©PIXABAY/BRIANPENNY

Harnessing Photonics for Machine Intelligence

Hanqing Zhu, Shupeng Ning, Hongjian Zhou, Ziang Yin, Ray T. Chen, Jiaqi Gu, and David Z. Pan

Abstract

The exponential growth of machine-intelligence workloads is colliding with the power, memory, and interconnect limits of the post-Moore era, motivating compute substrates that scale beyond transistor density alone. Integrated photonics is emerging as a candidate for artificial intelligence (AI) acceleration by exploiting optical bandwidth and parallelism to reshape data movement and computation. This review reframes photonic computing from a circuits-and-systems perspective, moving beyond building-block progress toward cross-layer system analysis and full-stack design automation. We synthesize recent advances through a bottleneck-driven taxonomy that delineates the operating regimes and scaling trends where photonics can deliver end-to-end sustained benefits. A central theme is cross-layer co-design and workload-adaptive programmability to sustain high efficiency and versatility across evolving application domains at scale. We further argue that Electronic-Photonic Design Automation (EPDA) will be pivotal, enabling closed-loop co-optimization across simulation, inverse design, system modeling, and physical implementation. By charting a roadmap from laboratory prototypes to scalable, reproducible electronic-photonic ecosystems, this review aims to guide the CAS community toward an automated, system-centric era of photonic machine intelligence.

Index Terms—Photonic computing, AI hardware, cross-layer co-design, electronic-photonic design automation.

Digital Object Identifier 10.1109/MCAS.2026.3688606
Date of current version: 10 August 2026

I. Reframing Photonics for AI Compute: An Emerging Substrate

Machine intelligence [1] has become the defining workload of modern computing, evolving from early deep neural networks that excelled

at pattern recognition [2] to today's foundation models. Underpinned by empirical scaling laws [3], [4], this evolution has shifted the field from isolated algorithmic novelty to systematic scaling of parameters and data, a trajectory that demands vastly more compute. Crucially, this paradigm now extends beyond training to inference: test-time scaling (TTS) [5], [6], [7] unlocks stronger capabilities by allocating additional computation at deployment, a trend accelerated by reasoning-centric models [8], [9]. This shift fundamentally alters the computational landscape: inference costs are transitioning from being linearly proportional to model size to scaling exponentially with the complexity of reasoning and agentic tasks [6], [8]. Consequently, compute availability has emerged as the central determinant of the "supply of intelligence," creating an urgent need to ensure that hardware performance does not become the bottleneck for the pace of machine intelligence.

For decades, the industry relied on Moore's Law [10] and Dennard scaling [11] to deliver near-automatic performance gains to meet the rising compute demand. However, this roadmap has now collided with fundamental physical limits in the single-digit-nanometer regime [12], [13]. As voltage scaling stagnates and quantum effects (e.g., tunneling) intensify, energy efficiency has failed to keep pace with density, making power, not transistor count, the dominant system limiter [14]. This results in the era of *dark silicon* [15], where thermal constraints prevent fully utilizing the available hardware. Consequently, transistor scaling alone is structurally incapable of sustaining the exponential growth of artificial intelligence (AI) workloads [16]. This breakdown motivates alternative computing substrates and architectures that can deliver scalable compute to support the growing intelligence.

While photonics has established itself as the definitive solution to the data-movement wall in interconnects [17], its utility is now expanding into the domain of computation itself [18], [19], [20]. We are witnessing a *resurgence of interest in optical computing,* distinguished by a fundamental strategic shift: rather than mimicking general-purpose digital logic, an approach demanding challenging cascadability and signal restoration [21], the community is increasingly targeting *special-purpose, predominantly analog* accelerators [18], [22], [23].

This pivot harnesses the intrinsic properties of light, such as high bandwidth, low latency, and massive parallelism, to perform efficient linear transformations, a capability that aligns precisely with the workload of modern deep learning. Since foundation models are dominated by dense Matrix-Vector Multiplication (MVM) yet exhibit remarkable tolerance to low-precision computations [24], [25], they are ideally suited to the analog domain. This algorithmic robustness allows optical cores to serve as *high-throughput, specialized primitives for the next generation of machine intelligence.*

Driven by this promise, recent years have witnessed rapid progress in optical computing prototypes, particularly photonic integrated circuit (PIC)-based optical neural networks (ONNs). This evolution is reinforced by massive physical gains: recent milestones include 3.8 TOPS throughput [26], sub-femtojoule energy efficiency [27], and the emerging realization of universal AI acceleration on photonic hardware [23]. While the field has predominantly demonstrated standard architectures, spanning multi-layer perceptron (MLP) [28], convolutional neural networks (CNNs) [26], [29], and spiking NNs (SNNs) [30], [31], it is now actively expanding toward *advanced* AI workloads, e.g., Transformer-style architectures, the foundation-model primitive. Notably, [32] presents one of the first explicit mappings of Transformer computation patterns onto a photonic substrate, redesigning the architecture to accommodate the distinct structure of attention operators

Yet, despite these compelling prototypes, it remains unclear *when* and *how* optical computing delivers a sustained *system-level* advantage. Prior surveys on photonic AI hardware and optical neural networks [20], [33], [34], [35] have largely emphasized photonic devices, ONN implementations, and accelerator prototypes. By contrast, our focus is on a complementary question: how these photonic innovations translate into deployable system benefits once mixed-signal interfaces, memory hierarchy, control overheads, workload mapping, and programmability constraints are explicitly considered. As a result, reported "optical-core" metrics do not directly translate into deployable performance, and comparisons across architectures and workloads remain difficult to interpret.

To close this gap, Section II establishes a *system-level benchmarking* baseline using cross-layer simulation (e.g., our SimPhony framework [36]). By modeling the full heterogeneous electronic-photonic datapath, from photonic tensor cores (PTCs) to digital-to-analog (DAC)/analog-to-digital(ADC) converters, memory, and laser power, we quantify where photonics is competitive

Hanqing Zhu, Shupeng Ning, Ray T. Chen, and David Z. Pan (corresponding author) are with The University of Texas at Austin, Austin, TX 78758 USA (e-mail: dpan@ece.utexas.edu).
Hongjian Zhou, Ziang Yin, and Jiaqi Gu (corresponding author) are with Arizona State University, Tempe, AZ 85287 USA (e-mail: Jiaqi.Gu@asu.edu).

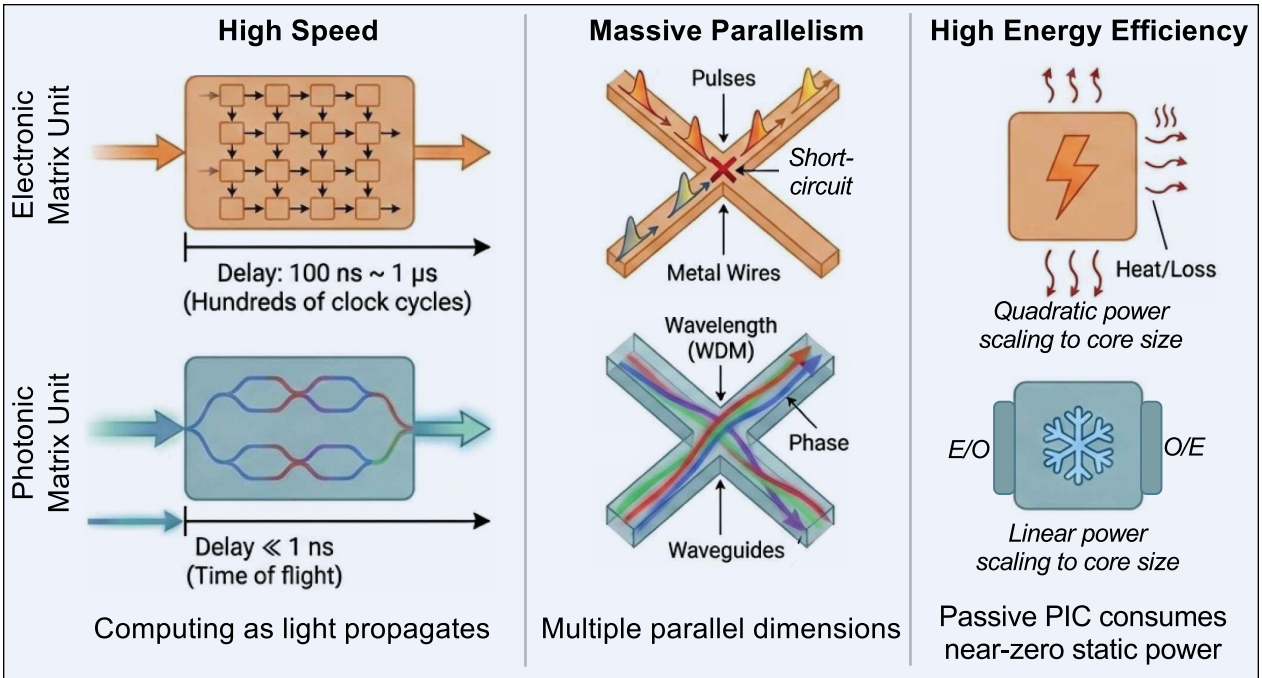

Figure 1. Physical advantages of photonic computing: ultra-low latency via RC-free propagation, high bandwidth density via multiplexing, and superior energy efficiency.

today under realistic assumptions. This benchmarking lens makes the "system tax" explicit and enables fair comparisons across representative PTC families and NN operators (e.g., linear layers and attention mechanisms).

Beyond point comparisons, however, a benchmark alone does not answer the forward-looking question: *what must improve for photonic computing to scale with rapidly evolving AI workloads?* Section III performs a *scaling analysis* to extract the critical dimensions that govern end-to-end efficiency, including area, parallelism (spatial/spectral/temporal), bit precision, and the degree to which interface costs can be amortized by reuse. Guided by these simulation-derived "pressure points," we organize the literature into a bottleneck-driven taxonomy, highlighting how recent efforts aim to scale photonic systems along the dimensions that matter most at the full system level.

Finally, scaling photonic AI from prototypes to deployable heterogeneous EPICs requires *full-lifecycle electronic-photonic design automation (EPDA).* Section IV reviews emerging EPDA capabilities across design stacks, including AI-assisted device simulation and inverse design, photonic-electronic circuit-level co-simulation, system-level modeling, and layout synthesis automation. We identify the remaining gaps and outline an EPDA roadmap toward future infrastructures to enable scalable, reproducible, and efficient design of photonic AI systems.

Roadmap and Organization. This survey arrives at a critical inflection point, as traditional electronic scaling struggles to meet the exponential demand for compute while integrated photonics matures into a viable heterogeneous accelerator. Unlike prior surveys that emphasize either device physics or isolated circuit techniques, we adopt a system-scalability lens: examining how photonics alters the computing landscape, when it provides a distinct advantage, what limits it, and what toolchains are required to make it practical.

The remainder of this paper is organized as follows:

- **Section** II establishes a system-level benchmarking baseline using cross-layer simulation, quantifying the regimes where photonics is competitive once mixed-signal interfaces, memory traffic, and optical link-budget constraints are included.
- **Section** III extracts the critical scaling dimensions that govern end-to-end performance and uses these simulation-derived pressure points to categorize photonic accelerator efforts into a bottleneck-driven taxonomy.
- **Section** IV reviews the EPDA stack required to scale photonic AI from prototypes to deployable EPICs, including modeling, verification, physical design, and calibration-aware co-simulation.

II. Quantifying Photonic Advantage: From Physical Promise to System-Level Benchmark

While the intrinsic physical advantages of optical computing, such as high bandwidth and low latency, are well understood, translating these attributes into deployable AI performance requires rigorously accounting for system-level constraints. In realistic hybrid architectures, the optical core never operates in isolation; its realizable throughput and efficiency are fundamentally governed by both the photonic devices and the electronic periphery, specifically the data converters (ADCs/DACs) and memory interfaces, required to sustain them.

In this section, we deconstruct the photonic advantage by tracing the signal chain from physics to system layers:

- **Section** II-A **(The Physics):** We revisit the **physical primitives,** isolating the specific properties of light (e.g., parallelism and passivity) that make it structurally efficient for linear algebra.
- **Section** II-B **(The System):** We conduct **system-level modeling and benchmarking.** By modeling the full datapath including mixed-signal overheads, we quantify photonic advantages and identify key system bottlenecks.

A. Physical Primitives: Why Light Can Be Efficient

Despite the rapid evolution of architectures, from CNNs to Transformers, the underlying computational backbone remains invariant. Whether computing convolutions, linear projections, or attention scores, the vast majority of modern AI inference relies on a single fundamental primitive: **Matrix-Vector Multiplication.** Below, we outline the primary physical attributes of light that allow it to execute these dense linear transformations with ultra-low latency, massive parallelism, and energy efficiency, beyond what is readily achievable with conventional electronics.

1) Low Latency via RC-Free Optical Propagation

A common *misconception* is that optical computing is faster simply because light travels "faster." In practice, the group velocity in silicon/silicon-nitride waveguides is typically 0.3–0.5 c, comparable to propagation in well-designed electrical transmission lines [37], [38]. The advantage is instead *how* delay scales with connectivity.

In dense CMOS interconnects, long wires and high fanout incur resistive-capacitive (RC) delays; the effective delay grows superlinearly with wire length in distributed RC regimes, and mitigating this requires repeater insertion, which adds area and power [39], [40]. Optical waveguides, in contrast, are "RC-free," i.e., the delay depends almost exclusively on the geometric path length (scaling linearly) and is effectively independent of capacitive loading. This enables optical signals to traverse centimeter-scale chips in sub-hundred-picosecond flight times, offering a critical latency advantage for the highfanout broadcasting and global data distribution required by modern neural networks.

2) High Bandwidth Density via Multiplexed Channels

A frequent critique of photonics is the "density gap." In modern CMOS, transistors scale to ~10^{10} cm^{-2}, while fabricated photonic components are typically orders of magnitude larger due to the diffraction limit [18]. However, *raw component density is not the right proxy for throughput.* In reality, the "truth" of photonic scaling lies in its ability to exploit dimensions of parallelism inaccessible to electronics.

For optical communication and computing, information is encoded onto electromagnetic waves that propagate through fibers or on-chip waveguides at carrier frequencies in the terahertz-petahertz range. Since they are not constrained by RC delay and Joule heating that confine practical electrical signaling rates to at most a few gigahertz (GHz) [41], [40], optical signals inherently afford a wider usable bandwidth. More importantly, the neutrality and bosonic nature of photons allow multiple optical modes to coexist within a shared waveguide without mutual exclusion or Coulombic repulsion [42], [43]. Orthogonal channels such as wavelength-division, polarization, spatial modes, and temporal encoding can be densely multiplexed with negligible crosstalk, yielding massive parallelism within a bulky photonic footprint [43], [44].

This multiplexing capability maps naturally onto the intrinsic fan-in and fan-out patterns of MVMs, alleviating the routing congestion, signal interference, and limited parallelism that increasingly constrain advanced CMOS platforms.

3) High Energy Efficiency via Linear Power Scaling

Photonic computing can offer superior energy efficiency by avoiding the resistive heating that fundamentally limits electronic AI accelerators. In CMOS technologies, dynamic power scales as $P_{\text{dyn}} \propto CV_{\text{dd}}^2 f$, and increases in operating frequency f typically require proportional increases in supply voltage V_{dd} to maintain switching speed. This voltage-frequency coupling leads to superlinear power scaling at high-performance operating points [41]. Additional frequency-dependent losses, such as skin-effect-induced current crowding in metal interconnects, further exacerbate ohmic heating [38].

By leveraging capacitive electro-optic mechanisms, EPIC achieves high modulation speeds while maintaining near-zero static power, avoiding the resistive leakage and thermal dissipation inherent to active electronic

transistors. Since the core matrix operations are subsequently performed via passive wave propagation, the dynamic power consumption is fundamentally decoupled from the computational complexity of the matrix interior. Unlike electronic processors, which suffer from quadratic energy growth with matrix size dictated by active switching for every arithmetic operation, EPIC confines energy expenditure to the I/O interfaces, thereby exhibiting a linear scaling trajectory with both operating frequency and input dimension. This advantage is further strengthened by passive photonic paradigms, including architectures based on phase-change materials (PCM) [45], diffractive structures [29], [46], and metasurfaces [47], which reduce electrical overhead by implementing linear transforms directly via propagation.

B. System-Level Benchmark: A Cross-Layer Simulation via SimPhony [36]

While photonic primitives offer intrinsic physical advantages, deployable utility is ultimately determined by the end-to-end *system tax* imposed by mixed-signal interfaces and memory, rather than isolated device metrics.

From an architecture standpoint, two metrics are especially diagnostic:

- **Effective Energy Efficiency (TOPS/W):** end-to-end throughput per wall-plug power, governing thermal feasibility and operating cost at scale.
- **Compute Density (TOPS/mm^2):** sustained throughput per integrated area, reflecting how much performance can be delivered per chip footprint once peripheral overheads are included.

Figure 2 maps representative photonic prototypes against state-of-the-art graphics processing units (GPUs) on the efficiency-density plane. The plot highlights rapid progress: successive demonstrations are pushing toward the upper-right, and several photonic tensor-core designs report regimes of tera operations per second per Watt (TOPS/W) and/or TOPS/mm^2 that exceed typical GPU operating points.

However, the reported photonic points span orders of magnitude, and even superficially similar photonic approaches can appear far apart. This spread is primarily driven by non-unified evaluation processes and assumptions, workload choice, precision and utilization

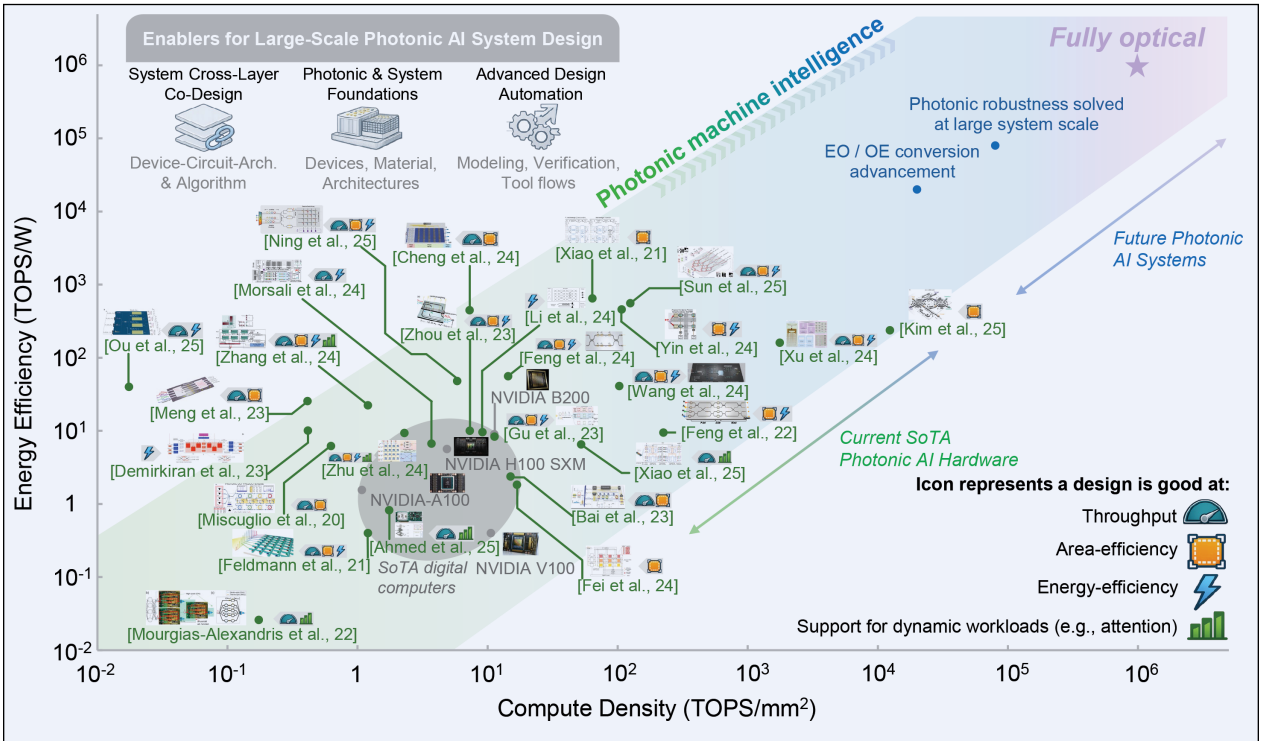

Figure 2. System-level energy efficiency versus compute density comparison between silicon-photonic accelerators and digital GPUs. The axes report effective energy efficiency (TOPS/W) and compute density (TOPS/mm^2) under reported or derived system-level assumptions. GPU points correspond to NVIDIA V100 [48], A100 [49], H100 SXM [50], and B200 [51], using vendor-reported peak INT8 performance and die area. Photonic points are compiled from prior silicon-photonic accelerator works [23, 28, 29, 32, 45, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72] and are annotated to reflect which scaling dimensions each design emphasizes (area efficiency, energy efficiency, robustness at scale, and dynamic-workload capability, such as attention).

assumptions, and whether key system taxes are consistently included (e.g., laser power under realistic insertion loss, ADC/DAC energy at the required bandwidth, and the control/calibration overheads needed to maintain operating points). Without a unified benchmarking methodology, it is difficult to isolate true bottlenecks or predict how architectural scaling will translate into deployable advantage.

1) Simulation Tool: a Cross-Layer Modeling Framework

To benchmark photonic AI accelerators under realistic constraints, we leverage SimPhony [36], a cross-layer modeling framework capable of simulating heterogeneous electronic-photonic systems from the device level up to architecture. Rather than relying on isolated figures of merit for the photonic core, SimPhony evaluates end-to-end performance by explicitly modeling the full signal chain: the photonic compute fabric, the mixed-signal interfaces (drivers, modulators, trans-impedance amplifiers (TIAs), ADCs/DACs), and memory required to sustain high-bandwidth operation.

As illustrated in Fig. 3, the framework composes a parametric system model from physical device and circuit building blocks. It then (i) generates optics-specific dataflows exploiting wavelength, time, and spatial parallelism; (ii) tracks memory traffic through multi-level buffers and off-chip memory; (iii) enforces optical link-budget constraints to determine the required laser power; and (iv) produces layout-aware area estimates and a granular energy breakdown (laser, modulation, readout, conversion, and memory). This holistic accounting

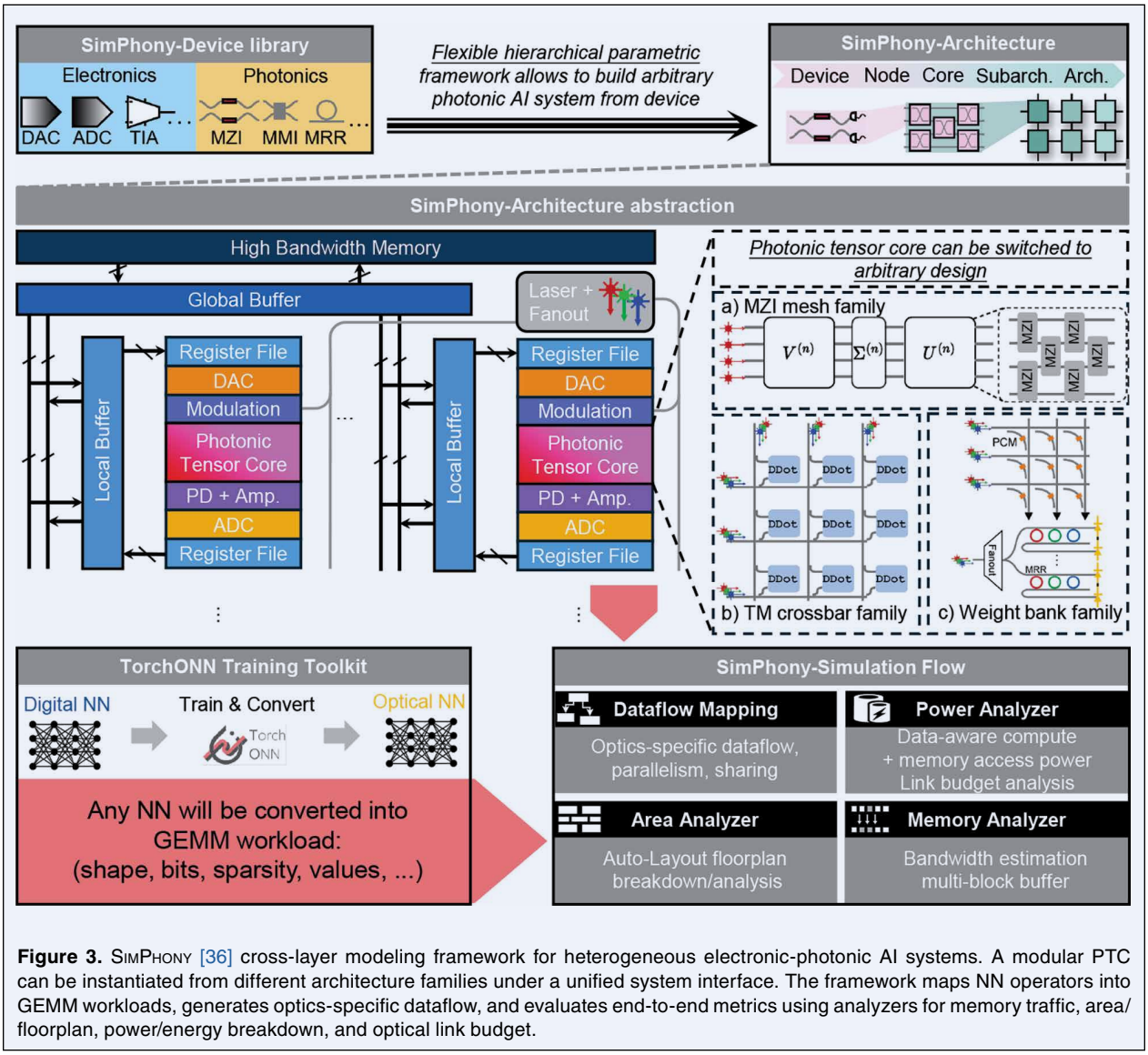

Figure 3. SimPhony [36] cross-layer modeling framework for heterogeneous electronic-photonic AI systems. A modular PTC can be instantiated from different architecture families under a unified system interface. The framework maps NN operators into GEMM workloads, generates optics-specific dataflow, and evaluates end-to-end metrics using analyzers for memory traffic, area/floorplan, power/energy breakdown, and optical link budget.

is essential because while many photonic architectures exhibit favorable *core-level* metrics, their deployable efficiency is strictly governed by "system taxes" that dominate at scale:

- **Memory Delivery and Utilization:** Bandwidth limits and the energy cost of data movement can throttle the optical core, often outweighing the compute energy itself.
- **Optical Loss → Laser Power:** Wall-plug laser power is strictly governed by the optical link budget, scaling exponentially with critical-path insertion loss and the target signal-to-noise ratio (SNR).
- **EO/OE Interfaces:** Data converter (DAC/ADC) energy scales linearly with sampling rate but exponentially with bit resolution.
- **Photonics-Aware Mapping:** Multi-dimensional parallelism only yields system benefits if the chosen dataflow effectively amortizes conversion and movement costs, rather than simply multiplying the number of required physical interfaces.

More detailed modeling formulations can be found in [36].

2) Simulation Setup: Photonic Tensor Core and Workload

The architecture-workload interaction. System-level energy and throughput are not intrinsic properties of a photonic core; rather, they emerge from the interaction between the *tensor core topology* (specifically, how it encodes weights) and the *workload characteristics* (specifically, how frequently weights change). This coupling is critical in the modern AI landscape, which is transitioning from static-weight workloads (e.g., CNNs) to dynamic, weight-free mechanisms (e.g., Attention) that require frequent operand refreshing. To capture this spectrum, we select representative architectures and workloads that span the design space from high-reuse static execution to reuse-free dynamic execution.

Representative PTC families. We categorize the diverse landscape of photonic AI systems into three families based on their physical weight-encoding mechanisms. We select one representative design for each to be evaluated under identical conditions (8 × 8 core, 12 wavelengths, 8-bit, 5 GHz):

- ***Mach-Zehnder Interferometer (MZI) Mesh Family (Coherent/Static-only):*** Exemplified by Clements-style arrays [28], [58], this family represents the coherent computing paradigm. It relies on unitary transformations and typically assumes slowly tunable phase shifters, making it highly efficient for static weights but challenging for rapid reconfiguration. We use the *coherent nanophotonic circuit* [28] as the exemplar.
- ***Weight-Bank Family (Incoherent/Both Dynamic and Static):*** Typified by broadcast-and-weight architectures like microring resonator (MRR) banks [73] and PCM arrays [45]. These designs offer high area density but, like MZI meshes, generally leverage static weight reuse to amortize the cost of precise thermal/PCM tuning. We use the *MRR weight bank* [52] as the representative design.
- ***Time-Multiplexed Crossbar Family (Both Dynamic and Static):*** Represented by engines like Lightening-Transformer [32] and TeMPO [56]. These architectures are explicitly designed for dynamic dataflows; they employ high-speed modulators for all operands and exploit time-multiplexing to minimize interface overheads during rapid weight updates. We use *Lightening-Transformer* [32] as the representative design.

Workloads: static linear versus dynamic attention. Early optical accelerators largely targeted CNN/MLP-style layers where a *trained weight matrix is reused* across many input activations, making weight-stationary mappings and slower/low-power weight-tuning mechanisms plausible. In contrast, modern Transformers are dominated by *attention micro-kernels* whose heaviest general matrix multiplications (GEMMs) include *activation-activation* products (e.g., QK^{T}), where both operands are token-dependent at runtime. This reduces reuse and stresses the very subsystems that often dominate the system tax, conversion bandwidth, modulation/update rate, and memory traffic, thereby invalidating assumptions that are benign under static layers. To capture this shift with a fair, compute-matched comparison, we evaluate two GEMM-equivalent operators with identical multiply-accumulate (MAC) counts:

- **Dynamic (Attention):** a representative self-attention micro-kernel with sequence length $S = 1024$ and embedding dimension $D = 512$, focusing on the QK^{T} product. Concretely, we model a batch of 4 query vectors multiplying a key matrix of width S, yielding 4×512×1024 MACs.
- **Static (Linear):** a compute-matched batch-4 linear projection of shape 512→1024, which also requires 4×512×1024 MACs. This serves as the weight-stationary baseline (representative of linear/conv projections after lowering) to contrast against attention-style dynamics.

3) Simulation Results and Insights

Figure 4 presents the simulated end-to-end energy breakdown across the three representative PTC families, decomposed into key components: modulation (drivers),

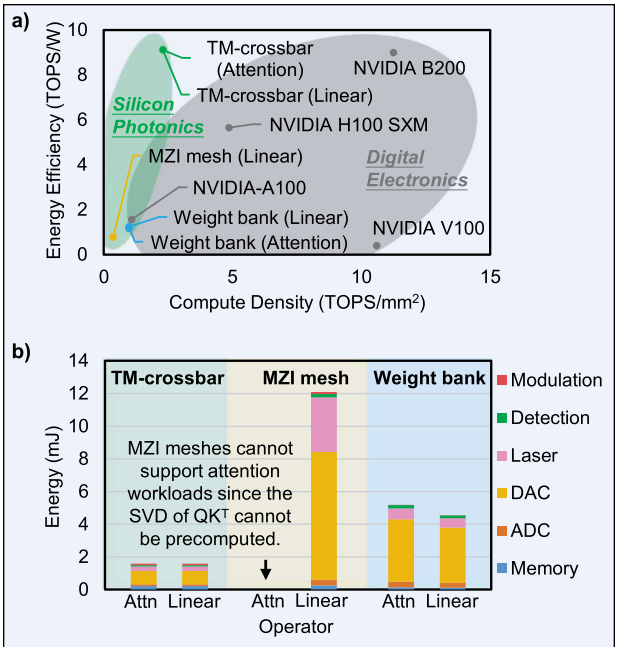

Figure 4. a) Energy efficiency and compute density comparison between photonic and digital electronic hardware. For a fair comparison, GPU performance is reported based on INT8 precision. b) End-to-end energy breakdown comparison across 3 representative PTC families under dynamic attention and static linear workloads with matched computations. All results are simulated using SimPhony [36]. *Arch Setting:* Total of 8 8×8 PTCs with 12 wavelengths, 8-bit precision, and a 5 GHz clock rate. *Workload:* Attention uses sequence length $S = 1024$ and embedding dimension $D = 512$; Linear layer projects hidden dimension from 512 to 1024 with a batch of 4. MZI mesh is incompatible with attention, thus omitted. Font sizes have been enlarged to improve overall readability.

detection (readout), laser input, data conversion (ADC/DAC), and memory. Furthermore, it benchmarks these architectures against state-of-the-art (SOTA) electronic baselines, ranging from the NVIDIA A100 and H100 to the emerging B200.

These results highlight three critical insights:

(1) Photonic Competitiveness Against SOTA Electronics. The most distinct takeaway is the performance of the **TM-Crossbar** architecture. Explicitly designed to mitigate system taxes, specifically data movement and cross-domain conversion, this design demonstrates a competitive position on the density-efficiency Pareto frontier relative to the NVIDIA A100. Notably, it achieves superior energy efficiency even compared to the newer NVIDIA B200. This validates the trajectory shown in Fig. 2: when architectures are optimized to minimize peripheral overheads, photonics retains its fundamental advantage, suggesting even greater potential as designs move toward fully optical, constraint-free implementations.

(2) System Efficiency Limits Beyond the Optical Core. In contrast, MZI meshes and MRR weight banks, which have been central to ONN research, can exhibit substantial *system-level overheads* when evaluated end-to-end on modern workloads. As illustrated by the breakdown in Fig. 4(b), overall efficiency in these designs is often constrained by the surrounding system stack: high-speed ADC/DAC conversion, data movement, and programming/control overheads. These costs become particularly prominent for dynamic and communication-heavy workloads (e.g., attention-style operations) where frequent reconfiguration and I/O amplify peripheral energy.

(3) The Rigidity of MZI Meshes in Dynamic Workloads. A critical limitation emerges regarding workload versatility: MZI meshes are fundamentally ill-suited for the Attention mechanisms. The MZI topology relies on a *static* linear transform programmed via Singular Value Decomposition (SVD) and phase decomposition. In Attention, however, the effective operator is input-dependent and changes at every token step with no static weights. Since the SVD and phase decomposition cannot be precomputed and reprogramming an $N \times N$ mesh requires updating $\mathcal{O}(N^2)$ thermal or electro-optic phase shifters at token-rate timescales, MZI meshes are thermally and control-limited to static linear projections. Consequently, they lack the opportunity to scale to the dynamic workloads that define the current AI era.

Outlook. It is important to note that the results in Fig. 4 represent "baseline" implementations of these topologies. The flourishing landscape of high-performance designs shown previously in Fig. 2 is populated by works that explicitly target these identified bottlenecks, introducing novel optimizations in area, energy efficiency, and flexibility. In the following section, we categorize these recent advancements by mapping them to the specific scaling dimension they improve.

III. Scaling Photonic AI: From Bottlenecks to Solutions

A. Identifying Bottlenecks via Key-Dimension Projections

In Section II-B3, we identified the "system tax" imposed by frequent data conversions and the rigidity of PTCs as primary barriers to efficient photonic computing. Taking a broader perspective, we first examine how these bottlenecks manifest across different layers of the photonic AI technology stack and summarize representative mitigation strategies. Building upon this qualitative foundation, we utilize our simulation framework to

sweep a broader set of architectural parameters, specifically *operating frequency, bit precision, wavelength parallelism, and tensor core size,* to determine how to effectively fuse greater computational power onto a single chip and guide future scaling strategies.

We map these parameters to their physical impact on system performance. Tensor core size dictates the *spatial density* of the compute fabric. Meanwhile, wavelength parallelism and operating frequency serve as complementary drivers of *effective throughput.* Wavelength parallelism acts as a spectral multiplier, increasing the number of concurrent input channels supported on the same physical chip, which scales throughput analogously to increasing the temporal clock frequency. Critically, by sweeping bit precision, we assess the feasibility of supporting high-fidelity arithmetic within this mixed-signal paradigm.

Figure 5 summarizes the simulated system-level energy breakdowns and energy efficiency trends across three PTC families, using the attention workload as a unified baseline. The results reveal distinct scaling behaviors across these dimensions:

Fusing Computational Density and Throughput: As illustrated in Fig. 5(a) and (b), increasing the computational density, whether spatially via larger tensor cores or spectrally via dense wavelength division multiplexing (DWDM), generally yields improved energy efficiency. By increasing the number of wavelength channels, the system processes more data in parallel within the same footprint, effectively amortizing fixed static power overheads (such as laser biasing and thermal control) across a higher aggregate throughput.

Frequency Saturation: While scaling the operating frequency [Fig. 5(c)] also improves throughput, the efficiency gains exhibit diminishing returns. At higher frequencies, the dynamic power consumption of high-speed drivers and readout circuits begins to scale (super-)linearly with the increased data rate, eventually neutralizing the efficiency benefits.

The Bit Precision Wall: Most notably, Fig. 5(d) highlights a severe limitation: *brute-force scaling of electronic bit precision is unsustainable.* As precision increases, energy consumption skyrockets while efficiency (TOPS/W) plummets. This is corroborated by the energy breakdown across all swept dimensions, where electro-optic conversion, dominated by DACs and optical modulation, emerges as the primary contributor to total energy consumption, consistently outweighing laser power and data movement. This confirms that the energy cost of high-resolution A/D conversion creates a fundamental "precision wall" for analog photonic computing.

B. Categorizing Photonic AI Progress via Bottleneck-Driven Analysis

1) Area Efficiency: Device and Architectural Density
Considering the relatively large footprint of optical components, device-level optimization can improve computing density by shrinking the overall area of EPICs. One device-level approach is to use multi-operand photonic primitives that collapse a length- k dot product into a single physical unit, enabling accumulation directly in the physical domain concurrent with the electro-optic mapping $x_{\text{out}} = f\left(\sum_{i=1}^{k} g(w_i, x_i)\right)$, where $g(\cdot)$ captures operand encoding and $f(\cdot)$ is the device transfer function. This strategy increases effective fan-in with a footprint comparable to that of single-operand optical synapses.

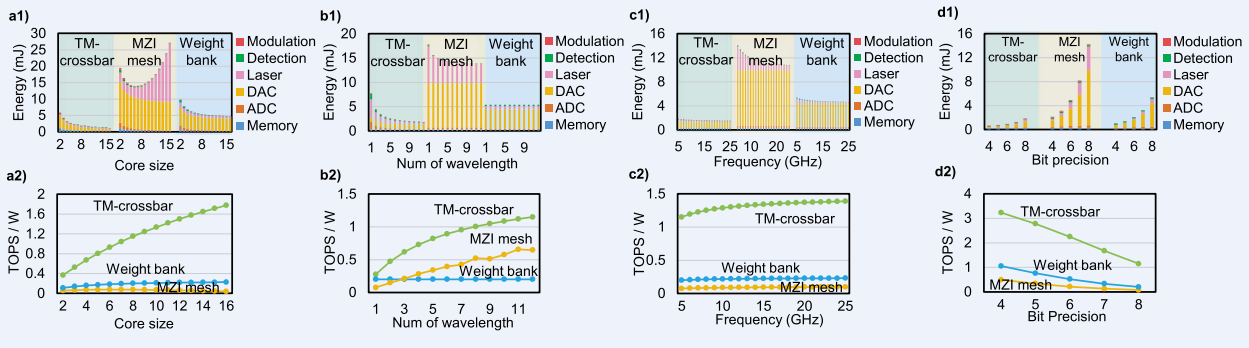

Figure 5. Impact of parameter scaling on energy breakdown and power efficiency across representative designs on **a single tensor core.** We evaluate three canonical photonic tensor core designs, the dynamic time-multiplexing crossbar [32], the Clements MZI mesh [28], and the weight bank [52], using simplified and standardized models. Energy breakdowns (left panels) and energy efficiency in TOPS/W (right panels) are reported as functions of a) core size, b) number of wavelengths, c) bit precision, and d) operating frequency. The baseline configuration is one 8×8 core, 12 wavelengths, 8-bit precision, and 5 GHz operation. All results are obtained using SimPhony [36] to ensure a fair and consistent comparison across architectures. All scaling explorations are performed using a 4-batch linear projection (512→1024) to ensure all PTC families can be evaluated.

Representative realizations include multi-operand MZIs, as well as MRRs and multimode inference devices (MMI) [70], [74], [75], [76], [77].

However, a key caveat is that multi-operand operation couples operands through the modulator nonlinearity, so individual contributions are less separable, which complicates calibration and training. At the same time, this nonlinearity can also be beneficial, since it can serve as an on-chip activation and reduce electrical post processing. Experiments with microring-based multi-operand neurons show that this nonlinear response can be trained to improve representational capacity, enabling comparable or better accuracy with fewer active modulators, which directly reduces both parameter count and tunable footprint [74], [75].

Beyond these device-level primitives, diffractive optical neural networks (DONNs) and related hybrid diffractive systems improve area efficiency through coordinated optimization across the device, circuit, and architecture levels. At the device level, modern nanofabrication enables metastructures and compact interferometric elements to serve as dense optical building blocks [46], [47], [71]. At the circuit level, these components are organized into cascaded passive propagation stages that perform computations via diffraction and wavefront shaping [78]. From an architectural perspective, this paradigm tailored to the strengths of optics enables parallel Fourier- and convolution-like primitives within an ultra-compact footprint, supplemented by optional sparse active tuning. Notably, this cross-layer codesign achieves linear scalability in both footprint and energy consumption relative to the input dimension, thereby circumventing the quadratic complexity overhead intrinsic to fully connected interferometric meshes. Recent hybrid diffractive–interference architectures further push efficiency while supporting larger-scale workloads [29], [79].

However, a primary limitation of passive diffractive architectures is that their weight configuration is fixed upon fabrication, rendering them task-specific and generally incapable of realizing arbitrary matrix transformations on demand. Consequently, many architectures adopt hybrid strategies that combine diffractive propagation with programmable structures. In addition, partially reconfigurable diffractive designs have been demonstrated using post fabrication tuning elements such as microheater arrays [72], [80]. Nevertheless, realizing fully programmable, universal weight matrices in purely diffractive systems remains challenging, highlighting a trade-off between the compact efficiency of passive diffractive computing and the flexibility enabled by active photonic components.

2) Throughput Scaling: Bandwidth and Multiplexing

To optimize ONNs for high-throughput computing, a prominent strategy from the circuit and architecture perspectives is to expand parallelism by exploiting multiple orthogonal optical degrees of freedom, such as wavelength, time, and spatial multiplexing. By encoding data across these distinct dimensions, the architecture allows a substantially higher volume of MAC operations to be executed "in flight" during a single propagation and detection cycle. An example is the universal optical vector convolution engine by Xu et al. [26] which concurrently leverages an integrated microcomb source to exploit temporal, wavelength, and spatial multiplexing, achieving >10 TOPS of computational throughput. More broadly, hyper-multiplexing architectures integrate space–time–wavelength parallelism to reach trillions of operations per second while requiring only $\mathcal{O}(N)$ modulator devices to scale to $\mathcal{O}(N^2)$ operations per cycle [68]. Related multiplexing-based approaches have also been reported [45], [81], [82]. Along a related direction, Yin et al. [83] demonstrated an ONN that combines WDM with mode-division multiplexing, using orthogonal spatial modes as additional parallel channels to further boost on-chip throughput without proportionally increasing the device count.

Complementary to exploiting orthogonal multiplexing dimensions, throughput can also be elevated at the device level by pushing the limits of electro-optic bandwidth, thereby enabling time-serial streams to be processed at higher symbol rates. While typical integrated PTCs report operation bandwidths in the GHz range, recent advancements in devices and material platforms have facilitated much more aggressive scaling. Ou et al. [68] and Lin et al. [84] demonstrated this potential by implementing a fully integrated tensor core based on thin-film lithium niobate (TFLN), which accommodates in-situ weight updates at speeds exceeding 40 GHz and a flexible fan-in.

3) Energy Efficiency: Mitigating the Conversion Tax

As noted above, in addition to area constraints, another major factor limiting ONN scalability is the energy overhead of electro-optic interfaces, as well as the practical challenges of calibration, control, and hardware complexity. While photonic platforms offer high-bandwidth processing, the power consumption required for driving modulators and performing data access and conversion (DAC/ADC) can dominate the total system energy budget, eroding the efficiency gains of optical computing. Consequently, a critical design objective is to minimize the information flux across the E-O boundary per inference, thereby restricting programmability to a sparse set of high-speed, low-power control elements.

One effective strategy from a cross-layer co-optimization perspective is to compress programmability into a compact active front end while preserving a largely passive photonic computing backbone. Returning to the context of diffractive ONNs, the diffractive backbone is typically passive and difficult to reconfigure at scale. Wang et al. [79] addressed this by placing a low-dimensional active modulation unit upstream of the passive diffractive cells. By injecting inputs through this compact active layer into the static diffractive volume, target transformations are synthesized via iterative time-domain updates. This approach strategically decouples the massive computational throughput from the control overhead, and shifts the programmability burden from a power-hungry, fully active spatial backbone to a minimal, rapid control layer. This significantly lowers the aggregate E-O interface cost while preserving the high-throughput advantages of diffractive propagation.

At the circuit level, another route to improving energy efficiency is to exploit the redundancy of modern DNNs so that fewer tunable parameters need to be physically encoded and controlled. Since high-accuracy solutions often reside in low-dimensional subspaces of the full weight space [85], [86], [87], compressed parameterizations can directly translate into fewer active photonic degrees of freedom and fewer interface channels, which in turn lowers power on E-O modulation. Feng et al. [54] proposed the optical subspace neural network (OSNN), factorizing weights as $W = B\Sigma P$ with diagonal Σ and implementing the unitary transforms B and P using hardware efficient butterfly meshes. By avoiding a fully programmable MZI mesh, the approach cuts the number of actively tuned elements by up to 7× while achieving 94.16% accuracy on MNIST. Similarly, Ning et al. [55] imposed a block circulant constraint to realize structured compression. Using a compact MRR based crossbar with WDM, the compression is embedded directly into the PIC topology, enabling up to a 75% reduction in trainable parameters and control overhead with negligible accuracy degradation across multiple tasks. Therefore, these results show that energy-efficient photonic acceleration can be approached across multiple levels, either by confining active programmability to a sparse front end or by reducing the number of actively controlled photonic degrees of freedom through structured parameterization.

While we discusses methods to mitigate this conversion tax by trading off precision for energy efficiency, supporting true high-precision computation remains critical for broader adoption. To bypass brute-force precision scaling, recent work increasingly adopts structural and hybrid co-design strategies that decompose high-precision arithmetic into multiple low-precision optical operations. For example, bit-slicing techniques can be combined with block floating-point (BFP) formats, where the photonic core executes low-bit mantissa multiplications while digital electronics handle accumulation and shared exponents [88]. Similarly, residue number system (RNS) formulations distribute computation across parallel, low-precision channels [89], [90]. By structurally mapping complex data formats onto low-resolution hardware, these approaches can recover near-FP32 accuracy while strictly utilizing energy-efficient, low-bit converters.

Considered local E-O interfaces, another route to mitigating the conversion tax is to reduce the energy required for state retention by incorporating non-volatile photonic devices or analog memory. On one hand, researchers seek to eliminate the static power dissipation of conventional optical components, which typically rely on volatile electro-optic or thermo-optic mechanisms to maintain their states. Non-volatile devices based on phase-change materials (PCMs) [45], [91], [92], ferroelectrics [93], or latching MEMS [94], [95] can retain information without a continuous power supply. However, existing non-volatile technologies, especially PCM-based implementations, still face endurance limitations that depend on the underlying programming mechanism, whether electrical, electrothermal, or optical. This raises concerns regarding their suitability for write-intensive workloads such as training [96], although recent efforts are actively addressing these durability constraints [97].

Additionally, complementary efforts focus on integrating foundry-compatible analog electronic memories, such as Dynamic Electro-Optic Analog Memory (DEOAM) [98], to relax the conversion bottleneck. In this approach, a capacitor is paired directly with an MRR to locally hold the drive voltage (data) on the modulator. Acting as a sample-and-hold element, this local memory allows DACs to be time-multiplexed across columns of devices rather than requiring a dedicated DAC for every MRR. The system updates weights row-by-row while the analog memory retains the signal on the MRRs, thereby reducing the DAC count from quadratic (N^2) to linear (N) complexity and significantly relaxing energy and bandwidth constraints. These approaches show that energy-efficient photonic acceleration can also be improved through device and local-interface innovations, either by eliminating static holding power through non-volatile state retention or by reducing repeated conversion overhead through analog storage.

4) Dynamic Workload Adaptation

To elucidate the limitations of existing architectures, we first distinguish between *static inference* and

dynamic workloads. Conventional inference relies on fixed weights, whereas dynamic workloads, most notably the attention mechanism in Transformers, require General Matrix Multiplication (GEMM) in which *both operands vary at runtime*.

A canonical example is self attention:

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{\text{QK}^\top}{\sqrt{d_k}}\right)V, \tag{1}$$

where the query (Q), key (K), and value (V) matrices are token dependent and generated on the fly. Unlike conventional layers, the QK^T operation induces dynamic, full range, all to all interactions between runtime generated operands.

More generally, we aim to support matrix multiplication of the form

$$\mathbf{Y} = \mathbf{WX} \tag{2}$$

where both **W** and **X** are dynamic, updated at runtime, and full range, containing both positive and negative values. These requirements impose stringent constraints that most legacy optical designs are fundamentally ill-equipped to satisfy.

Challenge 1: Reconfiguration Latency in Coherent MZI Meshes. Prior coherent architectures, such as Mach-Zehnder Interferometer (MZI) meshes [28], [58], struggle to efficiently support dynamic GEMM due to the high complexity of operand mapping. Unlike electronic crossbars, MZI meshes require unitary decompositions, most commonly SVD, to derive precise phase settings for each interferometric element. While acceptable for static weights, this process becomes a prohibitive runtime bottleneck when **W** changes every cycle. For example, computing the SVD and corresponding phase decomposition for a 12 × 12 matrix can take approximately 1.5 ms on a CPU, introducing system stalls that overwhelm the intrinsic speed of optical propagation. Moreover, to reduce footprint and insertion loss, these designs often rely on compact thermo-optic or non-volatile phase shifters, such as phase change materials. The programming latency of such devices, 10 ns to 10 μs, is orders of magnitude slower than the optical computation itself, rendering reconfiguration costs impossible to amortize under dynamic workloads.

Challenge 2: Sign Representation Overhead in Incoherent Architectures. Incoherent designs, such as MRR weight banks [73], face a fundamental limitation in representing full-range values. Because computation is performed through light intensity modulation, at least one operand must be non-negative. Supporting signed multiplication, therefore, requires decomposing operands into positive and negative components, for example, $X = X_+ - X_-$. A single multiplication $(X_+ - X_-)(W_+ - W_-)$ then expands into four sub operations, X_+W_+, X_+W_-, X_-W_+, and X_-W_-, which must be executed through time multiplexing or hardware duplication [99], [100]. This results in a > 2 to $4\times$ increase in hardware complexity and energy consumption. By significantly increasing modulation, DAC, and control overheads, this decomposition erodes the efficiency benefits that incoherent architectures typically derive from weight static dataflows.

Existing Solutions for Dynamic Workloads. With the emergence of Transformer-based foundation models, the dominant computational paradigm has shifted from static convolutions to dynamic attention mechanisms. In this regime, hardware must efficiently support interactions between runtime-generated operands, rather than between fixed weights and activations.

To address this challenge, one representative direction is to jointly optimize the photonic datapath together with workload-aware data movement and scheduling. Zhu and Gu [32] proposed *Lightening Transformer*, a high-speed optical accelerator that co-designs the photonic datapath with a specialized on-chip and off-chip tiling strategy. The system accelerates dynamic attention by leveraging coherent interference and WDM to achieve high spectral parallelism. To alleviate data movement bottlenecks, the architecture adopts a crossbar-style topology that maximizes intra-core operand reuse, employing dynamic orchestration of broadcast and tiling to amortize electro-optical conversion costs.

Related efforts have explored temporal orchestration as another route to supporting dynamic functionality. Kari et al. [101] proposed a time-multiplexed realization for executing dynamic dot products, sharing architectural principles with the optical tensor processor introduced by Yin et al. [57].

More broadly, dynamic reconfigurability can also be improved by shifting part of the programmability burden away from optical platform. Beyond attention accelerators, recent work from Lightmatter demonstrated a universal photonic processor capable of supporting a wide range of workloads [23], including natural language processing and deep reinforcement learning. To overcome the intrinsic rigidity of optical weight encoding, their design shifts weight programmability to the electrical domain using a differential photodetection unit coupled with a resistive, differential DAC. This approach enables full reconfigurability without incurring the latency penalties associated with tuning thermal or phase change optical elements.

More broadly, the bottleneck-driven analysis in this section reveals a recurring pattern: scalable photonic AI cannot be unlocked by improvements at a single abstraction level alone, because the dominant limitations are

inherently cross-layer. Gains achieved at one level are often capped, shifted, or even offset by constraints at another. For example, faster or denser photonic primitives at the device level do not automatically translate into higher system efficiency if circuit-level calibration, control, and interface overhead remain substantial. Similarly, architecture-level parallelism and data reuse can still be throttled by insufficient device bandwidth, optical loss, or limited programmability. Conversely, highly compact or largely passive architectures may improve efficiency, yet reduce flexibility and place a large burden on peripheral control and workload mapping. In this sense, device-level advances determine the footprint, bandwidth, loss, and programmability of photonic primitives, circuit-level techniques govern control overhead, signed arithmetic support, and interface efficiency, and architecture-level strategies decide how effectively conversion, communication, and memory costs can be amortized across workloads. Therefore, addressing the bottlenecks identified in Sections II and III increasingly requires systematic automation and cross-layer co-design.

IV. Enabling Photonic AI at Scale: Full-Lifecycle Electronic-Photonic Design Automation

Photonic artificial intelligence systems demand a level of design complexity and scale that can no longer be supported by conventional ad hoc, isolated, manual design flow. As photonics moves toward very-large photonic integration (VLPI) and heterogeneous EPICs, successful deployment increasingly depends on a full-lifecycle EPDA stack. We focus on how advances in co-simulation, inverse and automated design, and physical design automation are converging to enable photonic AI systems that are not only high-performing, but also manufacturable, robust, and deployable at scale.

A. EPDA: Device-/Circuit-Level Capabilities

In an EPDA stack, device- and circuit-level simulation is the *translation layer* between nanophotonic physics and system-level performance. For photonic machine intelligence, simulation is not merely verification, but also enables co-design across *devices, circuits, architectures, and learning algorithms.* Thus, "good simulation" must provide *scalability* to large design, *composability* into system, *parametric conditioning* over operating conditions, and *differentiability/uncertainty awareness* for inverse design and robustness, motivating AI-assisted alternatives.

1) AI-Assisted Photonic Device Simulation

AI-assisted simulation uses ML to approximate the Maxwell solution operator, addressing the oversimplification of compact models and the scalability limits of rigorous electro-magnetic (EM) solvers. These surrogates seek to preserve physical accuracy while delivering the speed, scalability, and differentiability needed for EPDA, especially for *system-aware* workflows such as parametric sweeps, composable circuit co-simulation, and differentiable inverse design. This EM surrogate maps a device description, typically discretized permittivity plus sources and boundary conditions, to EM responses. Table 1 summarizes representative approaches by domain (frequency versus time), physics priors, and scalability.

Prediction target: FoM versus EM field. *FoM surrogates* predict metrics (e.g., S-parameters, group index) without full-field reconstruction. They are fast but often *non-composable* and device-specific. *EM field surrogates* predict steady-state fields or time-domain evolution, enabling downstream multiple FoMs extraction and inverse design; their **composability, generalization,** and **differentiability** are essential to circuit- and system-level co-simulation.

Frequency-domain field surrogates: one-shot versus iterative. Most methods target steady-state *frequency-domain* solutions. *One-shot* models map devices/conditions to complex fields in one pass; *physics-driven/augmented* CNNs (e.g., MaxwellNet [102], WaveY-Net [103]) use residuals (e.g., $|Ae - b|$) to reduce artifacts while accelerating inference. A second line of works, *Operator learning* [108], approximates **parametric** Maxwell operators. NeurOLight [104] enables fast sweeps over wavelength, sources, and permittivity. PACE [105]) improves fidelity on challenging structures such as metalines and large interferometric devices. From an EPDA perspective, one-shot surrogates can reach $10^2 \sim 10^3\times$ speedups but often degrade on larger domains or complex scattering [102]; transfer learning [104] helps, yet robust generalization remains open.

To address domain scaling, *Iterative Maxwell neural solving* hybridizes learned local solves with classical loops (often via domain decomposition) and refines until criteria such as $|\mathbf{A}x-b| < \epsilon$, trading smaller speedups (often $\sim 10\times$) for robustness on large ($> 100\lambda$)domains.

Time-domain field surrogates. Time-domain surrogates learn FDTD-like spatiotemporal evolution [107], [109] for transient/broadband behavior, but must address *long-horizon* error accumulation. PIC^{2}O-Sim [107] uses causality-aware dynamic convolution aligned with Maxwell dynamics to achieve large speedups over FDTD (e.g., MEEP) with stable rollouts. **Physics priors and learning paradigms.** Across the aforementioned methods, the central question is *how physics is incorporated. Physics-driven* residual minimization reduces labels but can be optimization unstable. *Physics-augmented*

Table 1. Categorized AI-assisted photonic device full-field simulation approaches with key feature comparison. *N*-means normalized. MAE is the mean absolute error.

Domain	Approach	Model	Inputs | Condition	Physics Prior	Devices/ Scales	Speedup	Error	Comment
Frequency -domain	One-shot	MaxwellNet [102]	$\epsilon_r \mid \lambda, \Omega, J$, PML	Maxwell Loss	Lens/$< 10\lambda$	300-600× over COMSOL	~0.01 N-L2 Error	
		WaveY-Net [103] (U-Net)	$\epsilon_r \mid \lambda, \Omega, J$, PML	Maxwell Loss	Grating/$< 10\lambda$	700× over Direct Solver	3e-2 MAE	Fast Limited scalability
		NeurOLight [104] (FNO)	$\epsilon_r, \lambda, \Omega, J \mid$ PML	Wave Prior	Tunable/etched MMI/ ~ 10λ	>100× over Direct Solver	0.12 N-MAE	
		PACE [105] (FNO)	$\epsilon_r, \lambda, \Omega, J \mid$ PML	Wave Prior	Tunable/etched MMI, Metaline/ ~ 10λ	150-500× over Direct Solver	0.03-0.1 N-MAE	
	Iterative	FNO+F-GMRES [106]	$\epsilon_r, \lambda, \Omega, J \mid Ae - b \mid$, PML	Maxwell Residual	WDM, Coupler, Metalens/ ~ 100λ	10× over iterative GMRES	1e-3 L1 Error	Scalable Limited speedup
Time -domain	Autoregressive	PIC^{2}OSim [107] (CNN)	$\epsilon_r, \Omega, J \mid$ PML	Causality, Model	MRR, MMI, Metaline	300-600× over MEEP	3e-2 N-L2Norm	Broadband Error accumulates

training adds PDE/boundary and conservation/reciprocity constraints to improve physical validity. *Data-driven* operator learning (NeurOLight [104], PACE [105], PIC^{2}O-Sim [107]) embeds physics via **inputs, architectures** (local causality, global interference), and **training** (e.g., superposition augmentation [104]) for better generalization.

2) Photonic-Electronic Circuit-Level Co-Simulation

After device-level validation, circuit-level simulation captures component interactions for system modeling. A co-simulator must balance (i) enough fidelity to model nonidealities (loss/dispersion, reflections/feedback, modulation limits, noise, thermal drift) and (ii) enough speed for architecture exploration and learning-hardware co-design.

The prevailing workflow: extract-then-simulate. Today's dominant flow is *hierarchical:* devices are characterized by EM/measurement, reduced to compact models (e.g., frequency-dependent S-parameters), and composed in circuit simulators for sweeps and link budgets. Electronics (drivers, TIAs, control, DAC/ADC) are usually simulated separately in SPICE/Verilog(-A) with coarse interfaces. This ***split-flow*** **breaks down** for large photonic AI systems where mixed-domain interactions (loading, quantization/noise, feedback/calibration, thermal drift) jointly set performance, highlighting the need for unified, scalable photonic-electronic co-simulation [110].

Challenges in EPIC co-simulation. The difficulty is rooted in a *mismatch of native formalisms:* electronics solvers operate on voltages/currents in time-domain modified nodal analysis, while photonic circuits use complex waves and multiport scattering across wavelength/polarization. Bridging them needs stable, physically consistent interface models (modulators, detectors, impedance/parasitics) that translate electrical-optical variables [110], which is why many flows either couple domains manually or translate one into the other.

(1) Circuit-Level Unification via Behavioral Compact Models. Compact models make large-scale PIC simulation tractable and act as a *process development kit (PDK) contract,* typically via S-parameters for fast composition and sweeps. A common strategy implements photonics as behavioral models in electronics-native languages (Verilog-A [111], [112], SPICE [113]) so photonic blocks run inside mature electronic design automation (EDA) simulators, improving interface realism and mixed-signal verification [112]. This enables unified transient analysis and device-specific models for co-design with CMOS drivers/receivers, but introduces costs: model translation can be labor-intensive and

inconsistent with PDK models, and broadband response can require expensive sweeps (mitigated by chirp-based transients) [112]. Thus, compact-model unification is necessary but not sufficient for faithful mixed-domain validation at photonic-AI scale.

(2) Coupled-Domain Co-Simulation That Preserves Native Abstractions. An alternative couples domain-appropriate solvers through explicit interfaces instead of forcing a single representation, avoiding staged "handoff" co-simulation. This becomes crucial as systems exhibit *nonlinear, time-variant* electro-photonic interactions and feedback [111]. Recent work uses microwave-style *power waves* for bidirectional/reflection-aware modeling [114], and SPIPE couples a SPICE engine with an S-parameter photonic solver via physical modulator/photodetector interfaces [110], preserving transistor-level transients and scalable photonic S-matrix composition.

Perspectives and Future Directions. Looking forward, circuit-level EPIC simulation should evolve from merely "simulating a netlist" to *co-simulating and optimizing heterogeneous photonic-electronic systems under realistic conditions, which plays a key in EPDA flow to bridge physics-heavy photonic design with logic-heavy electronic design.* This requires simulation flows that not only resolve optical device and circuit behavior, but also incorporate RF interfaces, electronic control, and system-level optimization objectives. As photonic AI hardware scales, key directions include: **(1) Scalability:** co-simulation over thousands of devices with complex coupling (feedback, monitoring, reconfiguration); **(2) Radio frequency (RF)- and multi-wavelength awareness:** jointly modeling bandwidth/impedance/RF-optical effects with wavelength-dependent propagation and interference; **(3) AIassisted simulation acceleration:** using surrogate models or learned compact models to speed up computationally intensive photonic simulation and enable practical large-scale EPDA; **(3) Differentiability:** enabling end-to-end gradients for joint device/circuit optimization under AI-centric metrics; **(4) Layout awareness:** post-layout back-annotation via extracted parasitics and interconnect models for layout-dependent effects; and **(5) Measurement-in-the-loop:** measurement-informed digital twins that update compact models and uncertainty bounds to improve variation robustness.

B. EPDA: Architecture-Level Modeling

While device- and circuit-level simulation captures the physics of individual photonic components, it is insufficient for *system-valid* evaluation of photonic AI accelerators, where end-to-end efficiency emerges from the interaction among photonic compute units, electronic peripherals, conversion interfaces, memory hierarchy, interconnects, calibration/control, and workload mapping. Architecture-level EPDA provides the abstraction layer that connects component characteristics to system metrics under realistic workloads. Crucially, for photonic machine intelligence, architecture modeling must go beyond conventional "performance modeling" and explicitly represent optics-specific parallelism, analog error sources, and cross-domain overheads that can dominate at scale.

1) What Makes Photonic Architecture Modeling Different?

Architecture-level EPDA aims to model latency, throughput, energy, area, and accuracy trade-offs of heterogeneous EPIC systems without resolving electromagnetic fields. However, photonic AI systems violate several implicit assumptions that underpin many electronic-accelerator simulators:

- **Cross-Domain Overheads are First-Order.** Electrical-optical (E-O) interfaces (e.g., DAC/ADC, modulators, TIAs) can outweigh optical compute energy when scaled to high bandwidths or high precision.
- **Optics Introduces Non-Digital Error Behavior.** Phase noise, laser relative intensity noise, drift, interference, and analog accumulation yield *correlated* and often *data-dependent* errors, which cannot be captured by simple bit-flip or i.i.d. additive-noise models.
- **Parallelism is Multi-Dimensional.** Spatial replication, WDM, and broadcast/accumulate structures alter utilization, scheduling, and bottlenecks in ways that do not map cleanly to standard systolic or SRAM-centric models.

Therefore, a credible EPDA stack must (i) account for conversion and control costs, (ii) connect non-idealities to algorithm-level accuracy, and (iii) model workload-to-hardware mapping with photonics-aware primitives.

2) Adapting Electronic Architecture Simulators

Early efforts toward architecture-level photonic modeling leveraged structural similarities between photonic accelerators and analog compute-in-memory (CiM) systems. By adapting existing CiM architectural simulators, researchers demonstrated that photonic systems can be evaluated within a full-system context that accounts for DRAM access, on-chip buffering, and cross-domain data movement. This line of work provided an important system-level insight: even when optical-domain computation is highly efficient, data conversion and memory traffic can dominate total system energy, highlighting

the need for joint consideration of architecture and mapping strategies.

These CiM-inspired approaches primarily emphasize array-style accelerator organizations and dataflow-centric analysis, which are inherited from electronic architectures. While effective for capturing full-system behavior and enabling rapid design-space exploration, they are generally tailored to regular compute structures and abstract photonic hardware at a coarse architectural level. As a result, they are particularly well-suited for system-level comparisons and workload-driven analysis, rather than detailed exploration of photonic-specific architectural diversity.

3) Architecture-Specific Photonic Simulators

Beyond generic CiM-based modeling, several photonic accelerator efforts have developed custom architecture-level simulators tailored to specific optical computing paradigms. A representative example is the coherent photonic crossbar accelerator based on PCM, which employs a modified SCALE-Sim–based framework to model compute cycles, weight programming overhead, memory accesses, and peripheral electronics at the system level.

In this approach, cycle-accurate architectural modeling is combined with simulated/measured device characteristics, including losses, laser efficiency, ADC/DAC power, and SRAM/DRAM access energy, to evaluate end-to-end metrics such as throughput, energy efficiency, and chip area for large convolutional neural network workloads.

By explicitly modeling programming latency, batch size, and memory hierarchy effects, this class of simulators demonstrates how system-level constraints significantly influence the scalability of photonic accelerators beyond small arrays.

At the same time, these architecture-specific simulators are intentionally optimized around a given photonic design and operating regime. This specialization enables detailed and realistic evaluation of targeted architectures, while naturally limiting direct reuse for exploring a wide range of PTC topologies or performing broad cross-architecture comparisons.

4) Native Photonic Architecture Modeling Frameworks

More recently, photonic architecture modeling frameworks have emerged that aim to unify device behavior, circuit organization, and architectural execution within a unified flow (e.g., cross-layer approaches such as SimPhony [36]). Instead of assuming a fixed array abstraction like in electronic accelerators, these frameworks enable *parametric* construction of heterogeneous PTCs, encompassing mesh-, array-, and broadcast-style structures under a unified representation. At the architectural level, they explicitly model *photonic dataflow,* including multi-dimensional parallelism along spatial, spectral, and temporal dimensions, and hierarchical mixed-signal accumulation patterns. In addition, energy and performance estimation is often tied to configuration states, workload characteristics, and hardware budgets, allowing loss, laser power, and footprint to be reflected directly in architectural evaluation. By making such assumptions explicit, photonics-native frameworks facilitate more systematic and transparent architectural exploration across a wider design space.

Nevertheless, like architecture-specific simulators, these frameworks inevitably rely on behavioral abstractions and modeling assumptions whose validity depends on how non-idealities, calibration procedures, and control overheads are represented. Effects such as thermal crosstalk, fabrication variation, wavelength drift, and coherence-related accuracy degradation are still typically modeled only approximately. More accurate simulation will therefore require tighter integration with *device-level models, layout-induced parasitics, and measurement/test data.* For example, fabrication-induced errors should be captured through calibrated device models and process-aware extraction; placement- and interconnect-induced thermal crosstalk through parasitic extraction; and routing-induced path-length and phase mismatch through post-layout back-annotation into the simulator. This points to a *closed-loop EPDA flow* connecting design, layout, extraction, measurement, and simulation through standardized data-exchange interfaces and compact-model representations. Such a flow would provide fast feedback from implementation and test to system- and architecture-level simulation, improving model fidelity while preserving design-space exploration efficiency.

5) Open Problem and Future Direction

Architecture-level EPDA for photonic AI is still at an early stage: most existing studies provide valuable proof-of-concept system modeling and design space exploration [115], yet the field lacks a *full-lifecycle* methodology that connects workload intent to implementable heterogeneous EPIC systems with predictable performance and closed-loop co-design. Looking forward, the key opportunity is to elevate architecture-level EPDA from "estimating throughput on ideal blocks" to *application-to-hardware co-design and compile-time system synthesis,* grounded in physically realistic constraints and validated through cross-layer closure.

(1) Application-Architecture Co-Design. Future photonic accelerators will be judged by end-to-end workload

performance. This requires EPDA frameworks that understand the structure of modern applications, AI, scientific computing, and beyond, and expose architectural knobs that matter in practice: tensor operator support, precision formats, activation/dataflow movement, and control/calibration scheduling. A central research direction is to build *workload-aware photonic architectures,* where the choice of photonic compute primitive is guided by application structure.

(2) Workload-to-System Compiling and Mapping for Heterogeneous EPIC. A missing layer in many photonic architecture studies is a compiler-grade mapping stack that translates models into executable schedules under heterogeneous constraints: photonic resource allocation, memory hierarchy and data movement, conversion and control overheads, and timing constraints for reconfiguration. Prior work H^{3}PIMap leverages SimPhony as a performance evaluator to optimize workload mapping for 3-D hybrid photonic in-memory computing systems [116]. In the future, EPDA needs *workload-to-system compilation* that co-optimizes dataflow, tiling, placement of compute/memory, and reconfiguration policies.

(3) Rigorous System Performance Evaluation: From Ideal Operations to Signal Integrity and Robustness. Architecture-level evaluation must move beyond idealized TOPS/W estimates and simplistic independent noise injection. Photonic systems are constrained by signal integrity (SNR, effective number of bits (ENOB), dynamic range, bandwidth), polarization/wavelength/thermal management. A major open problem is to develop *architecture-appropriate abstractions* that capture these effects with high fidelity and efficiency and to support uncertainty-aware evaluation.

In summary, the next phase of architecture-level EPDA will be defined by *compiler-like workload mapping, physically grounded system evaluation,* and *closed-loop cross-layer co-design* that links architectural intent to implementable and robust heterogeneous photonic-electronic systems.

C. EPDA: Component-Level Inverse Design

1) Limitations of Manually Designed Devices

Conventional photonic device design largely follows a *forward-design* workflow: starting from canonical topologies (e.g., couplers, rings) and tuning a small set of geometric parameters via simulation sweeps. While effective for standard building blocks, it becomes a bottleneck when photonic AI systems require *compact footprints, complex functionality,* and *multi-metric optimization* (loss, bandwidth, extinction). Manual tuning typically searches only a *low-dimensional subspace of a chosen topology,* limiting discovery of non-intuitive structures and often forcing larger area or degraded performance under tight constraints. It is also *expert-dependent,* relying on significant intuition and trial-and-error that hinders accessibility and scalability.

These limitations diverge from the trend in electronics, where automation has turned design into a computation-driven workflow that scales with complexity [140], [141]. Photonics has not yet fully leveraged modern compute and AI/EDA advances, motivating *automated device synthesis* as an optimization problem that expands the search space and directly targets circuit- and system-level objectives.

2) Introduction to Inverse Design of Photonic Devices

To overcome the limits of manual trial-and-error, **inverse design** starts from a *target specification* [e.g., spectrum or figure of merit (FoM)] and automatically synthesizes device geometry within a design region (Fig. 6). A typical pipeline chooses design variables $\boldsymbol{\theta} \in \mathbb{R}^N$ and a parameterization mapping $\boldsymbol{\theta}$ to layout (pixels, splines, etc.), evaluates the FoM via a Maxwell solver or surrogate, and iteratively updates $\boldsymbol{\theta}$ using adjoint gradients or gradient-free search until convergence. By exploring high-dimensional, non-intuitive spaces, inverse design can produce compact structures that are hard to obtain by hand, often achieving similar functionality with a much smaller footprint (Fig. 7), enabling dense photonic integration.

Challenges in photonic inverse design. Although photonic inverse design has demonstrated strong potential, key challenges remain (Fig. 8): **(1) Manufacturability/yield:** irregular or pixelated layouts can violate foundry rules and increase process sensitivity (e.g., etch bias, line edge roughness, linewidth variation), thereby degrading yield; **(2) Simulation cost:** repeated high-fidelity solves across wavelengths/polarizations/corners are computationally expensive; **(3) Non-uniqueness/non-convexity:** many-to-one mappings, where distinct geometries can produce similar responses, and local minima make outcomes initialization-dependent; and **(4) Multi-objective trade-offs:** improving one metric can hurt other metrics (e.g., bandwidth, loss, or crosstalk), requiring principled objective balancing, constraint- and application-aware optimization.

Categorization of photonic inverse design methods. To address the challenges, inverse-design methods broadly fall into **(i) optimization-driven** and **(ii) AI-assisted** families [142], [143], which are often combined to balance exploration and efficiency.

Optimization-driven approaches typically include **heuristic/evolutionary** and **gradient-based** methods. Heuristics offer **global exploration** by maintaining a population of candidates (GA [124], [128], [144], PSO [117], [118], [119], [127], [145], DBS [129], [146]), well

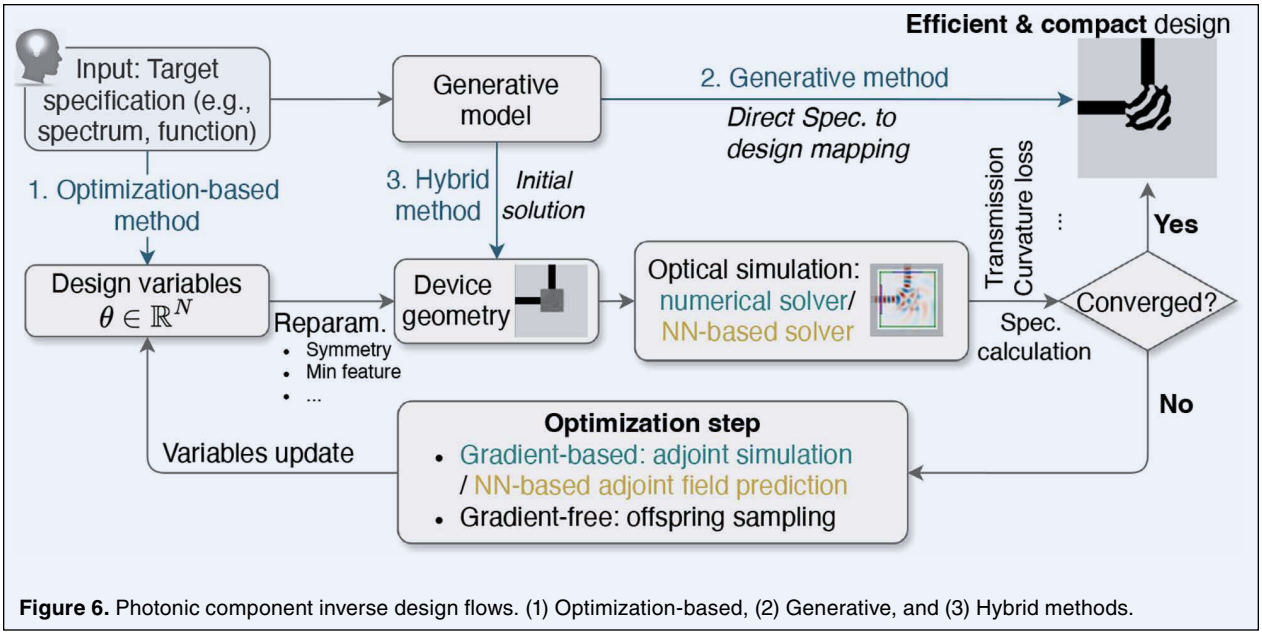

Figure 6. Photonic component inverse design flows. (1) Optimization-based, (2) Generative, and (3) Hybrid methods.

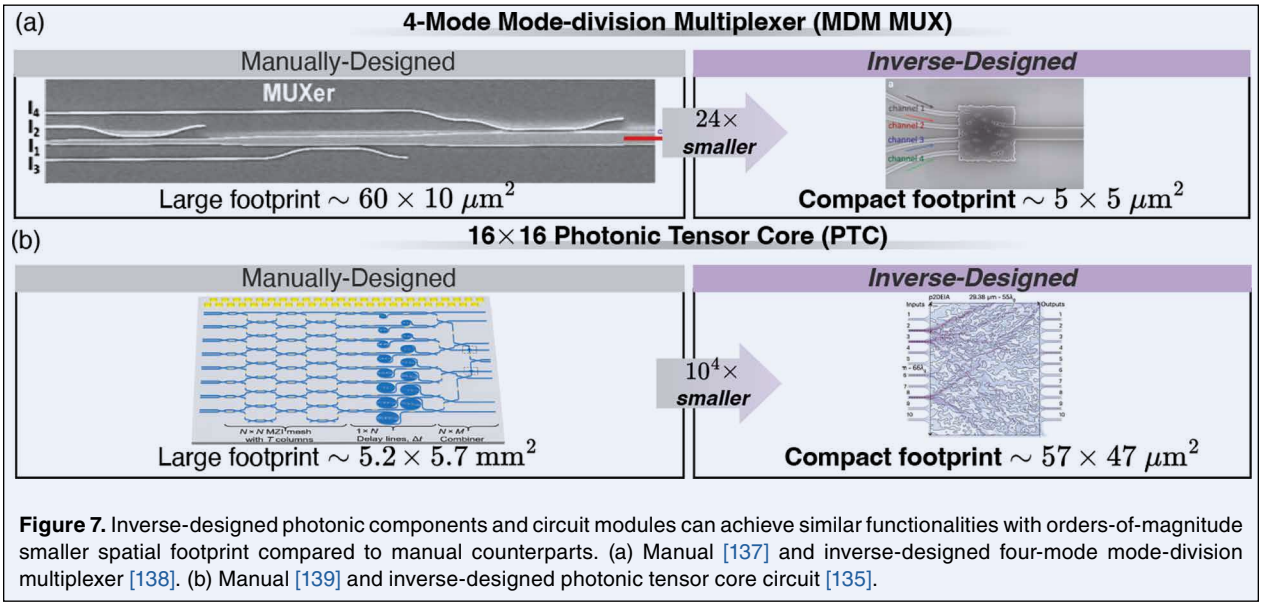

Figure 7. Inverse-designed photonic components and circuit modules can achieve similar functionalities with orders-of-magnitude smaller spatial footprint compared to manual counterparts. (a) Manual [137] and inverse-designed four-mode mode-division multiplexer [138]. (b) Manual [139] and inverse-designed photonic tensor core circuit [135].

suited to low-dimensional or discrete designs, but are often simulation-hungry (e.g., ~100 hours for GA in some cases [144]); Bayesian Optimization (BO) [121] improves sample efficiency but struggles as dimension grows.

Gradient-based inverse design typically uses adjoint method [147] to obtain $\partial \text{FoM}/\partial \boldsymbol{\theta}$ for thousands of parameters with one extra simulation, enabling large-scale topology optimization [130], [131], [133], [134], [135], [136], [148]; however, it is **local and initialization-sensitive,** and can yield non-manufacturable or variation-fragile patterns without constraints/ regularization [149].

AI-based methods typically include **predictive** and **generative** approaches, mainly for **cheaper evaluation** and **better proposal quality.** Predictive surrogates [117], [118], [133] approximate forward solvers for near-instant FoM evaluation inside search loops or for warm-starting in adjoint refinement, but can fail under distribution shift or sparse coverage. *Generative methods* [132], [150], [151] directly propose geometries conditioned on targets, reducing initialization sensitivity and exploring multi-modal solutions. This is especially valuable for ill-posed, many-to-one inverse problems; Candidates are then followed by physics verification

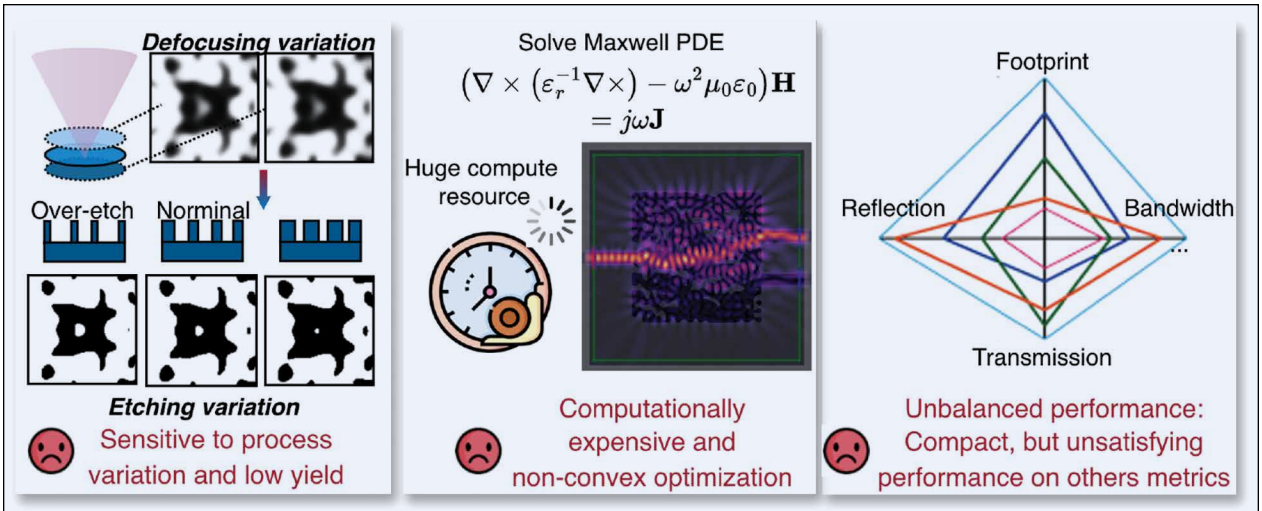

Figure 8. Challenges of photonic inverse design.

and local, constraint-aware refinement to ensure correctness and manufacturability.

3) Applications of Photonic Inverse-Designed Devices

Table 2 surveys **representative inverse-designed devices and their settings** [parameterization, optimizer, design of freedom (DoF), fab-awareness, and the approximate number of electromagnetic simulations (# Sims)]. Inverse design has been applied broadly to *passive* PIC building blocks such as wavelength routers [118] and filters [133], phase shifters [119], microring resonators [121], mode [122] and power splitters [132], [123], grating couplers [126], polarization rotators [128], multimode crossings [129], WDM demultiplexers [134], [148], and nonlinear optical switches [136].

Notably, *matrix-vector multiplication (MVM) units* [130], [135] for ONN and photonic accelerators have emerged as prominent application targets, because compact footprint and engineered spectral responses directly impact scalable, energy-efficient AI hardware. The paradigm also extends beyond on-chip PICs to fiber, lasers, and free-space optics [120], [124], [125]. In addition to passive components, inverse design has been demonstrated for *active or tunable* devices, including optical amplifiers [117], silicon modulators [127], and PCM-based MMIs [131], highlighting its relevance to both communication and computing-oriented photonic systems.

Across applications, DoF ranges from a few structural parameters to 10^3–10^4 in pixel/free-form geometries, which largely determines the optimizer choice. Heuristic methods (PSO/GA/DRL/DBS) are common for low-moderate DoF due to better global exploration but higher simulation cost, whereas adjoint-based gradient optimization dominates at high DoF by enabling thousands of variable updates.

Despite these successes, Table 2 also highlights a persistent gap between numerical optimality and manufacturable performance. Many pixel/free-form solutions introduce sub-resolution features and sharp geometries that are process-sensitive, motivating *fab-aware* inverse design (FAID) [133], [149], [152],[153],[154],[155],[156],[157],[158],[159]. Existing strategies range from *post-hoc* filtering/regularization, explicit min-feature and deformation constraints, and *in-loop* enforcement (e.g., differentiable lithography or DRC-aware optimization) to restrict search to manufacturable subspaces and improve robustness.

4) Prospective and Open Challenges

Looking forward, fabrication-aware inverse design must move beyond "idealized 2D optimization" toward *system-ready devices under realistic 3D process variation and coupled multiphysics.* Key directions include: **(1) Closing the simulation-fabrication gap with realistic variability models:** capturing 3D effects (sidewalls, roughness, etch/linewidth nonuniformity, index fluctuation) across *local* and *global* scales to translate uncertainty into robust margins. **(2) Scalable high-fidelity EM/multiphysics simulation:** 3D EM plus thermal/electrical/mechanical coupling is far costlier than 2D, limiting device size, spectral breadth, and free-form DoF. **(3) Fast yet accurate 3D-level AI solvers:** learned surrogates should deliver near-3D fidelity with reliable gradients and calibrated uncertainty for yield-aware optimization; large language model (LLM)-guided orchestration

TABLE 2. Representative inverse-designed photonic devices, comparing geometry parameterization schemes, optimization methods, degrees of freedom (DoF), design-region sizes, fabrication-awareness strategies, and the approximate number of simulations required (# Sims). Devices marked with † are active components. *Abbreviations:* PSO = particle swarm optimization; BO = Bayesian optimization; GA = genetic algorithm; DRL = deep reinforcement learning; DBS = direct-binary search; AGO = adjoint gradient optimization.

Device	Geometry	InvDes method	DoF	Design region	Fab-aware	# Sims
Optical amplifier† [117]	Structural	PSO + NN	7	N/A	×	~10,000
Wavelength router [118]	Structural	PSO + NN	5	N/A	Post-hoc	~1,000
Phase shifter [119]	Structural	PSO	3	Lenght ~ 3 μm	×	~1,000
Few-mode fiber [120]	Structural	Inversed NN	5	N/A	Post-hoc	~10,000
Microring resonator [121]	Boundary	BO	~10	Radius ~ 3 μm	Post-hoc	~100
Mode splitter [122]	Boundary	AGO	~200	14 × 2.5 μm^2	Post-hoc	~400
Power splitter [123]	Boundary	AGO	~200	6 × 2.7 μm^2	Δy bound	~200
Integrated lens [124]	Element-array	GA	~100	~ 5 × 10 μm^2	×	~1,000
Nanobeam laser [125]	Element-array	DRL	~200	20 × 0.7 μm^2	×	~1,000
Grating coupler [126]	Element-array	AGO	~200	12 × 0.5 μm^2	Feature-size	~600
Silicon modulator† [127]	Pixel-based	PSO	~30	N/A	×	~1,500
Polarization rotator [128]	Pixel-based	GA	~280	Length ~ 3 μm	×	~10,000
Four-mode crossing [129]	Pixel-based	DBS	~2,000	~ 12 × 12 μm^2	Hole diameter	~1,000
MVM unit [130]	Pixel-based	AGO	~1,000	4.8 × 2.88 μm^2	Pattern averaging	~400
PCM MMI† [131]	Pixel-based	AGO	~1,000	~ 40 × 8.5 μm^2	×	~500
Power splitter [132]	Pixel-based	Generative NN	~400	2.25 × 2.25 μm^2	×	~10,000
Wavelength filters [133]	Pixel-based	AGO + NN	~500	4 × 2 μm^2	Feature-size	~20,000
WDM demultiplexer [134]	Free-form	AGO	~10,000	2.8 × 2.8 μm^2	Multi- λ broadband	~400
MVM unit [135]	Free-form	AGO	~10,000	~ 11 × 10.3 μm^2	Low-index contrast	~500
Nonlinear optical switch [136]	Free-form	AGO	~20,000	~ 5 × 5 μm^2	Feature-size	~2,000
WDM demultiplexer [133]	Free-form	AGO	~10,000	6.4 × 6.4 μm^2	In-loop DRC	~1,000

may improve usability [160]. **(4) Multiphysics-in-the-loop actuation/control:** modeling practical tuning/modulation (heaters) with RC limits, loss, thermal crosstalk, and power to enable robust reconfigurable PICs. **(5)**

Device-circuit co-design and layout-aware constraints: To make inverse-designed components deployable at the circuit level, optimization must incorporate circuit context and layout constraints, such as routing parasitics, coupling dispersion, thermal proximity, and packaging stress. **(6) From device inverse design to circuit-level topology inverse design:** A natural next step is to extend automated design beyond device synthesis toward *circuit and module optimization,* where device arrangement, interconnection, and parameterization are included in the search space. Compared to device-level inverse design, circuit-level synthesis introduces a combinatorial discrete search space with complicated constraints, making it substantially more challenging. Recent work has begun to explore this direction via differentiable and multi-objective optimization to automatically search Pareto-optimal photonic tensor core topologies that improve expressivity, area/energy efficiency, and robustness [161], [162]. While still nascent relative to device inverse design, these results suggest a path toward "design compilers" for programmable photonic fabrics.

D. EPDA: Circuit/Chip-Level Layout Automation

After obtaining the PIC netlist and component layouts, designers often use schematic-driven layout (SDL) [163]: manually place components and connect them with waveguides and wires. As PICs scale to AI systems with many components, this manual loop becomes a bottleneck, labor-intensive, error-prone, and hard to iterate since small schematic changes trigger re-placement/re-routing and repeated design-rule checks (DRC). This motivates PIC layout automation for faster iteration, scalability, and improved quality.

1) Challenges of PIC Layout Automation

(1) Layout Sensitivity and Physics-Aware Constraints. PIC layout is *inherently* performance-driven: the geometry directly dictates system behavior. A simple waveguide can be a functional element whose length sets phase and whose curvature/crossings set loss and crosstalk. And placement/routing must enforce constraints such as thermal crosstalk mitigation during heater placement, path-length matching for MZMs, and cross-domain exclusions (e.g., avoiding long metal overlap that increases optical loss). **(2) Resource-Limited Physical Layout.** Large-scale EPICs are constrained by area and limited routing layers (often a single optical layer and a few metal layers), so routability depends on early corridor planning and minimizing crossings/detours, including minimizing waveguide crossings, reducing via usage and detours in metal routing. And chip packaging further pins optical I/O and pads to fixed locations, creating rigid boundary conditions that concentrate congestion and make routing topologically complex. **(3) High-Speed Circuit Layout Challenges.** High-speed links add transmission-line constraints (impedance/group-index matching), requiring area-hungry geometries (e.g., coplanar waveguides), careful handling of bends and metal fill, and shielding for differential pairs. **(4) Scalability and Fabrication Metrics.** At the thousand-component scale, automation is essential, but must account for yield (process variation, lithography impacts on n_{eff}). The goal is to balance area, insertion loss, and electrical bandwidth under manufacturing and physical constraints.

2) Tools for PIC Layout Automation

As summarized in Fig. 9, we categorize prior research on PIC physical design automation into three stages.

(i) Early-Stage, Small-Scale ONoC-Driven Tools (2007–2012). Motivated by *optical networks-on-chip* (ONoCs), early works emphasized *waveguide routing* and lightweight automation, including timing/congestion-aware optical routing for 3D system-on-packag [164], reusable parameterized libraries (OIL) [165], power-aware routing optimization (O-Router) [166], and scriptable layout generation (VANDAL) [167]. In parallel, channel-level Manhattan/non-Manhattan detailed routing was also explored to better capture geometry and crossings in constrained regions [168], [169], [170].

(ii) WRONoC Placement-and-Routing With Topology Awareness (2013–2021). Research expanded toward automated placement and routing (P&R) for wavelength-routed ONoCs (WRONoCs), with increasing *co-optimization* between logical topology and physical layout. PROTON [171] pioneered end-to-end photonic P&R by combining nonlinear placement with Lee-style routing, while PLATON [172] introduced scalable force-directed placement. To reduce crossings and improve routability, PlanarONoC used planar/graph-based reasoning [173], and later works pursued topology–layout co-optimization [174]. Complementary studies also addressed key layout tasks such as path-length matching [175], and structure-aware routing to reduce detours/crossings [176]. In parallel, the community also advanced the *schematic-driven* methodology [177], [178]: the flow starts from a circuit schematic, uses a PDK-backed component library for immediate simulation, and then generates layout based on the schematic, followed by DRC verification and post-layout parameter extraction to update the schematic. This paradigm is also widely adopted in industry; commercial toolkits (e.g., Synopsys OptoCompiler and Cadence Virtuoso) and open-source frameworks (e.g., GDSFactory) support GUI-, API-driven layout

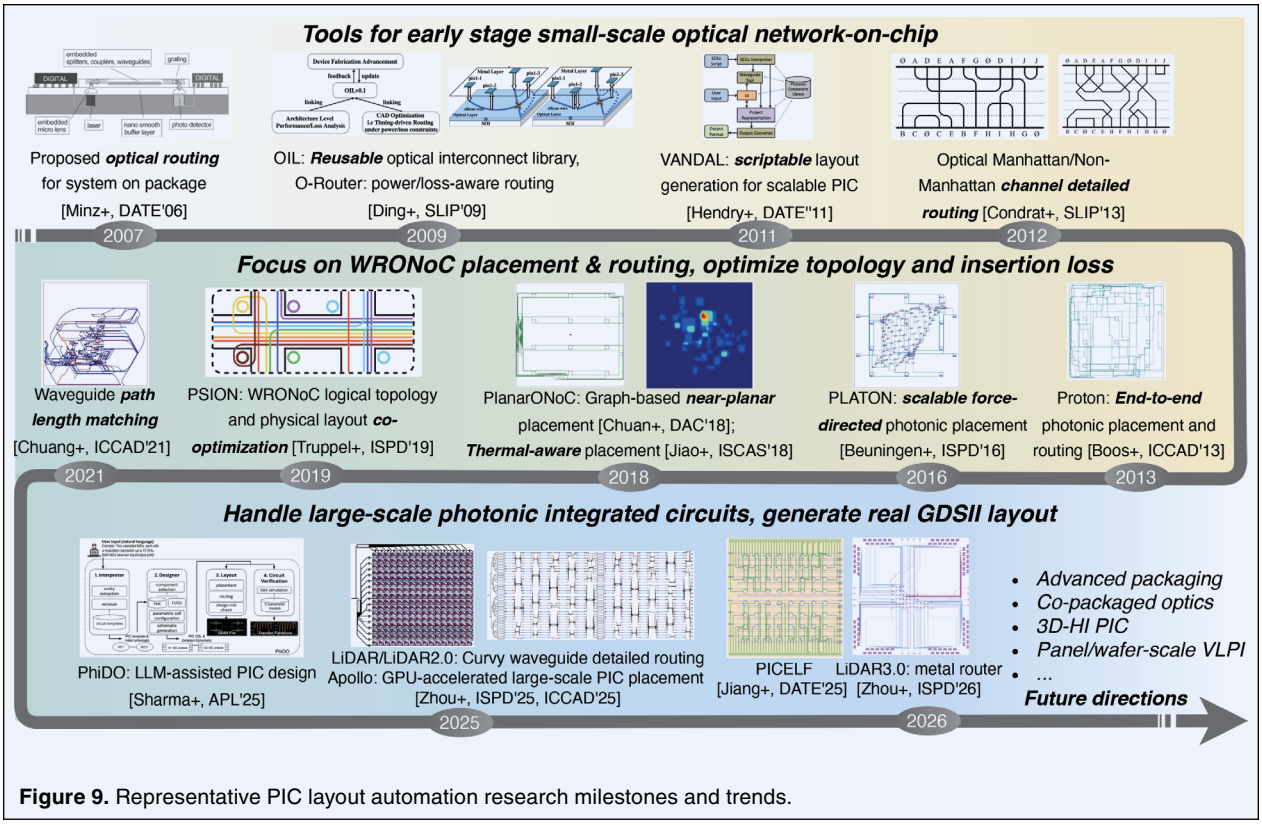

Figure 9. Representative PIC layout automation research milestones and trends.

generation, guided routing, and exporting netlists for simulation/verification [179].

(iii) Tackling Emerging Large-Scale PIC Design Automation (2022–present). With photonic computing pushing denser circuits, focus shifts to *circuit-scale* automation that outputs manufacturable GDSII. LiDAR/LiDAR2.0 [180], [181] advances detailed routing via dynamic crossing insertion and curvilinear routing, producing near-DRV-free final layout on WRONoC and photonic-computing designs. In parallel, the work [182] proposes an optical routing flow targeting phase/delay matching, using diffusion-based length matching with spiral detours. To further reduce loss under richer process options, the work [183] explicitly models hybrid waveguides and transitions, and optimizes insertion loss while enforcing matching constraints. As electrical nets scale, metal routing becomes a bottleneck [184]; to address photonics-specific keep-outs and spacing, recent work proposes congestion/DRC-aware global electrical planning with waveguide-aware assignment and guidance-driven detailed routing [185]. In addition, PICELF [184] targets the electronic routing by assigning the electrical pin via nonlinear binary programming and performing a fast two-stage router to produce DRC-clean metal layouts. Besides routing, Apollo [186] introduces GPU-accelerated, routing-informed placement with bending-aware objectives and explicit congestion/crossing modeling, achieving a 94.79% routing success rate on large-scale photonic-computing benchmarks. Beyond classical algorithmic EDA, emerging *agentic* workflows explore natural-language-to-GDSII automation. The PhIDO multi-agent framework [187] demonstrates an end-to-end pipeline that translates natural-language PIC requests into structurally valid layouts.

3) Prospective and Open Challenges

Looking forward, EPIC layout automation will be increasingly shaped by the emerging paradigm of **very-large photonic integration (VLPI),** where hundreds to thousands of photonic devices are co-integrated with electronics to form heterogeneous, programmable systems. In this context, physical synthesis should evolve into a unified framework that coordinates photonic devices, electronic circuits, interconnects, and package-level resources under a common methodology. In particular, EPDA must support **heterogeneous EPIC layout design** under advanced packaging and system-level constraints. A central direction is **photonic-electronic co-placement and co-routing,** where photonic components, electronic blocks, and their optical/electrical interconnects are optimized jointly rather than separately. This is particularly important in 2.5D/3D integration,

where PIC and EIC dies are connected through interposers, micro-bumps, and TSVs, and where bump assignment, inter-die routing congestion, signal integrity, fiber/laser coupling, thermal management, and power delivery must be considered together. To address these challenges, EPDA layout tools must enable joint optimization across PIC and EIC dies, interposer, and package hierarchies in a unified design space.

Second, **3D PIC and multi-layer photonics** present major opportunities and challenges. Moving beyond single-layer silicon waveguides to multi-layer and truly 3D interconnects can increase integration density, but requires new placement/routing abstractions (e.g., layer assignment, 3D waveguide vias), 3D-aware design rules, and cross-layer crosstalk and thermal constraints beyond current 2D workflows. Third, **robustness and first-pass manufacturability** will increasingly depend on **yield- and variability-aware optimization.** Incorporating process variations (linewidth/etch bias, thickness, overlay) and variability models directly into placement, routing, and device selection can reduce late-stage iterations and improve first-pass success. In this context, **machine-learning-guided heuristics** can accelerate exploration and provide better initial solutions (e.g., routability-aware placement seeds, rapid loss/crosstalk estimators), while still requiring physics-based refinement. Finally, scalable VLPI-based AI system deployment demands **verification and signoff beyond geometric DRC:** beyond geometric rule checking, future EPDA flows must support photonic layout-versus-schematic (LVS), proximity- and layout-dependent effect checking (e.g., waveguide coupling/crosstalk, crossing/bend penalties), and post-layout back-annotation for system-level co-simulation across optical, electrical, and thermal domains. Together, these EPDA capabilities are essential to translate future VLPI designs from layout to first-pass hardware with predictable system behavior, enabling photonic systems to scale with the rigor of electronic ICs.

V. Conclusion and Outlook

As machine intelligence becomes a pervasive infrastructure layer, compute demand is outpacing the energy and bandwidth gains of post-Moore silicon. This review has argued that photonic machine intelligence is entering a new phase, shifting from demonstrating physical feasibility to establishing system-level scalability and reproducible advantage. Realizing the potential of photonic computing requires a fundamental change in design philosophy. We conclude that the future of photonics lies not only in device foundation optimization, but also in a holistic system co-design where optical physics, electronic interfaces, and AI workloads are jointly optimized to create new system degrees of freedom. A primary insight of our analysis is that hardware must co-evolve with the rapid advancements in AI algorithms. To move beyond narrow, static inference, photonic AI systems must prioritize workload-flexible programmability and consistent computing fidelity, sustaining versatility and accuracy for dynamic, evolving workloads. This motivates a closed feedback loop between software and hardware, where device, circuit, system, algorithm, and physical implementation are designed together.

Sustaining this trajectory will require design workflows that scale with complexity. We identify EPDA as the pivotal enabler and new research focus for the community. The field must transition to a full-lifecycle design ecosystem that integrates (i) AI-assisted simulation and inverse design for compact, manufacturable components, (ii) rigorous cross-layer modeling for fair benchmarking and bottleneck attribution, and (iii) automated, verification-aware physical design for large-scale heterogeneous integration.

Outlook and Final Remark. Key directions for the field include: **(1) Standardized benchmarking and system evaluation**. Progress will increasingly hinge on community benchmarks that are physically rigorous (grounded in realistic device/interface models), system-comprehensive (including conversion, control, memory, interconnect), and workload-aware (exploring accuracy-efficiency tradeoffs on diverse applications and mapping strategies), enabling fair comparisons. **(2) Cross-layer co-design with versatility and robustness as key focus.** To remain relevant amid rapid algorithmic change, photonic AI must move from narrow, fixed-function demonstrations toward workload-flexible platforms, where robustness is treated as a first-class design constraint and addressed end-to-end, from device physics and mixed-signal interfaces to system control and algorithm-aware calibration/adaptation that sustains consistent accuracy over time. **(3) Open, reusable, full-stack EPDA infrastructure.** A decisive accelerant will be a full-stack EPDA toolflow, spanning simulation, inverse design, system modeling, and automated physical layout, that expedites design cycles, improves resilience and yield, and unlocks new design degrees of freedom by leveraging modern AI-driven methodologies and high-performance computing, ultimately turning lab-scale prototypes into reproducible ecosystems.

Acknowledgment

This work was supported in part by the Air Force Office of Scientific Research Multidisciplinary University Research Initiative under Grant FA9550-17-1-0071, in part

by the Air Force Office of Scientific Research (Photonics for AI and AI for Photonics, Texas Center for Optical Computing and Interconnects, and Equipment Donations from Nvidia) under Grant FA9550-23-1-0452. Hanqing Zhu, Shupeng Ning, Hongjian Zhou, and Ziang Yin contributed equally to this work.

Hanqing Zhu (Member, IEEE) received the B.E. degree in microelectronic science and engineering from Shanghai Jiao Tong University, Shanghai, China, in 2020, and the Ph.D. degree from the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX, USA, in 2026. He is currently a Member of Technical Staff at xAI. His current research interests include efficient and scalable AI algorithms and systems. He is excited about redesigning AI algorithms to be both theoretically grounded and practically efficient. His recent research emphasis is on large foundation models (efficient pre-training, test-time scaling, and reasoning). He has received the ML and Systems Rising Star in 2025, the MLSys Outstanding Paper Award (Honorable Mention) in 2025, the CVPR AI4CC Workshop Best Paper Award in 2025, and Texas ECE Achievement Award. Additionally, he won the Robert S. Hilbert Memorial Optical Design Competition in 2022 and secured first place in the IEEE/ACM MLCAD FPGA Macro-Placement Contest in 2023.

Shupeng Ning received the B.S. degree in microelectronics science and engineering from Nankai University, Tianjin, China, in 2020, and the Ph.D. degree in electrical and computer engineering from The University of Texas at Austin, Austin, TX, USA, in 2026. He is currently a Post-Doctoral Scholar Research Associate at California Institute of Technology, Pasadena, CA, USA. His research interests include photonic integrated circuits for high-performance computing and sensing applications. He received the Ben Streetman Prize in 2025 and won the Robert S. Hilbert Memorial Optical Design Competition in 2024.

Hongjian Zhou received the B.E. degree in electronic information engineering from Central South University, Changsha, China, in 2021, and the M.S. degree in electronic science and technology from ShanghaiTech University, Shanghai, China, in 2024. He is currently pursuing the Ph.D. degree with the School of Electrical, Computer and Energy Engineering, Arizona State University, AZ, USA, under the supervision of Prof. Jiaqi Gu. His interests include physical design, electronic-photonic design automation, and machine learning for EDA. He has received the Best Paper Award Nomination at ASP-DAC 2024 and ISPD 2026.

Ziang Yin received the B.S. degree in electrical and computer engineering and the M.S. degree in electrical and computer engineering from the University of Washington, Seattle, WA, USA, in 2022 and 2023, respectively. He is currently pursuing the Ph.D. degree with the School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, AZ, USA, where he works as a Graduate Research Assistant at Prof. Jiaqi Gu's Scope X Laboratory. He has authored or co-authored more than 20 peer-reviewed journal and conference papers in venues, including *Science Robotics*, *Journal of Applied Physics*, ACM/IEEE DAC, and ICCAD. His research interests span efficient machine learning and intelligent computing, including electronic-photonic accelerator architecture, cross-layer circuit-architecture-algorithm co-design, and AI/CAD techniques for optical and quantum systems. He is a Student Member of IEEE-HKN.

Ray T. Chen (Fellow, IEEE) received the B.S. degree in physics from National Tsing Hua University, Hsinchu, Taiwan, in 1980, and the M.S. degree in physics and the Ph.D. degree in electrical engineering from the University of California in 1983 and 1988, respectively. He is the Keys and Joan Curry/Cullen Trust Endowed Chair with the University of Texas at Austin (UT Austin), Austin, TX, USA. He is the Director of the Nanophotonics and Optical Interconnects Research Laboratory, Microelectronics Research Center. He is also the Director of the AFOSR MURI-Center for Silicon Nanomembrane involving faculty from Stanford, UIUC, Rutgers, and UT Austin. In 1992, he joined UT Austin to start the Optical Interconnect Research Program. From 1988 to 1992, he was a Research Scientist, the Manager, and the Director of the Department of Electro-Optic Engineering, Physical Optics Corporation, Torrance, CA, USA. From 2000 to 2001, he was the CTO, the Founder, and the Chairman of the Board of Radiant Research Inc., where he raised 18 million dollars A-Round funding to commercialize polymer-based photonic devices involving more than 20 patents, which were acquired by Finisar in 2002, a publicly traded company in Silicon Valley (NASDAQ: FNSR). He also serves as the Founder and the Chairman of the Board of Omega Optics Inc. since its initiation in 2001. Omega Optics has received over five million dollars in research funding. He

has supervised 41 post-doctoral researchers and graduated 62 Ph.D. students from his research group at UT Austin. Many of them are currently professors at the major research universities in the world. His research work has been awarded over 145 research grants and contracts from such sponsors as the Army, Navy, Air Force, DARPA, MDA, NSA, NSF, DOE, EPA, NIST, NIH, NASA, the State of Texas, and private industry. The research topics are focused on four main subjects: 1) nano-photonic passive and active devices for bio- and EM-wave sensing and interconnect applications; 2) thin film guided-wave optical interconnection and packaging for 2-D and 3-D laser beam routing and steering; 3) true time delay (TTD) wideband phased array antenna (PAA); and 4) 3-D printed micro-electronics and photonics. Experiences garnered through these programs are pivotal elements for his research and further commercialization. His group at UT Austin has reported its research findings in more than 970 publications, including over 100 invited papers and 74 patents. He has chaired or been a program-committee member for more than 130 domestic and international conferences organized by IEEE, SPIE (The International Society of Optical Engineering), OSA, and PSC. He has served as an editor, co-editor, or co-author for over twenty books. He has also served as a consultant for various federal agencies and private companies and delivered numerous invited talks to professional societies. He is a fellow of NAI, AIIA, OSA (Optica), and SPIE. He was a recipient of the 1987 UC Regent's Dissertation Fellowship and the 1999 UT Engineering Foundation Faculty Award for his contributions in research, teaching, and service. He received the Honorary Citizenship Award in 2003 from Austin City Council for his contribution in community service. He was also a recipient of the 2008 IEEE Teaching Award, the 2010 IEEE HKN Loudest Professor Award, and the 2013 NASA Certified Technical Achievement Award for contributions to moon surveillance conformable phased array antenna.

Jiaqi Gu (Member, IEEE) received the B.E. degree in microelectronic science and engineering from Fudan University, Shanghai, China, in 2018, and the Ph.D. degree from the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX, USA, in 2023. He is an Assistant Professor at the School of Electrical, Computer and Energy Engineering, Arizona State University (ASU), Tempe, AZ, USA. His current research interests include machine learning, efficient algorithms, and architecture design for high-performance AI, next-generation AI computing with emerging technology, and electronic-photonic design automation. He has published over 100 journal articles and refereed conference papers in the related field. He also served as a Technical Program Committee Member of DAC, ICCAD, ICCD, AAAI, and CLEO. He has received the Best Paper Award at ASP-DAC 2020, the Best Paper Finalist at DAC 2020, the Best Poster Award at NSF Workshop on Machine Learning Hardware in 2020, and the ACM Student Research Competition Grand Finals First Place in 2021, the Best Paper Award at IEEE TCAD 2021, won the Robert S. Hilbert Memorial Optical Design Competition 2022, the UT Austin-wide Outstanding Dissertation Award in Mathematics, Engineering, Physical Sciences, and Biological and Life Sciences in 2024, the Best Paper Award Candidate at DATE 2025, the Best Paper Nominee at ISPD in 2026, the NVIDIA Academic Grant Program Award in 2025, and the NSF CAREER Award in 2026. He served as a technical reviewer for over 20+ international journals/conferences, such as IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, TODAES, DAC, ICCAD, ISVLSI, GLVLSI, IEEE Transactions on Neural Networks and Learning Systems, NeurIPS, CVPR, ICML, ICCV, ECCV, AAAI, IROS, *Science*, *Nature Communications*, *Science Advances*, IEEE Journal of Selected Topics in Quantum Electronics, APL, and IEEE Photonics Technology Letters.

David Z. Pan (Fellow, IEEE) received the B.S. degree from Peking University in 1992 and the M.S. and Ph.D. degrees from the University of California, Los Angeles (UCLA), in 1998 and 2000, respectively. From 2000 to 2003, he was a Research Staff Member with the IBM T. J. Watson Research Center. His research interests include design automation for digital, analog, mixed-signal, RF ICs and emerging technologies, synergistic AI/IC co-optimizations, and design and technology/system co-optimizations. He has published over 550 technical papers in refereed journals and conferences, and is the holder of 10 U.S. patents. He has graduated 58 Ph.D./post-doctoral students at UT Austin, who are now holding key academic and industry positions. He is a fellow of ACM and SPIE. He has served in the Executive and Program Committees of many major conferences, including various leadership roles such as the ISPD 2007/2008 Program/General Chair, the ASP-DAC 2017 Program Chair, the ICCAD 2018/2019 Program/General Chair, the DAC 2023/2024 Technical Program Co-Chair/Chair, and the DAC 2026/2027 Vice/General Chair. He has received the 2013 SRC Technical Excellence Award, the DAC Top 10 Author in Fifth Decade, the DAC Prolific Author Award, the ASP-DAC Frequently Cited Author Award, the 21 Best Paper Awards (FCCM 2024, TCAD 2021, ISPD 2020,

ASP-DAC 2020, DAC 2019, GLSVLSI 2018, VLSI Integration 2018, HOST 2017, SPIE-AL 2016, ISPD 2014, ICCAD 2013, ASPDAC 2012, ISPD 2011, IBM Research Pat Goldberg Memorial Best Paper Award 2010 in CS/EE/Math, ASPDAC 2010, DATE 2009, ICICDT 2009, and SRC Techcon 1998, 2007, 2012, and 2015) and over 25 additional Best Paper Candidates or Outstanding/Honorable Mentions, Communications of the ACM Research Highlights in 2014, the ACM/SIGDA Outstanding New Faculty Award in 2005, the NSF CAREER Award in 2007, the UCLA Engineering Distinguished Young Alumnus Award in 2009, the UT Austin RAISE Faculty Excellence Award in 2014, the IBM Faculty Award four times, the SRC Inventor Recognition Award three times, the Cadence Academic Collaboration Award in 2019, the Nvidia Research Faculty Fellow in 2026, and a number of international CAD contest awards, among others. His students have won many awards, including the First Place of the ACM Student Research Competition Grand Finals twice in 2018 and 2021, the ACM/SIGDA Student Research Competition Gold Medal (three times), the ACM Outstanding Ph.D. Dissertation in EDA (twice), the EDAA Outstanding Dissertation Award (three times), and the UT Austin-wide Outstanding Dissertation Award in Mathematics, Engineering, Physical Sciences, and the Biological and Life Sciences in 2024. He has served as a Senior Associate Editor of *ACM Transactions on Design Automation of Electronic Systems* and an Associate Editor of *IEEE Design&Test*, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on Very Large Scale Integration, IEEE Transactions on Circuits and Systems—I: Regular Papers, and IEEE Transactions on Circuits and Systems—II: Express Briefs.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[2] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[3] J. Kaplan et al., "Scaling laws for neural language models," 2020, *arXiv:2001.08361*.

[4] J. Hoffmann et al., "Training compute-optimal large language models," 2022, *arXiv:2203.15556*.

[5] M. Bosma et al., "Chain-of-thought prompting elicits reasoning in large language models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 24824–24837.

[6] Y. Wu et al., "Inference scaling laws: An empirical analysis of compute-optimal inference for LLM problem-solving," in *Proc. Int. Conf. Learn. Represent.*, 2025, pp. 1–22.

[7] W. Cong and H. Zhu, "Can test-time scaling improve world foundation model?" 2025, *arXiv:2503.24320*.

[8] A. Jaech et al., "OpenAI o1 system card," 2024, *arXiv:2412.16720*.

[9] D. Guo et al., "DeepSeek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning," 2025, *arXiv:2501.12948*.

[10] G. E. Moore, "Cramming more components onto integrated circuits," *Proc. IEEE*, vol. 86, no. 1, pp. 82–85, Jan. 1998.

[11] M. Bohr, "A 30 year retrospective on dennard's MOSFET scaling paper," *IEEE Solid-State Circuits Soc. Newsletter*, vol. 12, no. 1, pp. 11–13, Feb. 2009.

[12] M. M. Waldrop, "More than Moore," *Nature*, vol. 530, no. 7589, pp. 144–148, Feb. 2016.

[13] F. Fang and N. Zhang, "Towards atomic and close-to-atomic scale manufacturing," *Int. J. Extreme Manuf.*, vol. 1, no. 1, Mar. 2019, Art. no. 012001.

[14] M. Horowitz, "Computing's energy problem," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, 2014, pp. 1–14.

[15] H. Esmaeilzadeh, E. Blem, R. St. Amant, K. Sankaralingam, and D. Burger, "Dark silicon and the end of multicore scaling," in *Proc. 38th Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2011, pp. 365–376.

[16] J. Shalf, "The future of computing beyond Moore's law," *Phil. Trans. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 378, p. 2166, Mar. 2020.

[17] Y. Wan et al., "Integrating silicon photonics with complementary metal-oxide-semiconductor technologies," *Nature Rev. Elect. Eng.*, vol. 3, pp. 1–17, Nov. 2025.

[18] P. L. McMahon, "The physics of optical computing," *Nature Rev. Phys.*, vol. 5, no. 12, pp. 717–734, Oct. 2023.

[19] X.-Y. Xu and X.-M. Jin, "Integrated photonic computing beyond the von neumann architecture," *ACS Photon.*, vol. 10, pp. 1027–1036, Feb. 2023.

[20] S. Ning et al., "Photonic-electronic integrated circuits for high-performance computing and AI accelerators," *J. Lightw. Technol.*, vol. 42, no. 22, pp. 7834–7859, Nov. 15, 2024.

[21] D. A. B. Miller, "Are optical transistors the logical next step?" *Nature Photon.*, vol. 4, no. 1, pp. 3–5, Jan. 2010.

[22] D. R. Solli and B. Jalali, "Analog optical computing," *Nature Photon.*, vol. 9, no. 11, pp. 704–706, Nov. 2015.

[23] S. R. Ahmed and R. Baghdadi, "Universal photonic artificial intelligence acceleration," *Nature*, vol. 640, no. 8058, pp. 368–374, 2025.

[24] A. Gholami et al., "A survey of quantization methods for efficient neural network inference," 2021, *arXiv:2103.13630*.

[25] T. Dettmers and M. Lewis, "GPT3.int8(): 8-bit matrix multiplication for transformers at scale," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 30318–30332.

[26] X. Xu et al., "11 TOPS photonic convolutional accelerator for optical neural networks," *Nature*, vol. 589, no. 7840, pp. 44–51, Jan. 2021.

[27] W. Heni and C. Haffner, "Plasmonic modulator enables <1 fJ/bit electro-optic conversion," *Science*, vol. 365, no. 6453, pp. 613–617, 2019.

[28] Y. Shen et al., "Deep learning with coherent nanophotonic circuits," *Nature Photon.*, vol. 11, no. 7, pp. 441–446, Jul. 2017.

[29] 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, Apr. 2024.

[30] I. Chakraborty and G. Saha, "Photonic in-memory computing primitive for spiking neural networks using phase-change materials," *Phys. Rev. Appl.*, vol. 11, no. 1, 2019, Art. no. 014063.

[31] J. Feldmann and N. Youngblood, "All-optical spiking neurosynaptic networks with self-learning capabilities," *Nature*, vol. 569, no. 7755, pp. 208–214, 2019.

[32] H. Zhu and J. Gu, "Lightening-transformer: A dynamically-operated optically-interconnected photonic transformer accelerator," in *Proc. HPCA*, Mar. 2024, pp. 686–703.

[33] B. J. Shastri et al., "Photonics for artificial intelligence and neuromorphic computing," *Nature Photon.*, vol. 15, no. 2, pp. 102–114, Feb. 2021.

[34] T. Fu and J. Zhang, "Optical neural networks: Progress and challenges," *Light, Sci. Appl.*, vol. 13, no. 1, p. 263, 2024.

[35] H. Zhang and Y. Song, "Integrated platforms and techniques for photonic neural networks," *npj Nanophotonics*, vol. 2, no. 1, p. 40, 2025.

[36] Z. Yin et al., "Simphony: A device-circuit-architecture cross-layer modeling and simulation framework for heterogeneous electronic-photonic ai system," 2024, *arXiv:2411.13715*.

[37] L. Thévenaz, "Slow and fast light in optical fibres," *Nature Photon.*, vol. 2, no. 8, pp. 474–481, Aug. 2008.

[38] D. M. Pozar, *Microwave Engineering: Theory and Techniques*. Hoboken, NJ, USA: Wiley, 2021.

[39] N. H. Weste and D. Harris, *CMOS VLSI Design: A Circuits and Systems Perspective*. London, U.K.: Pearson, 2015.

[40] S. Eyerman and L. Eeckhout, "Fine-grained DVFS using on-chip regulators," *ACM Trans. Archit. Code Optim. (TACO)*, vol. 8, no. 1, pp. 1–24, 2011.

[41] L. L. Ng et al., "Power consumption in CMOS circuits," in *Electromagnetic Field in Advancing Science and Technology*. London, U.K.: IntechOpen, 2022.

[42] D. J. Griffiths and D. F. Schroeter, *Introduction to Quantum Mechanics*. Cambridge, U.K.: Cambridge Univ. Press, 2018.
[43] B. E. Saleh and M. C. Teich, *Fundamentals of Photonics*. Hoboken, NJ, USA: Wiley, 2019.
[44] A. Rizzo et al., "Massively scalable kerr comb-driven silicon photonic link," *Nature Photon.*, vol. 17, no. 9, pp. 781–790, Sep. 2023.
[45] J. Feldmann et al., "Parallel convolutional processing using an integrated photonic tensor core," *Nature*, vol. 589, no. 7840, pp. 52–58, Jan. 2021.
[46] H. H. Zhu et al., "Space-efficient optical computing with an integrated chip diffractive neural network," *Nature Commun.*, vol. 13, no. 1, p. 1044, Feb. 2022.
[47] Z. Wang, L. Chang, F. Wang, T. Li, and T. Gu, "Integrated photonic metasystem for image classifications at telecommunication wavelength," *Nature Commun.*, vol. 13, no. 1, p. 2131, Apr. 2022.
[48] NVIDIA. (Jan. 2020). *NVIDIA V100 Tensor Core GPU Datasheet*. [Online]. Available: https://images.nvidia.com/content/technologies/volta/
[49] NVIDIA. (May 2022). *NVIDIA A100 Tensor Core GPU Datasheet*. [Online]. Available: https: //www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-nvidia-us-2188504-web.pdf
[50] NVIDIA. (Feb. 2023). *NVIDIA H100 Tensor Core GPU Datasheet*. [Online]. Available: https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/
[51] NVIDIA. (Jul. 2025). *PCF Summary for NVIDIA HGX B200 (Datasheet)*. [Online]. Available: https://images.nvidia.com/aem-dam/Solutions/documents/HGX-B200-PCF-Summary.pdf
[52] A. N. Tait et al., "Microring weight banks," *IEEE J. Sel. Topics Quantum Electron.*, vol. 22, no. 6, pp. 312–325, Nov. 2016.
[53] M. Miscuglio and V. J. Sorger, "Photonic tensor cores for machine learning," *Appl. Phys. Rev.*, vol. 7, no. 3, Sep. 2020, Art. no. 031404.
[54] C. Feng et al., "A compact butterfly-style silicon photonic-electronic neural chip for hardware-efficient deep learning," *ACS Photon.*, vol. 9, no. 12, pp. 3906–3916, Dec. 2022.
[55] S. Ning et al., "Hardware-efficient photonic tensor core: Accelerating deep neural networks with structured compression," *Optica*, vol. 12, no. 7, pp. 1079–1089, 2025.
[56] M. Zhang et al., "TeMPO: Efficient time-multiplexed dynamic photonic tensor core for edge AI with compact slow-light electro-optic modulator," *J. Appl. Physic*, vol. 135, no. 22, 2024, Art. no. 223105.
[57] Z. Yin et al., "SCATTER: Algorithm-circuit co-sparse photonic accelerator with thermal-tolerant, power-efficient in-situ light redistribution," in *Proc. ICCAD*, 2024, pp. 1–9.
[58] C. Demirkiran et al., "An electro-photonic system for accelerating deep neural networks," *ACM J. Emerg. Technol. Comput. Syst.*, vol. 19, no. 4, pp. 1–31, Sep. 2023.
[59] J. Gu et al., "SqueezeLight: A multi-operand ring-based optical neural network with cross-layer scalability," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 42, no. 3, pp. 807–819, Mar. 2023.
[60] X. Xiao et al., "TOMFuN: A tensorized optical multimodal fusion network," *APL Mach. Learn.*, vol. 3, no. 1, Mar. 2025, Art. no. 016121.
[61] S. Fei, A. Eldebiky, G. L. Zhang, B. Li, and U. Schlichtmann, "An efficient general-purpose optical accelerator for neural networks," in *Proc. 30th Asia South Pacific Design Autom. Conf.*, Jan. 2025, pp. 1070–1076.
[62] M. Morsali et al., "Oisa: Architecting an optical in-sensor accelerator for efficient visual computing," in *Proc. DATE*, 2024, pp. 1–6.
[63] D. Wang et al., "Ultrafast silicon photonic reservoir computing engine delivering over 200 TOPS," *Nature Commun.*, vol. 15, no. 1, p. 10841, Dec. 2024.
[64] X. Li et al., "NEOCNN: NTT-enabled optical convolution neural network accelerator," in *Proc. 38th ACM Int. Conf. Supercomputing*, May 2024, pp. 352–362.
[65] S. Sun et al., "Highly efficient photonic convolver via lossless mode-division fan-in," *Nature Commun.*, vol. 16, no. 1, p. 7513, Aug. 2025.
[66] X. Xiao et al., "Large-scale and energy-efficient tensorized optical neural networks on III–V-on-silicon MOSCAP platform," *APL Photon.*, vol. 6, no. 12, Dec. 2021, Art. no. 126107.
[67] W. Zhou et al., "In-memory photonic dot-product engine with electrically programmable weight banks," *Nature Commun.*, vol. 14, no. 1, p. 2887, May 2023.
[68] S. Ou et al., "Hypermultiplexed integrated photonics-based optical tensor processor," *Sci. Adv.*, vol. 11, no. 23, 2025, Art. no. eadu0228.
[69] J. Kim, "Photonic systolic array for all-optical matrix-matrix multiplication," *Laser Photon. Rev.*, vol. 20, no. 7, 2025, Art. no. 01995.
[70] C. Feng et al., "Integrated multi-operand optical neurons for scalable and hardware-efficient deep learning," *Nanophotonics*, vol. 13, no. 12, pp. 2193–2206, May 2024.
[71] X. Meng et al., "Compact optical convolution processing unit based on multimode interference," *Nature Commun.*, vol. 14, no. 1, p. 3000, May 2023.
[72] J. Cheng et al., "Multimodal deep learning using on-chip diffractive optics with in situ training capability," *Nature Commun.*, vol. 15, no. 1, p. 6189, Jul. 2024.
[73] A. N. Tait et al., "Neuromorphic photonic networks using silicon photonic weight banks," *Sci. Rep.*, vol. 7, no. 1, p. 7430, Aug. 2017.
[74] J. Gu et al., "Squeezelight: Towards scalable optical neural networks with multi-operand ring resonators," in *Proc. DATE*, Mar. 2023, pp. 238–243.
[75] S. Ning et al., "Microring-based multi-operand optical neurons with on-chip trainable nonlinearity," in *Proc. CLEO*, 2025, p. 120.
[76] J. Gu, H. Zhu, C. Feng, Z. Jiang, R. T. Chen, and D. Z. Pan, "M3ICRO: Machine learning-enabled compact photonic tensor core based on programmable multi-operand multimode interference," *APL Mach. Learn.*, vol. 2, no. 1, Mar. 2024, Art. no. 016106.
[77] J. Li et al., "End-to-end closed-loop optoelectronic computing breaking precision–accuracy coupling," *Adv. Photon.*, vol. 8, no. 1, p. 016, Dec. 2025.
[78] Z. Wang et al., "On-chip wavefront shaping with dielectric metasurface," *Nature Commun.*, vol. 10, no. 1, p. 3547, Aug. 2019.
[79] C. Wang et al., "Diffractive tensorized unit for million-TOPS general-purpose computing," *Nature Photon.*, vol. 19, no. 10, pp. 1078–1087, Oct. 2025.
[80] Y. Wang et al., "On-chip reconfigurable diffractive optical neural network based on Sb_2S_3," *Opt. Express*, vol. 33, no. 2, pp. 1810–1826, 2025.
[81] Y. Bai et al., "TOPS-speed complex-valued convolutional accelerator for feature extraction and inference," *Nature Commun.*, vol. 16, no. 1, p. 292, Jan. 2025.
[82] S. Xu et al., "High-order tensor flow processing using integrated photonic circuits," *Nature Commun.*, vol. 13, no. 1, p. 7970, Dec. 2022.
[83] R. Yin et al., "Integrated WDM-compatible optical mode division multiplexing neural network accelerator," *Optica*, vol. 10, no. 12, pp. 1709–1718, 2023.
[84] Z. Lin et al., "120 GOPS photonic tensor core in thin-film lithium niobate for inference and in situ training," *Nature Commun.*, vol. 15, no. 1, p. 9081, Oct. 2024.
[85] S. Han et al., "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," in *Proc. ICLR*, 2016, pp. 1–14.
[86] P. Molchanov et al., "Pruning convolutional neural networks for resource efficient inference," in *Proc. ICLR*, 2016, pp. 1–17.
[87] C. Ding et al., "CirCNN: Accelerating and compressing deep neural networks using block-circulant weight matrices," in *Proc. 50th Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, Oct. 2017, pp. 395–408.
[88] H. Gong et al., "LightMat-HP: A photonic-electronic system for accelerating general matrix multiplication with configurable precision," 2026, *arXiv:2604.12278*.
[89] C. Demirkiran et al., "A blueprint for precise and fault-tolerant analog neural networks," *Nature Commun.*, vol. 15, no. 1, p. 5098, Jun. 2024.
[90] C. Demirkiran et al., "Mirage: An RNS-based photonic accelerator for DNN training," in *Proc. ISCA*, Jul. 2024, pp. 73–87.
[91] C. Zhang et al., "Nonvolatile multilevel switching of silicon photonic devices with In_2O_3/GST segmented structures," *Adv. Optical Mater.*, vol. 11, no. 8, 2023, Art. no. 2202748.
[92] J. Xia et al., "Seven bit nonvolatile electrically programmable photonics based on phase-change materials for image recognition," *ACS Photon.*, vol. 11, no. 2, pp. 723–730, Feb. 2024.
[93] J. Geler-Kremer et al., "A ferroelectric multilevel non-volatile photonic phase shifter," *Nature Photon.*, vol. 16, no. 7, pp. 491–497, Jul. 2022.
[94] C. Errando-Herranz et al., "MEMS for photonic integrated circuits," *IEEE J. Sel. Topics Quantum Electron.*, vol. 26, no. 2, pp. 1–16, Mar. 2020.
[95] A. Unamuno and D. Uttamchandani, "MEMS variable optical attenuator with Vernier latching mechanism," *IEEE Photon. Technol. Lett.*, vol. 18, no. 1, pp. 88–90, Jan. 2006.
[96] L. Martin-Monier et al., "Endurance of chalcogenide optical phase change materials: A review," *Opt. Mater. Exp.*, vol. 12, no. 6, pp. 2145–2167, 2022.

[97] H. Zhu et al., "Elight: Toward efficient and aging-resilient photonic in-memory neurocomputing," *IEEE Trans. Comput. -Aided Design Integr. Circuits Syst.*, vol. 42, no. 3, pp. 820–833, Mar. 2022.
[98] S. Lam et al., "Neuromorphic photonic computing with an electro-optic analog memory," 2024, *arXiv:2401.16515*.
[99] F. Sunny et al., "CrossLight: A cross-layer optimized silicon photonic neural network accelerator," in *Proc. 58th ACM/IEEE Design Autom. Conf. (DAC)*, Dec. 2021, pp. 1069–1074.
[100] K. Shiflett et al., "Albireo: energy-efficient acceleration of convolutional neural networks via silicon photonics," in *Proc. ACM/IEEE 48th Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2021, pp. 860–873.
[101] S. R. Kari et al., "Realization of an integrated coherent photonic platform for scalable matrix operations," *Optica*, vol. 11, no. 4, pp. 542–551, 2024.
[102] J. Lim and D. Psaltis, "MaxwellNet: Physics-driven deep neural network training based on Maxwell's equations," *APL Photon.*, vol. 7, no. 1, Jan. 2022, Art. no. 011301.
[103] M. Chen, R. Lupoiu, C. Mao, D.-H. Huang, J. Jiang, P. Lalanne, J. A. Fan, "WaveY-Net: Physics-augmented deep-learning for high-speed electromagnetic simulation and optimization," *Proc. SPIE*, vol. 12011, Mar. 2022, Art. no. 120110C, doi: 10.1117/12.2612418.
[104] J. Gu et al., "Neurolight: A physics-agnostic neural operator enabling parametric photonic device simulation," in *Proc. NeurIPS*, 2022, pp. 1–14.
[105] G. Chen et al., "PACE: Pacing operator learning to accurate optical field simulation for complicated photonic devices," in *Proc. Adv. Neural Inf. Process. Syst. 37*. Red Hook, NY, USA: Curran Associates Inc., 2024, pp. 67535–67555.
[106] C. Mao and J. A. Fan, "Accurate and scalable deep Maxwell solvers using multilevel iterative methods," 2025, *arXiv:2509.03622*.
[107] P. Ma, H. Yang, Z. Gao, D. S. Boning, and J. Gu, "PIC2O-Sim: A physics-inspired causality-aware dynamic convolutional neural operator for ultra-fast photonic device time-domain simulation," *APL Photonics*, vol. 10, no. 3, Mar. 2025, Art. no. 036104.
[108] K. Azizzadenesheli et al., "Neural operators for accelerating scientific simulations and design," *Nature Rev. Phys.*, vol. 6, no. 5, pp. 320–328, Apr. 2024.
[109] H. Zhang et al., "Time-domain 3D electromagnetic fields estimation based on physics-informed deep learning framework," in *Proc. DATE*, 2025, pp. 1–7.
[110] Z. Gao et al., "SPIPE: Differentiable SPICE-level co-simulation program for integrated photonics and electronics," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 45, no. 3, pp. 1487–1498, Mar. 2026.
[111] C. Sorace-Agaskar et al., "Electro-optical co-simulation for integrated CMOS photonic circuits with verilogа," *Opt. Exp.*, vol. 23, no. 21, p. 27, Oct. 2015.
[112] M. J. Shawon and V. Saxena, "Rapid simulation of photonic integrated circuits using verilog–A compact models," in *Proc. IEEE 62nd Int. Midwest Symp. Circuits Syst. (MWSCAS)*, Aug. 2019, pp. 424–427.
[113] Y. Fu, Y. Liu, N. Wong, and J. Xu, "Invited paper: SPICE-compatible modeling and design for electronic-photonic integrated circuits," in *Proc. 30th Asia South Pacific Design Automat. Conf. (ASPDAC)*. New York, NY, USA: Association for Computing Machinery, 2025, pp. 135–140, doi: 10.1145/3658617.3703152.
[114] D. Azhigulov et al., "Enabling data-driven and bidirectional model development in verilog—A for photonic devices," *Opt. Exp.*, vol. 32, no. 17, pp. 29965–29975, Aug. 2024.
[115] M. Li et al., "O-HAS: Optical hardware accelerator search for boosting both acceleration performance and development speed," in *Proc. ICCAD*, 2021, pp. 1–9.
[116] Z. Yin et al., "H3PIMAP: A heterogeneity-aware multi-objective DNN mapping framework on electronic-photonic processing-in-memory architectures," 2026, *arXiv:2503.07778*.
[117] T. Zhao et al., "Highly efficient inverse design of semiconductor optical amplifiers based on neural network improved particle swarm optimization algorithm," *IEEE Photon. J.*, vol. 15, no. 2, pp. 1–9, Apr. 2023.
[118] Z. Wang et al., "Efficient inverse design method of AWG based on BPNN-PSO algorithm," *Opt. Commun.*, vol. 552, Feb. 2024, Art. no. 130080.
[119] J. Liao et al., "Inverse design of ultra-compact and low-loss optical phase shifters," *Photonics*, vol. 10, no. 9, p. 1030, Sep. 2023.
[120] S. Chebaane et al., "Machine learning-based inverse design of raised cosine few mode fiber for low coupling," *Opt. Quantum Electron.*, vol. 56, no. 1, p. 56, Jan. 2024.
[121] Z. Gao, Z. Zhang, and D. S. Boning, "Automatic synthesis of broadband silicon photonic devices via Bayesian optimization," *J. Lightw. Technol.*, vol. 40, no. 24, pp. 7879–7892, Sep. 15, 2022.
[122] J. Liao et al., "Inverse design of highly efficient and broadband mode splitter on SOI platform," *Chin. Optics Lett.*, vol. 22, no. 1, 2024, Art. no. 011302.
[123] Y. Liu et al., "Inverse design of multi-port power splitter with arbitrary ratio based on shape optimization," *Nanomaterials*, vol. 15, no. 5, p. 393, Mar. 2025.
[124] J. Marqués-Hueso et al., "Genetic algorithm designed silicon integrated photonic lens operating at 1550 nm," *Appl. Phys. Lett.*, vol. 97, no. 7, Aug. 2010, Art. no. 071115.
[125] R. Li et al., "Deep reinforcement learning empowers automated inverse design and optimization of photonic crystals for nanoscale laser cavities," *Nanophotonics*, vol. 12, no. 2, pp. 319–334, Jan. 2023.
[126] N. V. Sapra et al., "Inverse design and demonstration of broadband grating couplers," *IEEE J. Sel. Topics Quantum Electron.*, vol. 25, no. 3, pp. 1–7, May 2019.
[127] Z. Zhu et al., "PSO-aided inverse design of silicon modulator," *IEEE Photon. J.*, vol. 16, no. 2, pp. 1–5, Apr. 2024.
[128] Z. Yu, H. Cui, and X. Sun, "Genetic-algorithm-optimized wideband on-chip polarization rotator with an ultrasmall footprint," *Opt. Lett.*, vol. 42, no. 16, p. 3093, Aug. 2017.
[129] T. Muratsubaki et al., "Direct-binary-search algorithm for fabrication-tolerant photonic-crystal-like subwavelength structures and its application to a four-mode waveguide crossing in 2 μm waveband," *Jpn. J. Appl. Phys.*, vol. 61, no. 4, Apr. 2022, Art. no. 042003.
[130] K. Wang et al., "Inverse-designed photonic computing core for parallel matrix-vector multiplication," *J. Lightw. Technol.*, vol. 42, no. 22, pp. 8061–8071, Nov. 15, 2024.
[131] C. Wu et al., "Reconfigurable inverse-designed phase-change photonics," *APL Photonics*, vol. 10, no. 1, Jan. 2025, Art. no. 016113.
[132] Y. Tang et al., "Generative deep learning model for inverse design of integrated nanophotonic devices," *Laser Photon. Rev.*, vol. 14, no. 12, Dec. 2020, Art. no. 2000287.
[133] S. Mao et al., "Multi-task topology optimization of photonic devices in low-dimensional Fourier domain via deep learning," *Nanophotonics*, vol. 12, no. 5, pp. 1007–1018, Mar. 2023.
[134] A. Y. Piggott et al., "Inverse design and demonstration of a compact and broadband on-chip wavelength demultiplexer," *Nature Photon.*, vol. 9, no. 6, pp. 374–377, Jun. 2015.
[135] V. Nikkhah et al., "Inverse-designed low-index-contrast structures on a silicon photonics platform for vector-matrix multiplication," *Nature Photonics*, vol. 18, no. 5, pp. 501–508, May 2024.
[136] T. W. Hughes et al., "Adjoint method and inverse design for nonlinear nanophotonic devices," *ACS Photon.*, vol. 5, no. 12, pp. 4781–4787, Dec. 2018.
[137] J. Wang et al., "Silicon mode (de) multiplexer enabling high capacity photonic networks-on-chip with a single-wavelength-carrier light," *Optics letters*, vol. 38, no. 9, pp. 1422–1424, 2013.
[138] K. Y. Yang et al., "Multi-dimensional data transmission using inverse-designed silicon photonics and microcombs," *Nature Commun.*, vol. 13, no. 1, p. 7862, Dec. 2022.
[139] Y. Xie et al., "Complex-valued matrix-vector multiplication using a scalable coherent photonic processor," *Sci. Adv.*, vol. 11, no. 14, p. 7475, Apr. 2025.
[140] L. T. Wang et al., *Electronic Design Automation: Synthesis, Verification, and Test*. San Mateo, CA, USA: Morgan Kaufmann, 2009.
[141] G. Huang et al., "Machine learning for electronic design automation: A survey," *ACM TODAES*, vol. 26, no. 5, pp. 1–46, 2021.
[142] Y. Su et al., "Machine learning-assisted design automation of integrated photonic devices," in *Proc. Int. Symp. Electron. Design Autom. (ISEDA)*, May 2025, pp. 718–723.
[143] R. Marzban, A. Adibi, and R. Pestourie, "Inverse design in nanophotonics via representation learning," *Adv. Opt. Mater.*, vol. 14, no. 1, p. 02062, Nov. 2025.
[144] Y. Xie et al., "Design of an arbitrary ratio optical power splitter based on a discrete differential multiobjective evolutionary algorithm," *Appl. Opt.*, vol. 59, no. 6, pp. 1780–1785, 2020.
[145] E. Zhang et al., "Improved particle swarm optimization with less manual intervention for photonic inverse design," *IEEE Photon. Technol. Lett.*, vol. 35, no. 24, pp. 1355–1358, Dec. 15, 2023.
[146] M. Hansi et al., "Silicon-based on-chip optical devices based on direct binary search algorithm," *Opto-Electron. Eng.*, vol. 52, no. 11, p. 250, 2025.

[147] D. Givoli, "A tutorial on the adjoint method for inverse problems," *Comput. Methods Appl. Mech. Eng.*, vol. 380, Jul. 2021, Art. no. 113810.
[148] M. F. Schubert et al., "Inverse design of photonic devices with strict foundry fabrication constraints," *ACS Photon.*, vol. 9, no. 7, pp. 2327–2336, Jul. 2022.
[149] P. Ma et al., "Boson^{-1}: Understanding and enabling physically-robust photonic inverse design with adaptive variation-aware subspace optimization," in *Proc. DATE*, Apr. 2025, pp. 1–7.
[150] W. Kim et al., "Inverse design of nanophotonic devices using generative adversarial networks," *Eng. Appl. Artif. Intell.*, vol. 115, Oct. 2022, Art. no. 105259.
[151] W. Ma et al., "Probabilistic representation and inverse design of metamaterials based on a deep generative model with semi-supervised learning strategy," *Adv. Mater.*, vol. 31, no. 35, Aug. 2019, Art. no. 1901111.
[152] M. Chen, J. Jiang, and J. A. Fan, "Design space reparameterization enforces hard geometric constraints in inverse-designed nanophotonic devices," *ACS Photon.*, vol. 7, no. 11, pp. 3141–3151, Nov. 2020.
[153] E. Gershnabel et al., "Reparameterization approach to gradient-based inverse design of three-dimensional nanophotonic devices," *ACS Photon.*, pp. 815–823, Oct. 2022.
[154] A. M. Hammond et al., "High-performance hybrid time/frequency-domain topology optimization for large-scale photonics inverse design," *Opt. Exp.*, vol. 30, no. 3, pp. 4467–4491, 2022.
[155] E. Khoram, X. Qian, M. Yuan, and Z. Yu, "Controlling the minimal feature sizes in adjoint optimization of nanophotonic devices using b-spline surfaces," *Opt. Exp.*, vol. 28, no. 5, pp. 7060–7069, 2020.
[156] E. W. Wang et al., "Robust design of topology-optimized metasurfaces," *Optical Mater. Exp.*, vol. 9, no. 2, pp. 469–482, 2019.
[157] F. Wang, J. S. Jensen, and O. Sigmund, "Robust topology optimization of photonic crystal waveguides with tailored dispersion properties," *J. Opt. Soc. Amer. B*, vol. 28, no. 3, pp. 387–397, 2011.
[158] M. Schevenels et al., "Robust topology optimization accounting for spatially varying manufacturing errors," *Comput. Methods Appl. Mech. Eng.*, vol. 200, nos. 49–52, pp. 3613–3627, 2011.
[159] A. M. Hammond et al., "Photonic topology optimization with semiconductor-foundry design-rule constraints," *Opt. Exp.*, vol. 29, no. 15, pp. 23916–23938, 2021.
[160] M. Kim, H. Park, and J. Shin, "Nanophotonic device design based on large language models: Multilayer and metasurface examples," *Nanophotonics*, vol. 14, no. 8, pp. 1273–1282, Feb. 2025.
[161] J. Gu et al., "ADEPT: Automatic differentiable design of photonic tensor cores," in *Proc. DAC*, Jul. 2022, pp. 1–6.
[162] Z. Jiang et al., "ADEPT-Z: Zero-shot automated circuit topology search for Pareto-optimal photonic tensor cores," in *Proc. ASPDAC*, 2025, pp. 1077–1083.
[163] W. Bogaerts and L. Chrostowski, "Silicon photonics circuit design: Methods, tools and challenges," *Laser Photon. Rev.*, vol. 12, no. 4, Apr. 2018, Art. no. 1700237.
[164] J. R. Minz, S. Thyagara, and S. K. Lim, "Optical routing for 3-D system-on-package," *IEEE Trans. Compon. Packag. Technol.*, vol. 30, no. 4, pp. 805–812, Dec. 2007.
[165] D. Ding and D. Z. Pan, "OIL: A nano-photonics optical interconnect library for a new photonic networks-on-chip architecture," in *Proc. SLIP*, Jul. 2009, pp. 11–18.
[166] D. Ding et al., "O-router: An optical routing framework for low power on-chip silicon nano-photonic integration," in *Proc. DAC*, Jul. 2009, pp. 264–269.
[167] G. Hendry et al., "Vandal: A tool for the design specification of nano-photonic networks," in *Proc. Design, Autom. Test Eur.*, Mar. 2011, , pp. 1–6.
[168] C. Condrat, P. Kalla, and S. Blair, "A methodology for physical design automation for integrated optics," in *Proc. IEEE 55th Int. Midwest Symp. Circuits Syst. (MWSCAS)*, Aug. 2012, pp. 598–601.
[169] S. Sharma and S. Roy, "Optimizing bend loss in optical waveguide channel routing on photonic integrated circuits," *J. Comput. Electron.*, vol. 22, no. 1, pp. 350–363, Dec. 2022.
[170] C. Condrat et al., "Channel routing for integrated optics," in *Proc. SLIP*, Jun. 2013, pp. 1–8.
[171] A. Boos et al., "PROTON: An automatic place-and-route tool for optical networks-on-chip," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, Nov. 2013, pp. 138–145.
[172] A. von Beuningen and U. Schlichtmann, "PLATON: A force-directed placement algorithm for 3D optical networks-on-chip," in *Proc. Int. Symp. Phys. Design*, Apr. 2016, pp. 27–34.
[173] Y.-K. Chuang et al., "PlanarONoC: Concurrent placement and routing considering crossing minimization for optical networks-on-chip," in *Proc. 55th ACM/ESDA/IEEE Design Autom. Conf. (DAC)*, Jun. 2018, pp. 1–6.
[174] Y. T. Chen et al., "CPONoC: Critical path-aware physical implementation for optical networks-on-chip," in *Proc. ASPDAC*, 2025, pp. 1251–1256.
[175] F.-Y. Chuang and Y.-W. Chang, "On-chip optical routing with waveguide matching constraints," in *Proc. IEEE/ACM Int. Conf. Comput. Aided Design (ICCAD)*, Nov. 2021, pp. 1–6.
[176] Z. Zheng et al., "ToPro: A topology projector and waveguide router for wavelength-routed optical Networks-on-Chip," in *Proc. ICCAD*, Nov. 2021, pp. 1–9.
[177] L. Chrostowski et al., "Schematic driven silicon photonics design," *Proc. SPIE*, vol. 9751, pp. 9–22, Mar. 2016.
[178] L. Chrostowski et al., "Design and simulation of silicon photonic schematics and layouts," *Proc. SPIE*, vol. 9891, pp. 185–195, May 2016.
[179] J. Matres et al. (2024). *Gdsfactory*. [Online]. Available: https://github.com/gdsfactory/gdsfactory
[180] H. Zhou et al., "LiDAR: Automated curvy waveguide detailed routing for large-scale photonic integrated circuits," in *Proc. Int. Symp. Phys. Design*, 2025, pp. 64–72.
[181] H. Zhou et al., "LiDAR 2.0: Hierarchical curvy waveguide detailed routing for large-scale photonic integrated circuits," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, early access, Oct. 22, 2025, doi: /TCAD.2025.3624197.
[182] Y. Wu et al., "Automatic routing for photonic integrated circuits under delay matching constraints," in *Proc. Design, Autom. Test Eur. Conf. (DATE)*, Mar. 2025, pp. 1–2.
[183] Y. Wu et al., "Constraints-aware adaptive routing with hybrid waveguides for photonic integrated circuits," in *Proc. IEEE/ACM Int. Conf. Comput. Aided Design (ICCAD)*, Oct. 2025, pp. 1–8.
[184] X. Jiang et al., "PICELF: An automatic electronic layer layout generation framework for photonic integrated circuits," in *Proc. Design, Autom. Test Eur. Conf. (DATE)*, Mar. 2025, pp. 1–7.
[185] H. Zhou et al., "Photonics-aware planning-guided automated electrical routing for large-scale active photonic integrated circuits," 2025, *arXiv:2509.23764*.
[186] H. Zhou et al., "Apollo: Automated routing-informed placement for large-scale photonic integrated circuits," in *Proc. IEEE/ACM Int. Conf. Comput. Aided Design (ICCAD)*, Oct. 2025, pp. 1–9.
[187] A. Sharma, Y. Fu, V. Ansari, R. Iyer, F. Kuang, K. Mistry, R. I. Aishy, S. Ahmad, J. Matres, D. R. Englund, and J. K. S. Poon, "AI agents for photonic integrated circuit design automation," *APL Mach. Learn.*, vol. 3, no. 4, Dec. 2025, Art. no. 046113.
[188] L. Bajic. (Feb. 2026). *The Path To Ubiquitous AI*. [Online]. Available: https://taalas.com/the-path-to-ubiquitous-ai/
[189] T. Wang et al., "Image sensing with multilayer nonlinear optical neural networks," *Nature Photon.*, vol. 17, no. 5, pp. 408–415, 2023.
[190] Y. Baek et al., "Edge intelligence through in-sensor and near-sensor computing for the artificial intelligence of things," *npj Unconventional Comput.*, vol. 2, no. 1, p. 25, Oct. 2025.
[191] K. Shiflett et al., "Flumen: Dynamic processing in the photonic interconnect," in *Proc. 50th Annu. Int. Symp. Comput. Archit.*, Jun. 2023, pp. 1–13.
[192] M. Yang, Z. Zhong, and M. Ghobadi, "On-fiber photonic computing," in *Proc. 22nd ACM Workshop Hot Topics Netw.*, Nov. 2023, pp. 263–271.