r)\, e_i^{\text{adj}}(r) = -\frac{\partial L}{\partial e_i(r)}$$

where $\varepsilon_{\text{r}}(r)$ is the complex relative permittivity. Once the original and adjoint fields are obtained, the gradient in one solver is given by

$$\frac{\partial L_i}{\partial \varepsilon_{\text{imag}}(r)} = 2\omega^2 \varepsilon_0 \text{Im}\left\{e_i^{\text{adj}}(r) e_i(r)\right\}$$

where Im represents the imaginary part of the complex value. The global gradient is calculated by the summation of the results from all parallel solvers:

$$\frac{\partial L}{\partial \varepsilon_{\text{imag}}(r)} = \sum_{i=1}^{N} \frac{\partial L_i}{\partial \varepsilon_{\text{imag}}(r)}$$

To reduce the number of evaluations of the original and adjoint fields, the imaginary index is updated using a limited-memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) optimization algorithm that improves the convergence rate without substantially increasing the memory requirements by providing an approximate inverse Hessian matrix[38]. Supplementary Video 1 shows the optimization process. The simulations are performed in a processor with dimensions of 150 μm × 90 μm. The grid size used for the FDFD method is 100 nm × 100 nm for a signal with a free-space wavelength of 1,500 nm. The pixel resolution of the spatial imaginary-index modulation is limited to 2 μm × 2 μm, consistent with the feasibility in our experiments (Supplementary Section 3).

Online algorithm

Different from the real-index inverse design, we apply approximations to the imaginary-index-driven inverse design to substantially simplify the gradient extraction. Starting from the precise adjoint method, the excitation source of the adjoint field at the output port j is

$$b_i^{\text{adj}}(r_j) = -\frac{\partial L_i}{\partial e_i(r_j)} = -\frac{\partial L_i}{\partial O_{j,i}} \frac{\partial O_{j,i}}{\partial e_i(r_j)} \propto -\frac{\partial L_i}{\partial O_{j,i}} e_i^*(r_j)$$

where the output power at port j is $O_{j,i} \propto \sum_{r_j} e_i^*(r_j) e_i(r_j)$. The precise gradient can be divided into an amplitude term and a phase term:

$$\frac{\partial L_i}{\partial \varepsilon_{\text{imag}}(r)} \propto \text{Im}\left\{e_i^{\text{adj}}(r)\, e_i(r)\right\} = \left|e_i^{\text{adj}}(r)\, e_i(r)\right| \sin\left[\varphi_i^{\text{adj}}(r) + \varphi_i(r)\right]$$

where φ_i^{adj} and $\varphi_i(r)$ are the phase of the adjoint field and the original field, respectively. The sine term for the phase is important, as it controls the sign of the value and thus determines the imaginary index and thus either gain or loss for the next iteration. At the position of the output ports, the phase relation of these two fields is fixed, because the excitation source of the adjoint field is proportional to the conjugation of the original field. By considering the phase difference of −π/2 between the excitation and the field, we obtain

$$\sin\left[\varphi_i^{\text{adj}}(r_j) + \varphi_i(r_j)\right] = \pm 1$$

where the sign on the right side remains the same as the sign of $\partial L_i/\partial O_{j,i}$. Due to the dimensions of our device (about two orders of magnitude greater than the wavelength) and the relatively weak imaginary-index modulation, point-source approximations, located at the position r_i and r_j marked by red and blue circles in Extended Data Fig. 1a, can be safely applied for the incidence of the original and adjoint field. In this way, we can approximately write the phase term as

$$\sin\left[\varphi_i^{\text{adj}}(r) + \varphi_i(r)\right] = \pm\cos\left[k_{\text{eff}} R(r)\right]$$

where k_{eff} is the effective wavevector and $R(r)$ is defined as $R(r) = |r - r_i| + |r - r_j| - |r_i - r_j|$. Because the excitation of the original field is normalized and the adjoint field intensity is proportional to the error, $\text{err}_{j,i} = \partial L_i/\partial O_{j,i} = (O_{j,i} - T_{ji} I_i)$, which can be calculated by the measured output and input power in the experiments, the gradient is simplified as

$$\frac{\partial L_i}{\partial \varepsilon_{\text{imag}}(r)} \approx C \sum_j \text{err}_{j,i} \cos\left[k_{\text{eff}} R(r)\right]$$

where C is a constant. The cosine term gives a series of elliptical contours lines (Extended Data Fig. 1b). By considering the actual pumping resolution, we are only interested in the sparse pattern near the line connecting the corresponding input and output ports. A spatial map

$$F_{ji}(\mathbf{r}) = \begin{cases} \cos\left[k_{\text{eff}} R_{ji}(r)\right], & \text{if } R_{ji}(r) \le R_0 \\ 0, & \text{else} \end{cases}$$

is finally used to describe the approximate gradient:

$$\frac{\partial L_i}{\partial \varepsilon_{\text{imag}}(r)} \approx C \sum_j \text{err}_{j,i} F_{ji}(r)$$

Because the spatial maps $\{F_{ji}(r)\}$ are analytical and can be preloaded, the system can be optimized based only on the measurements of the light power in each port, without the application of the computationally expensive offline algorithm. Supplementary Video 2 shows the optimization process using the online algorithm, where the simulated output powers at each port are used to mimic the measurements. A range parameter $R_0 = \frac{5}{4}\lambda_{\text{eff}}$ is used (marked by the white dashed ellipse in Extended Data Fig. 1b). Although the convergent speed is slower than the offline algorithm, a perfect performance is also reached at the end, convincingly validating the performance of the online algorithm for the optimization.

Sample preparation

A wafer consisting of 220-nm-thick InGaAsP multiple quantum wells on an InP substrate was used to fabricate the lithography-free photonic processor and its connected signal input/output modules. On this active semiconductor platform, we performed electron-beam lithography (EBL) to pattern the sample, including the central lithography-free area, as well as the microlasers for signal input and the grating couplers for signal output. Hydrogen silsesquioxane solution in methyl isobutyl ketone was used as a negative resist. After exposure, the wafer was

developed using tetramethylammonium hydroxide solution (MFCD-26) and rinsed in deionized water. The exposed and developed resist thus served as a mask for the subsequent inductively coupled plasma reactive ion etching by BCl_3:Ar plasma. After dry etching of the InGaAsP, the remaining resist was removed by immersing the sample in buffered oxide etchant. A 3-μm-thick cladding layer of Si_3N_4 was then deposited on the patterned structures alongside the unpatterned main processor area, using plasma-enhanced chemical vapour deposition. Finally, the sample was bonded to a piece of glass slide, and the InP substrate was selectively removed by wet etching with a mixture of HCl and H_3PO_4.

Measurements of the transmission matrix

The optical set-up is shown in Extended Data Fig. 2a. The pumping beam is from a nanosecond pulse laser (wavelength of 1,064 nm). The pump is divided into two paths by a beamsplitter. The one modulated by SLM 1 is used to generate the pattern that programs the transmission matrix of the processor (average pumping power of 1.5 mW). A toolbox (OTSLM[39]) for structured light methods is used to generate the hologram for the target pumping pattern based on the Gerchberg–Saxton (GS) algorithm. The numerical aperture (NA) of the ×10 objective is 0.45, which guarantees a pumping pixel resolution of 2 μm (Extended Data Fig. 2b). The other path, modulated by SLM 2, is used for microring laser excitation and input signal encoding. The radii and widths of the microring lasers are designed to lase at a single longitudinal mode around 1,500 nm. A typical spectrum collected at one input channel is shown in Extended Data Fig. 2c. The lasing wavelengths of the microrings for different channels are slightly detuned, with a measured range of 3 nm. The signal emitted from the chip is collected by a ×20 objective, and the photoluminescence (PL) is first filtered by the bandpass filter centred at 1,500 nm with a bandwidth of 12 nm. The intensities of the output ports are captured by an infrared charge-coupled device camera. The image with only the pumping pattern from SLM 1 applied (no microlaser is excited) is captured as a reference. A reference subtraction was used eliminate the influence of the PL. The PL subtraction method was confirmed by the spectrum collected at one output port (Extended Data Fig. 2d), with the off-resonant PL signal remaining the same level with and without an input signal from a microlaser. To operate in a real-time manner, the signals are recorded by the camera rather than spectrum measurements during online optimization. Extended Data Fig. 2e–h presents images for when the microlasers for different input channels are excited one by one under the same pumping pattern, optimized by the online algorithm. The power in each output port is integrated over the area marked by the light dashed boxes and the output vector is then normalized over all output ports. The normalized values are used as the transmission matrix elements instead of the absolute transmission, defined as the power ratio of output to input. The transmission matrix elements extracted from the four images are shown in Fig. 3 (Fig. 3e, iteration 9).

In situ training for vowel recognition

A total of 128 samples of audio data for four vowel classes (from different male and female individuals) were randomly picked from the vowel dataset. The data were randomly divided into a training set and a testing set. The audio files were recorded with a sampling rate of 16 kHz. To remove the redundant information, the bark spectra were extracted by a feature extraction function from the MATLAB Audio Toolbox. The eight features from the spectra were used as the input vector to the processor.

In the in situ training for vowel recognition in the experiments, the loss function for the training is $L = \frac{1}{2}\sum_k \sum_j \mathrm{err}_{j,k}^2$, with the error defined by $\mathrm{err}_{j,k} = O_{j,k} - O_{j,k}^{\mathrm{T}}$. Here $O_{j,k}$ and $O_{j,k}^{\mathrm{T}}$ are the measured and target power intensities in output port j for training audio data k. The target is defined as $O_{k,j}^{\mathrm{T}} = \beta + (1-4\beta)\delta_{jl_k}$, where l_k is the true label for the training data k, and β is a constant. For one layer of the linear matrix operation, the output contrast for different vowel classes is expected not to reach a significantly high level, so a bias of $\beta = 0.15$ is introduced to make the training focus more on the overall performance. Similar to the optimization of the power transmission matrix, the online algorithm was applied to update the imaginary index according to only the measurement results of the light power at the inputs and outputs after each training epoch. Note that the pumping pattern was only updated according to the training set, and the testing set was only applied to record the accuracy in each epoch. The in situ training process can be described by the flowchart shown in Extended Data Fig. 3. The training was performed until the training accuracy reached a desired target, a_t.

Data availability

Data that support the findings of this study are available at https://doi.org/10.6084/m9.figshare.22320649.v1.

Code availability

The computer codes that support the plots within this paper and other findings of this study are available from the corresponding author upon reasonable request.

References

38. Moritz, P., Nishihara, R. & Jordan, M. A linearly-convergent stochastic L-BFGS algorithm. In *Proc. 19th International Conference on Artificial Intelligence and Statistics* (eds Gretton, A. & Robert, C. C.) 249–258 (PMLR, 2016).
39. Lenton, I. C. D., Stilgoe, A. B., Nieminen, T. A. & Rubinsztein-Dunlop, H. OTSLM toolbox for Structured Light Methods. *Comput. Phys. Commun.* **253**, 107199 (2020).

Acknowledgements

We acknowledge support from the Defense Advanced Research Projects Agency (DARPA) Young Faculty Program (W911NF-21-1-0340), Army Research Office (ARO; W911NF-21-1-0148) and National Science Foundation (NSF; ECCS-2023780). This work was carried out in part at the Singh Center for Nanotechnology, which is supported by the NSF National Nanotechnology Coordinated Infrastructure Program under grant no. NNCI-1542153.

Author contributions

T.W. and L.F. conceived the project. T.W. and M.M. developed the algorithms and performed simulations. T.W. fabricated the samples and conducted optical measurements. L.F. guided the research. All authors contributed to discussions and paper preparation.

Competing interests

The authors declare no competing interests.

Additional information

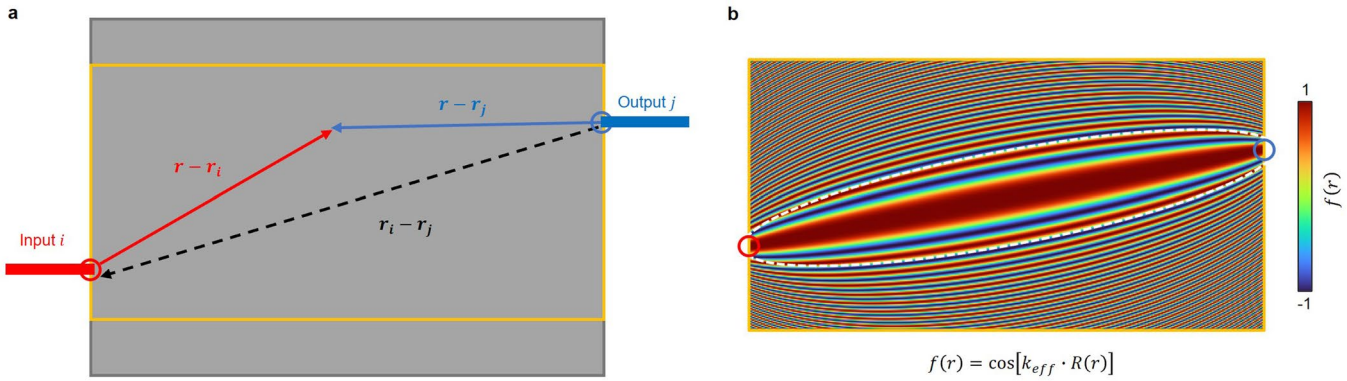
Extended data is available for this paper at https://doi.org/10.1038/s41566-023-01205-0.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41566-023-01205-0.

Correspondence and requests for materials should be addressed to Liang Feng.

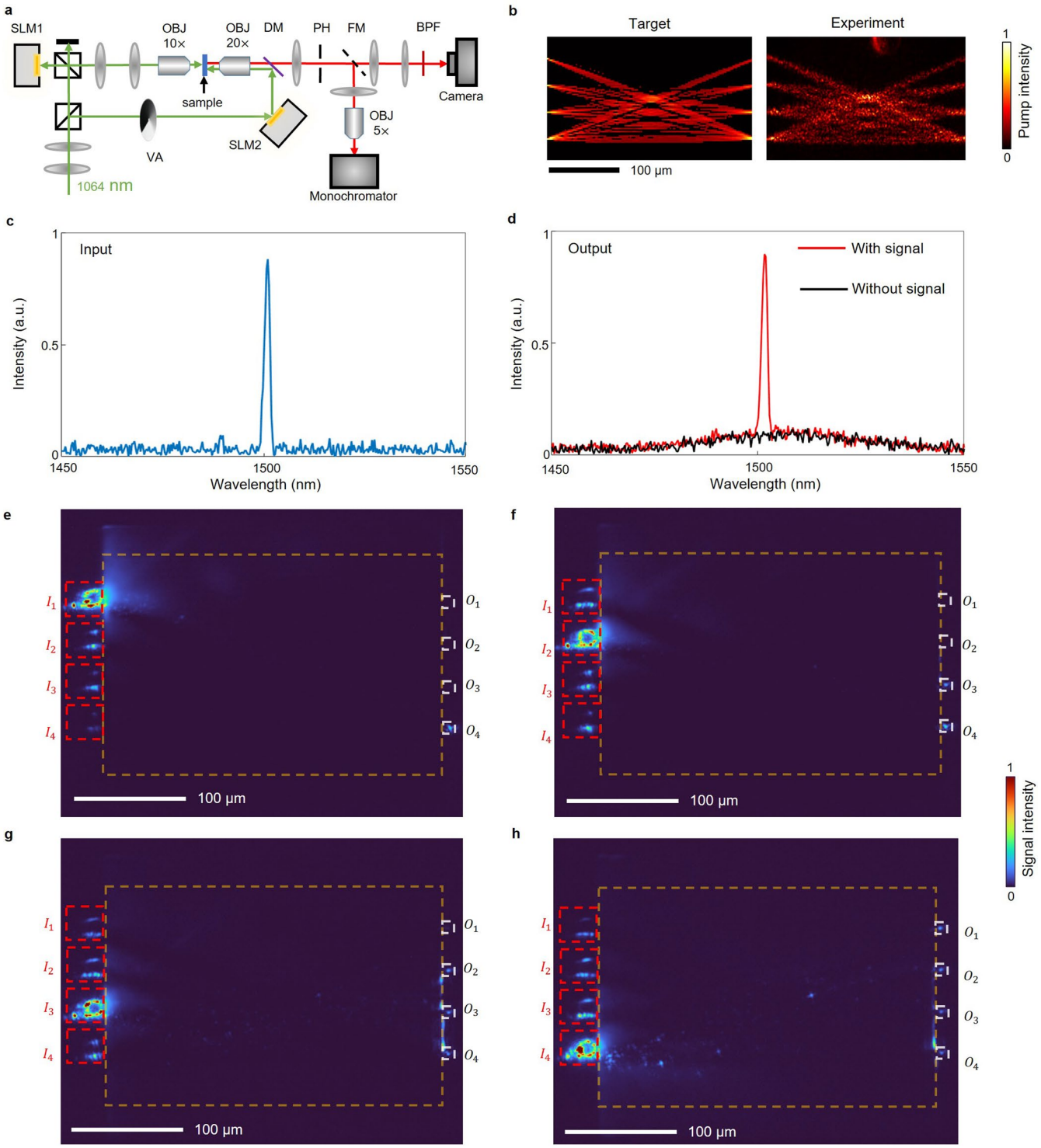
Peer review information *Nature Photonics* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41566-023-01205-0



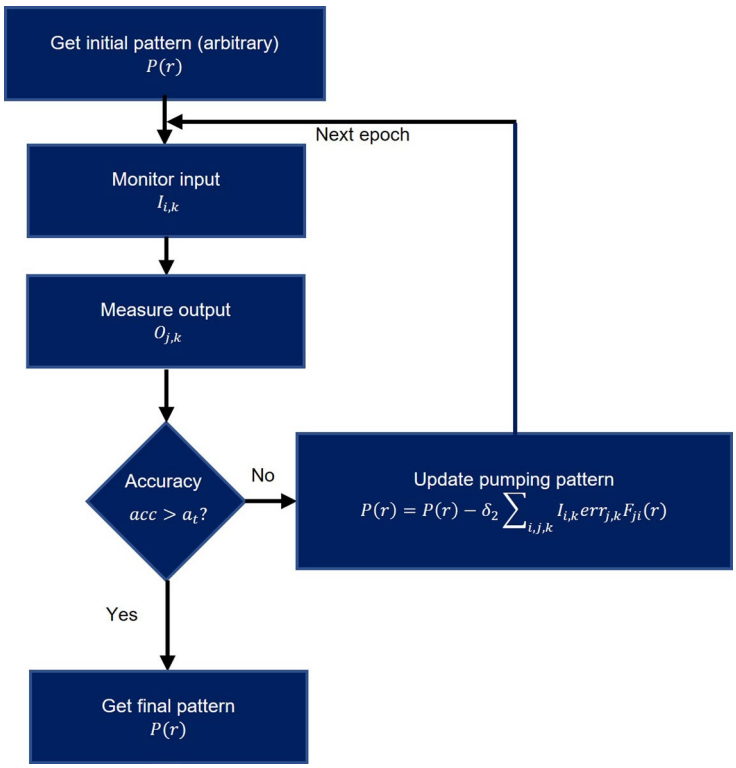
Extended Data Fig. 1 | Illustration of the online algorithm. a, The illustration of the geometry related to input port *i* and output port *j*. **b**, The spatial function *f*(*r*). The isovalue contours are the ellipses with 2 focal points at the point sources of the original (red circle) and adjoint field (blue circle). The contours become denser in the place far from the line connecting 2 ports. The white dashed ellipse shows the range $R(r) \leq R_0 = \frac{5}{4}\lambda_{eff}$, which is used for simulations in Supplementary Video 2.

https://doi.org/10.1038/s41566-023-01205-0



Extended Data Fig. 2 | Transmission measurements. a, Dual-pump optical setup. The 1064 nm pump laser (green trace) is split into two paths for the patterned pumping and the microlaser excitation. The signal around 1500 nm (red trace) is collected by the infrared camera. VA: variable attenuator, OBJ: objective lens, DM: dichroic mirror, PH: pinhole, FM: flip mirror, BPF: band pass filter. **b**, Target pumping pattern and the pattern generated in experiment. The light spot on the top of the experimental pattern is the zero-order beam from SLM, which does not affect the performance as it is far away from the center. **c**, Spectrum collected at an input port. **d**, Spectrum collected at one output port with (red) and without (black) microring lasers excited. **e**-**h** Images with different excitation channels. The red and white boxes mark the position of individual microring laser and the output grating. The yellow box indicates the whole imaginary-index-driven area.

https://doi.org/10.1038/s41566-023-01205-0



Extended Data Fig. 3 | Flow chart of the in-situ training. The initial pattern can be an arbitrary connection between the inputs and the outputs. In each epoch, the inputs and outputs related to all the samples in the dataset are measured. The pumping pattern is updated based on the measurements in the epoch until the accuracy reaches the target.