Closed-Loop Latent Surrogate Modeling of Thermally Coupled VCSEL Dynamics for Waveform-Level Circuit Simulation

Alireza Pourafzal[1], Siavash Mowlavi[2], Muralikrishnan Srinivasan[3], Lars Svensson[2], Peter A. Andrekson[4], and Henk Wymeersch[1]

[1] Department of Electrical Engineering, Chalmers University of Technology, Gothenburg, Sweden
[2] Microwave Electronics Laboratory, Department of Microtechnology and Nanoscience, Chalmers University of Technology, Gothenburg, Sweden
[3] Indian Institute of Technology (BHU), Varanasi, India
[4] Photonics Laboratory, Department of Microtechnology and Nanoscience, Chalmers University of Technology, Gothenburg, Sweden

Abstract

We present a compact data-driven surrogate model for waveform-level simulation of thermally coupled VCSEL dynamics, motivated by fabrication-stage studies in which repeated transient evaluation with high-fidelity electro-thermal VCSEL models can become costly and cumbersome. The proposed approach identifies an empirical finite-memory representation from simulator-generated waveform data, and then learns a shared nonlinear latent dynamical model with lightweight temperature-specific affine readouts. Training is performed using a closed-loop rollout objective so that the surrogate is optimized for recursive self-generated operation rather than only one-step prediction. Simulation results based on a high-fidelity electro-thermal VCSEL model at 20 Gbaud show that the surrogate can reproduce optical waveforms accurately over 1200-sample free-running rollouts across temperatures from -40°C to 80°C. The learned encoder contains $35,680$ trainable parameters, corresponding to approximately 0.136 MB in single precision, while each temperature-specific affine readout adds only 33 parameters. The results further show that a shared latent representation transfers across temperatures through lightweight per-temperature readout calibration, and also transfers effectively across bias conditions with minimal adaptation. These findings support closed-loop latent dynamical surrogate modeling as a compact approach for repeated waveform-level evaluation of thermally coupled VCSEL dynamics.

Keywords: vertical-cavity surface-emitting lasers (VCSELs); electro-thermal modeling; data-driven surrogate modeling; waveform-level simulation; waveform-level circuit evaluation; transient waveform evaluation; closed-loop rollout training

1 Introduction

Vertical-cavity surface-emitting lasers (VCSELs) are key enabling technologies of short-reach optical interconnects because of their low cost, compact footprint, low power consumption, and ability to support high-speed direct modulation [1–4]. These features make VCSEL-based links well suited to data centers and high-performance computing systems, where dense, energy-efficient, and high-capacity interconnects are essential [1, 5]. However, as optical engines are integrated closer to heat-generating electronic circuitry, the thermal operating conditions of such links become increasingly challenging [6, 7]. Similar constraints also arise in harsh-environment applications such as automotive optical links, where wide temperature ranges must be tolerated [1, 5–7]. These thermal variations can significantly alter key VCSEL characteristics, including threshold current, emitted optical power, wavelength, thermal resistance, and modulation response [8–10]. Therefore, accurate temperature-aware VCSEL modeling becomes essential for pre-fabrication circuit simulation, waveform verification, and robust link design.

Temperature modeling becomes especially significant during pre-fabrication circuit design, where the VCSEL must be represented inside the simulation loop of the driver and associated front-end circuitry [11]. In such design workflows, the VCSEL is not simply a static optical source or equivalent load, but a

nonlinear electro-optical device whose transient response depends on biasing, modulation waveform, and thermal operating point [12]. Static current-voltage or current-power characteristics alone are insufficient for this purpose, since the VCSEL response under modulation is governed by coupled carrier-photon and thermal dynamics. Designers therefore require models that can reproduce realistic optical output waveforms under large-signal excitation, long pseudo-random bit sequence (PRBS) operation, and temperature sweeps [13]. These transient simulations are essential for assessing waveform quality, pattern dependence, and operating-margin robustness before fabrication.

A major challenge, however, is that high-fidelity physics-based VCSEL models are computationally expensive. Detailed electro-optical and electro-thermal descriptions can capture the relevant device dynamics with high accuracy, but their direct use inside large circuit simulations often leads to long runtimes and, in some cases, numerical or convergence difficulties [14]. The burden becomes particularly severe in long transient simulations with PRBS excitation, where realistic verification requires extended sequences to capture pattern-dependent distortions and thermally induced waveform variations [15]. Thus, although such detailed models remain indispensable, they are often too complex for rapid iterative study during fabrication-stage driver development [16].

This motivates a useful distinction between three modeling directions. High-fidelity electro-thermal and rate-equation-based models provide detailed physical descriptions but can be costly for repeated long transient evaluation. Compact and circuit-level models, including equivalent-circuit and hardware-description-language implementations, are more convenient for circuit simulators and driver co-simulation, but their accuracy depends on the retained physical effects and parameter calibration [17, 18]. Data-driven models can learn input–output behavior directly from simulated or measured waveforms, but without suitable dynamical structure or training objectives, they may behave as black-box local regressors rather than stable waveform generators.

Therefore, it is imperative to develop models that preserve the dominant waveform-level dynamics of the VCSEL while remaining compact enough for repeated transient evaluation. Ideally, such a model should remain accurate under large-signal operation, capture thermal dependence explicitly, support stable waveform generation over extended temporal windows under recursive prediction, and admit efficient evaluation within practical design studies. Machine learning offers a promising direction in this regard. More broadly, ML has emerged as an attractive alternative to purely model-based communication design when accurate analytical models are difficult to derive, overly complex to optimize over, or insufficiently adaptive to changing operating conditions [19]. In optical communication systems, receiver-side algorithms for equalization, synchronization, data detection, and impairment compensation have already been learned by either mimicking conventional signal-processing pipelines or training deep neural networks more directly [20–24]. On the transmitter side, learning-based pre-equalization and predistortion have also attracted increasing interest as a means to cope with bandwidth limitations and nonlinearities more effectively [25–29]. These developments suggest that suitably structured learned surrogates may provide an effective compromise between physical fidelity and computational efficiency. Similar computational-efficiency pressures also appear in other areas of photonics, where numerical optimization methods are used to reduce the cost of repeated device evaluation, for example in grating-based optical-component design [30].

At the same time, the present problem differs from many existing ML tasks in optical communications and photonics. Here, the goal is not primarily end-to-end transceiver optimization, adaptive equalization, static classification, waveform synthesis, or one-step waveform fitting, but the construction of a compact surrogate that can faithfully emulate thermally dependent VCSEL waveform dynamics for circuit-level transient simulation. The surrogate must preserve recursive prediction quality over extended waveform durations, remain valid across temperatures, and stay sufficiently simple for repeated integration into a circuit-design workflow. This introduces a stability gap: a predictor trained only on measured past outputs may perform well in one-step testing, but during autonomous waveform generation it is driven by its own previous predictions. Small local errors can then accumulate and lead to amplitude drift, phase drift, or unstable waveform evolution over long rollouts. Addressing this gap is especially important in transmitter-side modeling, where learning and optimization are already known to be challenging when a differentiable and sufficiently accurate device model is not readily available [31].

Although neural-network-based and other learned approximations of VCSEL behavior have been explored [16, 29, 32], the present setting poses a different challenge. The objective here is not merely to fit

device behavior locally or improve a specific communication metric, but to construct a compact surrogate that can be embedded in repeated circuit-level transient simulations and operated autonomously over extended waveform durations. This requires a model that remains stable under recursive generation, captures the effect of temperature on waveform evolution, and remains lightweight enough for practical use in design loops. Nevertheless, detailed physics-based representations that explicitly preserve broader electro-thermal dynamics can remain computationally burdensome in repeated simulation and optimization settings [14].

In this work, we address this problem through a data-driven surrogate model for thermally coupled VCSEL dynamics. Rather than repeatedly simulating the full internal electro-thermal state evolution at every time step, the proposed approach learns an input–output dynamical representation from waveform data generated by a high-accuracy reference simulator. The model is constructed by first identifying a compact empirical memory horizon, then learning a shared nonlinear latent representation of the waveform dynamics, and finally adapting this shared representation across operating conditions through lightweight temperature-specific affine readouts. In this way, the surrogate is designed to preserve the dominant dynamical structure required for realistic waveform generation while providing a more compact representation suitable for repeated transient simulation. The main contributions of this work are as follows:

- We formulate fabrication-stage VCSEL surrogate modeling as a closed-loop waveform-generation problem under thermal variation, shifting the focus from static device characterization or one-step prediction to stable recursive transient simulation over extended temporal windows.
- We develop a finite-memory latent surrogate architecture that combines an empirically identified observable state, a shared nonlinear dynamical representation, and lightweight operating-condition-specific affine readouts, enabling dominant waveform dynamics to be retained while keeping operating-condition adaptation computationally simple.
- We train and evaluate the surrogate in closed loop, showing with high-fidelity electro-thermal simulator data that it can generate stable optical waveforms over extended waveform durations, that a shared latent representation transfers across temperatures through lightweight per-temperature readout calibration, and that the learned representation also transfers across bias conditions with minimal additional adaptation.

The remainder of the paper is organized as follows. Section 2 describes the fabrication-stage modeling requirements and the reference high-accuracy VCSEL generator model. Section 3 presents the proposed surrogate modeling framework, including memory identification, latent-model construction, and training methodology. Section 4 reports simulation results and evaluates waveform-prediction accuracy, cross-temperature transfer under lightweight adaptation, and transfer across operating bias. Section 5 concludes the paper and outlines directions for future work.

2 Fabrication-Stage Modeling Requirements

2.1 Necessity of a Generator Model

During the fabrication-stage design of VCSEL driver integrated circuits, the laser must be represented inside the circuit simulator together with the driver, bias network, and surrounding front-end components. In this setting, the VCSEL is not merely a static optical source or equivalent load, but a nonlinear electro-optical device whose transient response depends on the applied current waveform, operating bias, and temperature. As a result, reliable driver verification requires a model that can reproduce the time-domain optical response of the device under realistic electrical excitation.

This requirement arises from the types of simulations routinely performed before tape-out. Circuit designers typically evaluate DC operating points, small-signal characteristics, large-signal transient behavior, and parametric variations across bias current, temperature, and process corners. Among these, transient simulations play a particularly important role because they determine whether the combined driver–VCSEL system can deliver the required waveform quality under practical signaling conditions. In particular, waveform-level simulations are used to assess optical swing, eye opening, pattern dependence,

and sensitivity to nonlinear and thermal effects. For this purpose, static current–voltage (IV) or current–power (IP) characteristics are insufficient, since they do not capture the coupled carrier–photon dynamics that govern the modulated response of the VCSEL.

A further challenge is that these transient simulations are not performed only once. They must be repeated throughout the design cycle as the driver architecture evolves, for example when modifying the output stage, bias circuitry, equalization network, or pre-emphasis structure. In short-reach optical links, such simulations are commonly driven by long pseudo-random bit sequences (PRBS) using signaling formats such as NRZ or PAM4 in order to expose intersymbol interference, pattern-dependent effects, and operating-margin limitations. Consequently, the model embedded in the simulation loop must act as a *generator model*, namely, a model that can recursively produce the optical output waveform sample by sample under a prescribed input-current sequence.

Long pseudo-random test patterns make this simulation burden especially severe. For example, at 50 Gb/s, a PRBS-7 sequence (127 bits) spans only about 2.5 ns, whereas PRBS-15 (32,767 bits) already spans about 0.65 µs. For PRBS-31, the full sequence length reaches 2.15×10^9 bits, corresponding to approximately 43 ms. This exponential growth in transient duration quickly makes exhaustive waveform simulation impractical in transistor-level circuit-design environments, especially when extracted parasitics and multiple operating corners must be included.

Physics-based VCSEL models [14, 33] can reproduce these behaviors with high fidelity because they incorporate the relevant carrier, photon, thermal, and parasitic dynamics. However, this accuracy comes at a considerable computational cost. In practical circuit-design environments, repeated transient simulations involving long data patterns, extracted parasitics, and multiple operating conditions can become prohibitively slow, and the resulting simulations may also suffer from numerical or convergence difficulties. The burden grows further as the sequence length increases, since realistic verification often requires sufficiently long patterns to reveal waveform distortions.

2.2 Conventional High-Accuracy Generator Model

The conventional high-accuracy generator model used as a reference in this work is a physics-based large-signal equivalent-circuit VCSEL model introduced in [14, 33]. The model is formulated as an electro-thermal and optical equivalent circuit specifically intended for circuit-level simulation of directly modulated datacom VCSELs. The model explicitly represents the underlying physical dynamics through a set of strongly interdependent subcircuits corresponding to the electrical input interface, carrier reservoirs, photon population, and internal temperature.

Internally, the model tracks carrier populations in both the separate confinement heterostructure (SCH) barriers and quantum wells, as well as the photon population in the optical cavity, using coupled dynamic relations derived from the device physics. These subsystems interact bidirectionally with a thermal block that captures self-heating effects and the resulting feedback on material and device parameters. Temperature influences the model through both the ambient boundary condition and temperature-dependent parameters such as gain-related terms, recombination coefficients, absorption, mirror losses, and electrical resistance, which are described in the original work using polynomial dependencies around room temperature [14, 33]. Accordingly, device characteristics such as threshold current and modulation bandwidth are not fixed constants in the reference model but vary with operating bias and temperature through the coupled electro-thermal model. For this reason, rather than reporting a single nominal set of such values, we refer the reader to the original generator-model descriptions in [14]. The resulting structure enables the model to reproduce large-signal modulation behavior across bias and temperature conditions while remaining compatible with circuit simulators.

In a typical transient waveform simulation, the internal operation of the model in the circuit simulator can be summarized as follows. First, a DC operating point is established for the selected bias current and ambient temperature, providing steady-state carrier, photon, and thermal conditions. A time-varying modulation current waveform (e.g., OOK or PAM signaling) is then applied on top of the bias current as the electrical excitation. During transient simulation, the coupled electro-thermal subcircuits evolve simultaneously, continuously updating the carrier populations in the SCH and quantum wells, the photon population in the cavity, and the internal temperature state. Self-heating dynamically modifies the internal temperature, which in turn updates temperature-dependent material and device parameters.

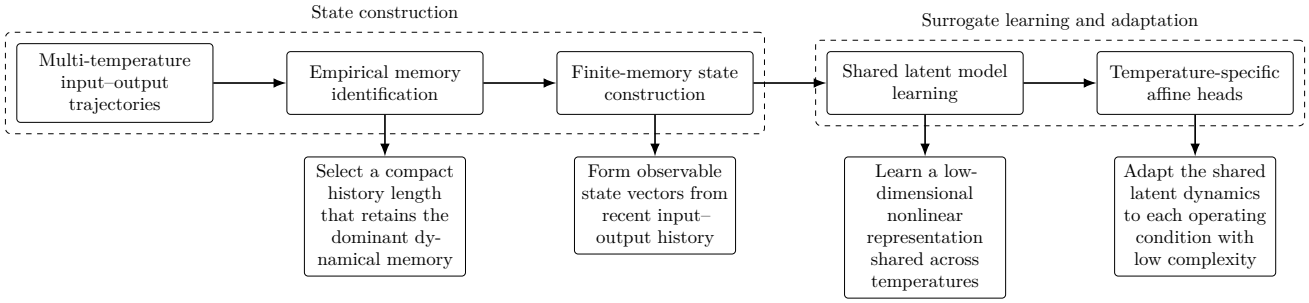

Figure 1: Overview of the surrogate-modeling pipeline. Input–output trajectories collected across temperatures are first used to identify an empirical memory horizon. This memory length defines the observable finite-memory state used by the surrogate. A shared nonlinear latent representation is then learned from these states, after which lightweight temperature-specific affine heads are fitted to adapt the common latent dynamics to each operating condition.

Finally, the optical output waveform is obtained from the photon escape rate through the cavity mirrors. In the model, the emitted optical power is calculated from this photon rate together with the photon energy, as described in [14, 33].

3 The Proposed Lightweight Surrogate Modeling Framework

The proposed surrogate construction consists of three main components. First, we determine how much past input–output history must be retained to predict the next optical-power sample accurately. This step provides an empirical memory order and defines the observable state used by the surrogate. Second, using this finite-memory representation, we learn a compact nonlinear latent model that is shared across temperatures. The shared model is trained in closed loop so that it captures the waveform evolution under recursive prediction rather than only one-step regression. Third, after the shared representation has been learned, we fit lightweight temperature-specific affine readout heads that adapt the common latent dynamics to each operating condition with minimal additional complexity. The following subsections describe these steps in order: memory identification, surrogate construction, training, and evaluation. Figure 1 summarizes the overall pipeline.

3.1 Empirical Identification of Dynamical Memory

A key modeling question is how much past waveform history must be retained in order to predict the next optical-power sample accurately. Retaining too little history leads to systematic modeling error as relevant information is discarded. In contrast, retaining overly long histories increases model complexity without improving predictive accuracy.

To determine a suitable history length, we estimate an *empirical memory horizon* directly from data using predictive error analysis [34, Sec. 2.9]. Consider a measured trajectory obtained at a fixed operating point characterized by bias current I_{mA}, symbol rate R, and temperature T. The resulting waveform consists of paired sequences

$$\{x_t\}_{t=1}^{N}, \qquad \{y_t\}_{t=1}^{N},$$

where x_t denotes the injected current sample and y_t denotes the measured optical power.

For a candidate memory length $L \in \mathbb{N}$, define the observable state

$$\mathbf{s}_t^{(L)} = \begin{bmatrix} y_t & \cdots & y_{t-L+1} & x_t & \cdots & x_{t-L+1} & x_{t+1} \end{bmatrix}^{\mathsf{T}} \in \mathbb{R}^{2L+1}, \tag{1}$$

which aggregates the recent output history, input history, and the next current sample.

Using this representation, the prediction problem becomes

$$y_{t+1}^{(L)} \approx f(\mathbf{s}_t^{(L)}). \tag{2}$$

To quantify the predictive value of different memory lengths, we fit linear predictors using ridge regularization and evaluate their test mean-squared error $\mathrm{MSE}(L)$ (defined as the average squared discrepancy between the predicted and true next sample on the test segment; see Appendix A). As L increases, the available information grows and the achievable prediction error typically decreases until it saturates once the retained history captures the dominant input–output memory of the device. We therefore select the smallest memory length whose predictive performance is near-optimal. Let

$$\mathrm{MSE}_{\min} = \min_{L' \in \mathcal{L}} \mathrm{MSE}(L'), \tag{3}$$

where $\mathcal{L}$ denotes a grid of candidate memory lengths. The empirical memory order is then defined as

$$L^\star = \min \Big\{ L \in \mathcal{L} : \mathrm{MSE}(L) \le (1+\eta)\mathrm{MSE}_{\min} \Big\}, \tag{4}$$

where η is a small tolerance. Implementation details of the predictive-error analysis are provided in Appendix A.

Remark 1. The rule in (4) selects the smallest history length that achieves prediction performance within a prescribed margin of the best observed value. This corresponds to identifying the effective observable memory of the dynamical system.

3.2 Finite-Memory Surrogate Model

The physical VCSEL simulator evolves a hidden electro–thermal state $\mathbf{z}_t$ according to

$$\mathbf{z}_{t+1} = \mathcal{F}(\mathbf{z}_t, x_{t+1}), \qquad y_{t+1} = \mathcal{G}(\mathbf{z}_{t+1}), \tag{5}$$

where $\mathbf{z}_t$ collects the internal physical variables of the device, such as carrier populations, cavity photon population, and thermal state. In the underlying high-fidelity model, the state-update map $\mathcal{F}(\cdot)$ represents the time-domain evolution induced by the coupled carrier–photon–thermal dynamics under the applied current sample x_{t+1}, while the readout map $\mathcal{G}(\cdot)$ converts the internal optical state to the observable output waveform, namely the emitted optical power. This abstraction is consistent with physics-based large-signal VCSEL models based on coupled rate equations with thermal feedback, as described in Section 2.1 and VCSEL generator models [14, 33].

In the surrogate model, however, these internal variables are not observed directly. Instead, we seek a reduced input–output description that reproduces the waveform evolution using only measurable quantities. The empirical memory analysis in Section 3.1 indicates that, for prediction purposes, the relevant dynamical information can be well approximated by a finite history of past inputs and outputs. We therefore replace the hidden state $\mathbf{z}_t$ by the observable finite-memory state $\mathbf{s}_t = \mathbf{s}_t^{(L^\star)}$ and model the one-step dynamics as

$$y_{t+1} \approx f(\mathbf{s}_t). \tag{6}$$

3.2.1 Latent representation

To obtain a compact approximation of the unknown mapping $f(\cdot)$, we introduce a nonlinear encoder

$$\mathbf{h}_t = \phi(\mathbf{s}_t; \theta) \in \mathbb{R}^d, \qquad d \ll 2L^\star + 1. \tag{7}$$

The predicted output is then given by a temperature-dependent affine head

$$\hat{y}_{t+1}^{(T)} = \mathbf{w}_T^\top \mathbf{h}_t + b_T. \tag{8}$$

Remark 2 (Modeling rationale). The encoder $\phi(\cdot;\theta)$ is intended to capture the nonlinear dynamical structure of the VCSEL that is common across operating conditions, while the affine head models operating-point–dependent effects such as changes in bias current and temperature. This decomposition is motivated by practical VCSEL design workflows. In circuit and device simulations, engineers often evaluate the same device across a range of bias points and temperatures. Ideally, the underlying dynamical generator $f(\cdot)$ should remain robust to these operating-point variations, while only a small

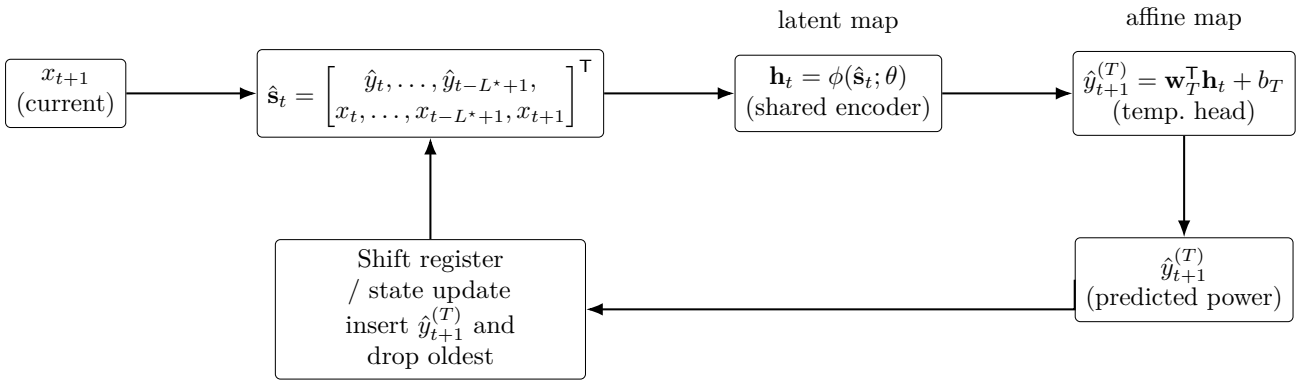

Figure 2: Closed-loop surrogate generator. The shared encoder $\phi(\cdot;\theta)$ maps the finite-memory state $\hat{\mathbf{s}}_t$ to a low-dimensional latent $\mathbf{h}_t$, and a temperature-specific affine head $(\mathbf{w}_T, b_T)$ produces $\hat{y}_{t+1}^{(T)}$. The prediction is fed back through a shift-register state update to form $\hat{\mathbf{s}}_{t+1}$ for the next step.

set of parameters needs to adapt to the specific condition. By learning a shared nonlinear representation through $\phi(\cdot;\theta)$ and restricting the operating-point dependence to a simple affine readout, the model allows the latent dynamical structure to be reused across operating conditions. Consequently, the surrogate does not need to relearn the entire input–output mapping for each temperature or bias setting, but adapts via a low-dimensional parameter change in the output layer.

We implement $\phi(\cdot;\theta)$ as a shallow feedforward MLP,

$$\phi(\mathbf{s};\theta) = \mathbf{W}_3\,\sigma\Big(\mathbf{W}_2\,\sigma(\mathbf{W}_1\mathbf{s} + \mathbf{b}_1) + \mathbf{b}_2\Big) + \mathbf{b}_3, \tag{9}$$

with pointwise nonlinearity $\sigma(\cdot)$ and latent dimension d. The feedforward choice matches the modeling premise that all relevant dynamical information is contained in the finite observable state $\mathbf{s}_t$; no additional recurrence is introduced beyond the shift-register structure of (1). Restricting depth and latent dimension enforces compactness and keeps per-step cost independent of the rollout horizon. Under standard universal approximation results, shallow MLPs approximate continuous mappings on compact domains [35], providing a finite-dimensional approximation of the unknown input–output operator.

3.2.2 Closed-loop generator

The surrogate is ultimately used as a waveform generator. Starting from an initial history window, it is iterated recursively so that previously generated outputs are fed back into the finite-memory state. To describe this closed-loop operation, let

$$\hat{\mathbf{s}}_t = \begin{bmatrix} \hat{y}_t & \cdots & \hat{y}_{t-L^\star+1} & x_t & \cdots & x_{t-L^\star+1} & x_{t+1} \end{bmatrix}^\mathsf{T} \tag{10}$$

denote the simulated state formed from the known input sequence and the previously generated outputs. The next optical-power sample is then produced as

$$\hat{y}_{t+1}^{(T)} = \mathbf{w}_T^\mathsf{T}\phi(\hat{\mathbf{s}}_t;\theta) + b_T. \tag{11}$$

Equations (10)–(11) define a deterministic finite-memory dynamical generator driven by the input sequence $\{x_t\}$. Once initialized, the model produces the waveform recursively, sample by sample, without access to future measured outputs. This deployment mode is central to the intended use of the surrogate for fast waveform simulation, and it motivates the rollout-based training strategy introduced in Section 3.3. The resulting closed-loop architecture is illustrated in Fig. 2.

Remark 3 (Temperature information). The temperature T in $(\mathbf{w}_T, b_T)$ is assumed known in this work. This is natural in the present setting, since the surrogate is intended for pre-fabrication waveform generation and circuit-level simulation, where the operating temperature is part of the simulation scenario and is specified a priori. Thus, the temperature-conditioned readout does not introduce an

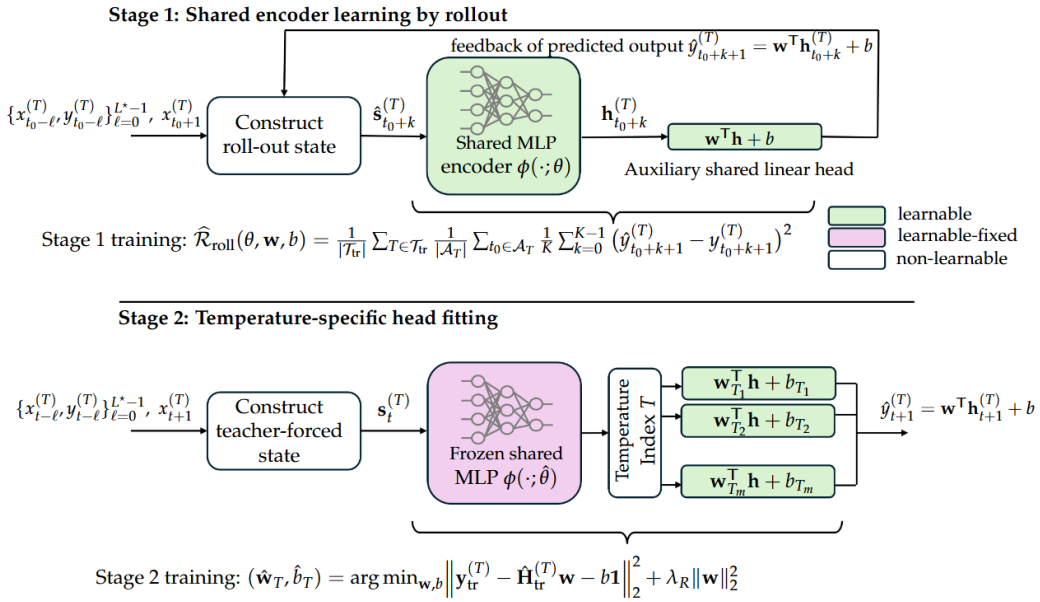

Figure 3: Two-stage training procedure of the proposed surrogate. In Stage 1, a measured history window is used only to initialize the rollout, after which the model evolves recursively using its own predictions. A shared MLP encoder is trained together with an auxiliary shared linear head by minimizing the rollout loss over the training temperatures. The shared head is used only to provide a scalar supervision signal and is discarded after Stage 1. In Stage 2, the learned encoder is frozen and teacher-forced latent features are used to fit a temperature-specific linear head by ridge regression.

additional sensing requirement within the design workflow. If desired, temperature could also be inferred from the initial waveform segment, since the emitted optical signal exhibits temperature-dependent changes in statistical complexity and features such as Fuzzy entropy [36] can be used to estimate the VCSEL operating temperature directly from payload-driven waveforms [37].

3.3 Training Procedure

The training procedure consists of two stages. In Stage 1, a shared nonlinear encoder is learned across the training temperatures so as to capture the latent dynamical structure common to all operating conditions. In Stage 2, the encoder is frozen and a lightweight temperature-specific affine readout head is fitted in the learned latent space. The first stage is trained using rollout loss, since the surrogate is ultimately deployed in recursive closed-loop mode, while the second stage reduces to a convex ridge-regression problem. An overview of the training procedure is given in Fig. 3.

3.3.1 Stage 1: Rollout-based shared encoder learning

The central challenge in training the surrogate is that its deployment mode is recursive. During waveform simulation, the predictor does not operate on states built from measured past outputs, but on states that are progressively formed using its own previous predictions. Consequently, the relevant training objective is not only one-step predictive accuracy under measured states, but multi-step accuracy under the model-induced closed-loop dynamics.

To define this training problem, we first extract supervised state–target pairs from the measured trajectories. For each temperature T, let

$$\mathcal{D}_T = \{(\mathbf{s}_t^{(T)}, y_{t+1}^{(T)})\}_{t \in \mathcal{J}_T} \tag{12}$$

denote the set of admissible state–target pairs constructed from the measured waveform history, where $\mathcal{J}_T$ is the corresponding index set. These pairs define the data-induced one-step regression problem. To learn a latent representation shared across temperatures, we attach a single affine output layer, common to all training temperatures, to the encoder and consider the auxiliary predictor

$$F_{\theta,\mathbf{w},b}(\mathbf{s}) \triangleq \mathbf{w}^\top \phi(\mathbf{s};\theta) + b, \tag{13}$$

where $(\mathbf{w}, b)$ is shared across temperatures. The role of this affine head is purely auxiliary: it provides a common scalar supervision signal that guides the encoder toward a latent representation that is informative for waveform evolution across temperatures.

Stage 1 optimizes the shared encoder under recursive closed-loop evolution rather than only one-step prediction. At this stage, the model consists of the encoder $\phi(\cdot;\theta)$ together with the auxiliary shared head $(\mathbf{w}, b)$ in (13). For a training temperature $T \in \mathcal{T}_{\mathrm{tr}}$ and an admissible anchor index t_0, the recursion is initialized using the measured history

$$\hat{y}^{(T)}_{t_0-\ell} = y^{(T)}_{t_0-\ell}, \qquad \ell = 0, \dots, L^\star - 1,$$

so that the initial state is anchored in the data. After initialization, the model is run in closed loop:

$$\hat{y}^{(T)}_{t_0+k+1} = \mathbf{w}^{\mathsf{T}} \phi\Big(\hat{\mathbf{s}}^{(T)}_{t_0+k}; \theta\Big) + b, \qquad k = 0, \dots, K-1, \tag{14}$$

where $\hat{\mathbf{s}}^{(T)}_{t_0+k}$ is obtained by replacing the output history in the finite-memory state with the previously generated samples. Thus, the predictor is repeatedly evaluated on its own induced states, exactly as in deployment.

The corresponding K-step rollout loss is

$$\mathcal{L}_{\mathrm{roll}}(\theta, \mathbf{w}, b; T, t_0) = \frac{1}{K} \sum_{k=0}^{K-1} \Big(\hat{y}^{(T)}_{t_0+k+1} - y^{(T)}_{t_0+k+1}\Big)^2. \tag{15}$$

Averaging over training temperatures and admissible anchors yields the Stage 1 objective

$$\widehat{\mathcal{R}}_{\mathrm{roll}}(\theta, \mathbf{w}, b) = \frac{1}{|\mathcal{T}_{\mathrm{tr}}|} \sum_{T \in \mathcal{T}_{\mathrm{tr}}} \frac{1}{|\mathcal{A}_T|} \sum_{t_0 \in \mathcal{A}_T} \mathcal{L}_{\mathrm{roll}}(\theta, \mathbf{w}, b; T, t_0), \tag{16}$$

where $\mathcal{A}_T$ denotes the set of anchor indices for temperature T whose history window and K-step rollout remain inside the training-time segment.

Minimization of (16) is performed by backpropagation through the unrolled recursion (14). The encoder is therefore trained not only according to one-step predictive accuracy, but according to how its latent representation behaves under autonomous closed-loop evolution. This is the key point of Stage 1: the encoder is explicitly optimized so that repeated one-step predictions remain accurate when its own outputs are fed back into future states.

Remark 4. The teacher-forced approach evaluates the predictor on the data-induced state distribution, whereas recursive waveform generation evaluates it on model-induced states generated by previous predictions. This distribution mismatch is commonly referred to as exposure bias [38]. Training with rollout loss mitigates this effect by optimizing the shared representation directly under the same recursive mechanism used at deployment.

After Stage 1, the auxiliary shared head $(\mathbf{w}, b)$ is discarded and only the learned encoder parameters $\hat{\theta}$ are retained.

3.3.2 Stage 2: Temperature-specific affine head fitting

With the encoder fixed, Stage 2 adapts the shared latent representation to each temperature through a lightweight affine readout. For each temperature T, we construct latent features from teacher-forced states,

$$\hat{\mathbf{h}}^{(T)}_t = \phi(\mathbf{s}^{(T)}_t; \hat{\theta}), \qquad t \in \mathcal{J}_T, \tag{17}$$

and define the final temperature-dependent predictor as

$$\hat{y}^{(T)}_{t+1} = \mathbf{w}_T^{\mathsf{T}} \hat{\mathbf{h}}^{(T)}_t + b_T. \tag{18}$$

Since the encoder is fixed, estimation of $(\mathbf{w}_T, b_T)$ reduces to a linear regression problem in latent space. Let

$$\hat{\mathbf{H}}^{(T)}_{\mathrm{tr}} = \Big[(\hat{\mathbf{h}}^{(T)}_t)^{\mathsf{T}}\Big]_{t \in \mathcal{J}_T} \in \mathbb{R}^{n_T \times d}, \qquad \mathbf{y}^{(T)}_{\mathrm{tr}} = \Big[y^{(T)}_{t+1}\Big]_{t \in \mathcal{J}_T} \in \mathbb{R}^{n_T}, \tag{19}$$

and augment the latent design matrix with a bias column,

$$\hat{\mathbf{H}}_{\mathrm{tr},b}^{(T)} = \left[\hat{\mathbf{H}}_{\mathrm{tr}}^{(T)} \quad \mathbf{1}\right]. \tag{20}$$

The temperature-specific head is then estimated by ridge regression,

$$(\hat{\mathbf{w}}_T, \hat{b}_T) = \arg\min_{\mathbf{w},b} \left\| \mathbf{y}_{\mathrm{tr}}^{(T)} - \hat{\mathbf{H}}_{\mathrm{tr}}^{(T)}\mathbf{w} - b\mathbf{1} \right\|_2^2 + \lambda_R \|\mathbf{w}\|_2^2, \tag{21}$$

which admits the closed-form solution

$$\hat{\boldsymbol{\beta}}_T = \left((\hat{\mathbf{H}}_{\mathrm{tr},b}^{(T)})^{\mathsf{T}} \hat{\mathbf{H}}_{\mathrm{tr},b}^{(T)} + \lambda_R \mathbf{P} \right)^{-1} (\hat{\mathbf{H}}_{\mathrm{tr},b}^{(T)})^{\mathsf{T}} \mathbf{y}_{\mathrm{tr}}^{(T)}, \tag{22}$$

where

$$\hat{\boldsymbol{\beta}}_T = \begin{bmatrix} \hat{\mathbf{w}}_T \\ \hat{b}_T \end{bmatrix}, \qquad \mathbf{P} = \mathrm{diag}(\mathbf{I}_d, 0), \tag{23}$$

so that the regression coefficients are penalized but the intercept is not.

In summary, Stage 1 learns a shared latent dynamical representation through rollout training, while Stage 2 adapts that representation to each operating temperature through a lightweight closed-form affine fit. This separation is computationally attractive because the nonlinear optimization is carried only by the encoder, whereas the temperature-specific adaptation reduces to a small convex problem in latent space.

3.4 Evaluation Protocol

Evaluation is designed to assess how well a shared latent dynamical representation transfers across temperatures while preserving causal time ordering within each trajectory. To this end, the shared encoder is first trained once on a source set of temperatures and then frozen. For evaluation, a temperature-specific affine head is fitted separately for each temperature sweep using only its training-time prefix, and performance is measured on the corresponding strict test suffix.[1]

Within each trajectory, we split the samples contiguously in time as

$$\mathcal{I}_{\mathrm{tr}}^{(T)} = \{1, \ldots, \lfloor \rho N_T \rfloor\}, \qquad \mathcal{I}_{\mathrm{te}}^{(T)} = \{\lfloor \rho N_T \rfloor + 1, \ldots, N_T\}, \tag{24}$$

where $\rho \in (0,1)$ denotes the training fraction and we use $\rho = 0.7$. All state standardization is performed using training statistics only and then applied unchanged to non-training states.

Under this protocol, Stage 1 provides a fixed shared encoder. Stage 2 then fits a temperature-specific affine head for each sweep using only the prefix indexed by $\mathcal{I}_{\mathrm{tr}}^{(T)}$. All reported metrics are computed on the strict test suffix indexed by $\mathcal{I}_{\mathrm{te}}^{(T)}$.

For one-step evaluation, let

$$\hat{y}_{t+1}^{\mathrm{TF},(T)} = \hat{\mathbf{w}}_T^{\mathsf{T}} \phi(\mathbf{s}_t^{(T)}; \hat{\theta}) + \hat{b}_T, \qquad t+1 \in \mathcal{I}_{\mathrm{te}}^{(T)}, \tag{25}$$

denote the teacher-forced predictor evaluated on measured states. The corresponding one-step mean-squared error is

$$\mathrm{MSE}_{\mathrm{TF}}^{(T)} = \frac{1}{|\mathcal{J}_{\mathrm{te}}^{(T)}|} \sum_{t \in \mathcal{J}_{\mathrm{te}}^{(T)}} \left(\hat{y}_{t+1}^{\mathrm{TF},(T)} - y_{t+1}^{(T)} \right)^2, \tag{26}$$

where

$$\mathcal{J}_{\mathrm{te}}^{(T)} = \{t : \ t+1 \in \mathcal{I}_{\mathrm{te}}^{(T)}, \ t \geq L^\star\}.$$

For free-running evaluation, the model is initialized at the train–test boundary and then iterated recursively over the test suffix. For a rollout horizon K, the corresponding rollout mean-squared error is

$$\mathrm{MSE}_{\mathrm{roll}}^{(T)}(K) = \frac{1}{K} \sum_{k=0}^{K-1} \left(\hat{y}_{t_0+k+1}^{(T)} - y_{t_0+k+1}^{(T)} \right)^2, \tag{27}$$

[1] Accordingly, the temperature results should be interpreted as evaluating transfer of a shared latent representation under lightweight per-temperature adaptation, rather than fully zero-shot temperature generalization.

where $t_0 = \lfloor \rho N_T \rfloor$ denotes the last sample in the training prefix. In the reported results we use the corresponding root-mean-squared errors,

$$\mathrm{RMSE}_{\mathrm{TF}}^{(T)} = \sqrt{\mathrm{MSE}_{\mathrm{TF}}^{(T)}}, \qquad \mathrm{RMSE}_{\mathrm{roll}}^{(T)}(K) = \sqrt{\mathrm{MSE}_{\mathrm{roll}}^{(T)}(K)}.$$

Performance is reported using both one-step teacher-forced prediction error and free-running rollout error. The former measures conditional next-sample accuracy under measured states, whereas the latter measures the autonomous behavior of the closed-loop generator under recursive prediction. Since the surrogate is intended for waveform simulation, rollout performance is the primary evaluation metric.

4 Simulation Results

4.1 Dataset and Implementation Details

The simulations use aligned input–output waveform pairs generated by the high-fidelity electro-thermal VCSEL simulator described in Section 2.2. Each waveform sweep corresponds to a fixed operating condition defined by the bias current, symbol rate, and ambient temperature. The full simulator-generated dataset covered a broader range of operating conditions than those retained in the present evaluations. However, because the underlying physics-based generator model can become numerically unstable for some operating-point combinations, particularly at elevated temperatures and higher symbol rates, we restricted attention to the subset of sweeps for which the reference simulator produced stable and physically reliable waveform trajectories. Operating points for which the simulator exhibited instability or unreliable convergence were discarded, since such trajectories do not provide a suitable ground truth for surrogate training and evaluation. After alignment and cropping, each retained sweep provides a paired trajectory of injected current and optical power samples of length 48,000.

In this work we focus on the 20 Gbaud operating point under PAM4 modulation and consider two bias currents, $I \in \{6, 10\}$ mA. This choice provides an initial large-signal validation setting for assessing closed-loop waveform generation while isolating the temperature- and bias-transfer behavior of the proposed surrogate. For each bias current, multiple temperature sweeps are available over a wide operating range. The retained temperature sweeps used in the reported 20 Gbaud experiments span the set

$$\{-40, -35, -30, -25, -20, -15, -10, -5, 0, 5, 10, 15, 20, 25, 30, 35, 40, 50, 80\}\ ^\circ\mathrm{C}, \tag{28}$$

with nonuniform spacing, and each temperature contributes one aligned trajectory under otherwise identical operating conditions. For the held-out temperature evaluation, we use $\{-40, 5, 80\}^\circ$C as held-out temperatures and use the remaining sweeps for training.

The surrogate uses the empirically selected memory order $L^\star$, which yields an input dimension of $2L^\star + 1$. The shared encoder $\phi(\cdot;\theta)$ is implemented as a feedforward MLP with two hidden layers of widths 128 and 64, followed by a latent layer of dimension $d = 32$. ReLU activations are used after each affine layer in the encoder.

Stage 1 rollout training is performed at the source operating bias $I = 6$ mA (the bias-transfer experiment later uses the $I = 10$ mA domain as target). The rollout horizon during training is set to $K_{\mathrm{train}} = 50$, while rollout evaluation uses $K_{\mathrm{eval}} = 1200$. The encoder is trained for 20 epochs with 200 mini-batch iterations per epoch and mini-batch size 128, using the Adam optimizer with learning rate 10^{-3}, momentum parameters $(\beta_1, \beta_2) = (0.9, 0.999)$, and numerical constant $\epsilon_{\mathrm{Adam}} = 10^{-8}$. To stabilize training, gradients are clipped by global norm with threshold 5.0.

During Stage 1, the encoder is coupled to an auxiliary shared scalar output layer used only for rollout training. To prevent divergence under recursive prediction, we optionally bound the output using a saturating nonlinearity. The predicted output is bounded through

$$\hat{y} \leftarrow y_{\max} \tanh(\hat{y}/y_{\max}), \tag{29}$$

with $y_{\max} = 3.0$, to suppress unstable excursions during recursive generation. After Stage 1, the auxiliary output layer is discarded and only the learned encoder is retained.

Stage 2 fits a temperature-specific affine head in the learned latent space using ridge regression with regularization parameter $\lambda_R = 10^{-2}$. Unless otherwise stated, all simulations use a contiguous time split with training fraction $\rho = 0.7$.

4.2 Numerical Method and Preprocessing

The numerical workflow starts from waveform traces produced by the high-fidelity electro-thermal reference simulator. For each operating point, the simulator provides two time-domain quantities: the injected anode current and the emitted optical power. These quantities are exported separately, and each exported file may contain several temperature sweeps. Therefore, before any surrogate modeling is performed, the raw simulator outputs must be converted into consistent paired trajectories.

The current and optical-power traces are paired through the operating-point metadata. Each retained trajectory is therefore associated with a unique operating condition (I, R, T), where I denotes the bias current, R the nominal baud rate, and T the ambient temperature.

A key numerical requirement is that the electrical excitation and optical response be represented on a common time basis. After pairing, the optical-power trace is therefore mapped onto the current-waveform time grid, yielding synchronized input–output samples (x_t, y_t). This is important because the finite-memory state in (1) assumes that the input and output histories refer to the same discrete-time trajectory.

In addition to this common-grid representation, we correct residual sample-level timing offsets between the exported electrical and optical traces. Such offsets can occur because the two quantities are obtained from separate simulator outputs. If left uncorrected, the delay would be absorbed into the learned memory and could bias both the empirical memory selection and the closed-loop rollout evaluation. We therefore align the trajectories using their transition structure, choosing the lag that maximizes the correlation between the first differences of the current and optical-power waveforms. This focuses the synchronization on waveform transitions rather than on absolute optical-power levels, which vary with temperature and bias. When the estimated lag is consistent across an operating-point subset, a common lag is used for all corresponding temperature sweeps to avoid introducing artificial temperature-dependent timing variations.

4.3 Empirical Memory Analysis

We next examine the effective dynamical memory of the VCSEL response. For a fixed operating point (20 Gbaud and $I = 6$ mA), the history length L is swept over the range $L \in \{2, 4, \ldots, 128\}$. For each candidate L, a ridge-regularized one-step predictor is trained using the finite-memory state defined in (1). To ensure that the selected memory order generalizes across operating conditions, prediction performance is evaluated by holding out one temperature sweep at a time during the memory-selection step.

Figure 4 reports the resulting Leave-one-temperature-out (LOTO) test error as a function of the history length. The blue curve shows the mean prediction error across held-out temperature sweeps, while the red curve shows the worst-case error. As expected, increasing the history length initially improves prediction accuracy by incorporating additional dynamical information. However, the improvement becomes progressively smaller as L grows, indicating that the dominant system memory has already been captured.

To make this behavior more explicit, Fig. 5 shows the relative excess prediction error with respect to the minimum achievable mean error. The curve decreases rapidly for small values of L and then enters a broad diminishing-returns region, indicating that much of the predictive information is already captured by moderate history lengths. Beyond this region, further increases in L yield progressively smaller gains, but small improvements remain visible in the cross-temperature criterion.

Following the selection rule in (4), we choose the smallest history length whose mean LOTO error lies within a tolerance of $\eta = 1\%$ of the minimum observed value. Under this criterion, the selected memory order is $L^\star = 98$. We retain this value as a conservative operating point, since the final surrogate is used in recursive rollout mode, where underestimating the observable history can be more harmful than a moderate increase in input dimension.

Taken together, these results indicate that the dominant waveform memory is captured with a finite and practically manageable observable history, while the selected value $L^\star = 98$ favors predictive robustness over the most aggressive possible compression in subsequent closed-loop waveform generation.

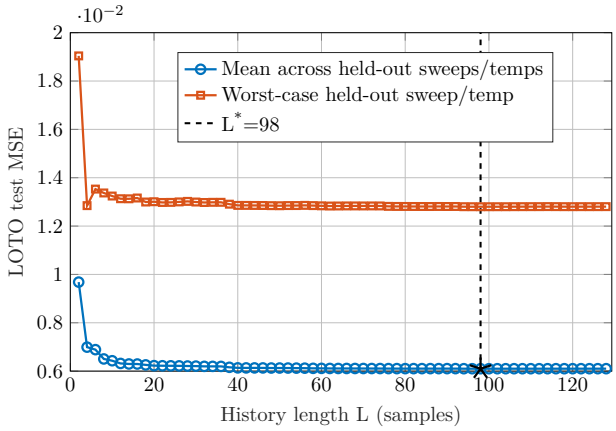

Figure 4: Leave-one-temperature-out (LOTO) prediction error as a function of the history length L. The blue curve shows the mean prediction error across held-out temperature sweeps, while the red curve shows the corresponding worst-case error. The dashed vertical line indicates the selected memory order $L^\star = 98$, obtained as the smallest history length whose mean LOTO test error lies within the tolerance rule in (4).

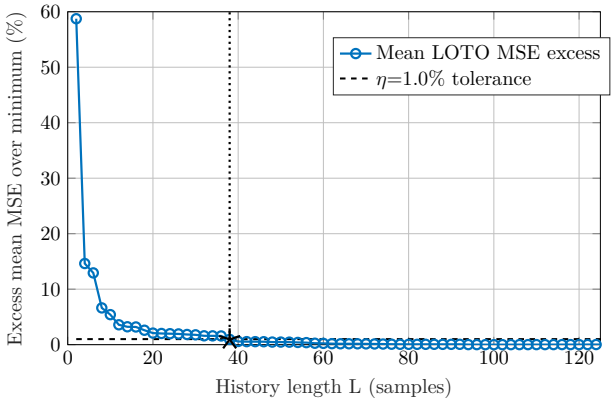

Figure 5: Relative excess prediction error with respect to the minimum achievable mean error. The dashed horizontal line indicates a tolerance level of $\eta = 1\%$. The curve rapidly enters a diminishing-returns region, indicating that much of the predictive information is captured by moderate history lengths, while further increases in L provide only progressively smaller gains.

4.4 Learning Behavior of the Shared Encoder

We first examine the optimization behavior of the shared encoder under the unrolled training objective. The encoder is trained on the training-temperature sweeps, while performance is monitored on the held-out temperatures $\{-40, 5, 80\}$°C using strict time holdout. At each iteration, optimization is performed over a finite rollout horizon of K_{train} steps, so that the latent representation is shaped directly by closed-loop prediction error rather than only one-step teacher-forced loss.

Fig. 6 shows the training loss as a function of iteration. The loss decreases rapidly at early iterations and then gradually stabilizes, indicating that the dominant input–output structure is learned early and subsequently refined through continued unrolled optimization. No instability or divergence is observed, suggesting that the finite-memory state and the chosen network size lead to a well-behaved training problem.

Fig. 7 reports the rollout RMSE on held-out temperatures versus epoch. Both the mean and worst-case held-out errors decrease during training, although moderate fluctuations remain, particularly in the worst-case curve. This is expected because the evaluation is carried out on unseen temperatures and in free-running rollout mode, which is more sensitive to error accumulation than one-step prediction.

Overall, these results show that the shared encoder can be trained stably under the unrolled objective and that the learned latent representation improves closed-loop prediction even on operating conditions not used during training. At the same time, the persistent gap between the mean and worst-case held-out errors indicates that temperature-dependent variation remains significant, motivating the temperature-specific output heads introduced next.

4.5 Dimensionality and Temperature Sensitivity of the Latent Representation

To further examine the structure of the learned representation, we compare the statistical variability of the raw finite-memory states with that of the latent representations produced by the frozen encoder. In particular, we quantify two aspects: (i) the effective dimensionality of each representation and (ii) how strongly this dimensionality varies across temperature.

Using the state sequence $\{\mathbf{s}_t\}$ in (1) and the corresponding latent sequence $\{\mathbf{h}_t\}$ in (7), we estimate the covariance matrix of each representation across time for every temperature sweep. Let $\{\lambda_i\}$ denote the eigenvalues of the corresponding covariance matrix. We quantify effective dimensionality through

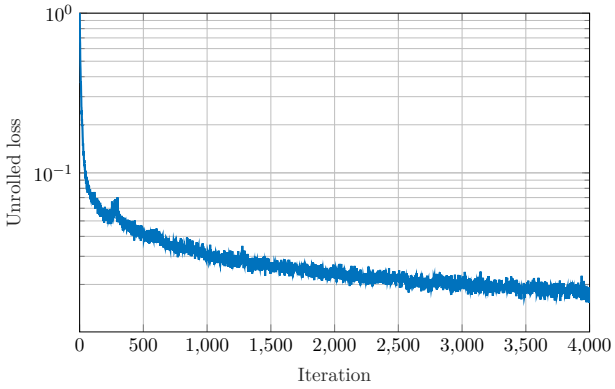

Figure 6: Training loss of the unrolled shared encoder as a function of iteration. The loss decreases rapidly at early iterations and then gradually stabilizes, indicating stable optimization of the latent dynamical model.

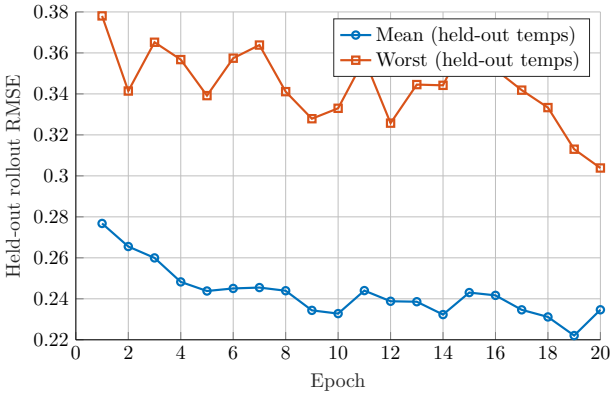

Figure 7: Held-out rollout RMSE versus epoch under strict time holdout. The mean and worst-case errors across the held-out temperatures both decrease during training, showing that the learned shared encoder improves closed-loop prediction on unseen temperatures.

Table 1: Effective dimensionality of the raw-state and latent representations across temperature. Deviations are computed relative to the temperature-averaged effective dimension of each representation.

T (°C)	−40	−30	−20	−10	0	10	20	30	40	50	80	Mean	Std	Range
$d_{\text{eff}}^{\text{raw}}$	8.240	8.169	7.996	7.939	7.899	7.961	7.908	7.894	7.929	7.981	8.234	7.996	0.117	0.346
$d_{\text{eff}}^{\text{lat}}$	6.904	6.937	6.955	6.833	6.745	6.830	6.814	6.938	6.908	6.869	6.900	6.877	0.074	0.219
Dev. raw	0.244	0.174	0.000	−0.056	−0.097	−0.035	−0.088	−0.101	−0.067	−0.015	0.238	–	–	–
Dev. lat.	0.027	0.060	0.078	−0.044	−0.132	−0.047	−0.063	0.062	0.031	−0.008	0.023	–	–	–

the participation ratio

$$d_{\text{eff}} = \frac{(\sum_i \lambda_i)^2}{\sum_i \lambda_i^2}, \tag{30}$$

which measures how many principal directions effectively contribute to the variance of the representation [39, Eq. (4)]. When the variance is concentrated in fewer dominant directions, d_{eff} becomes smaller, indicating a more compact representation.

Table 1 reports the effective dimensionality of the raw-state and latent representations across operating temperatures. The latent representation consistently exhibits a lower effective dimensionality than the raw state. Averaged across temperature, the effective dimension decreases from 7.996 for the raw state to 6.877 in the latent space, indicating that the encoder concentrates the waveform variability into fewer dominant directions.

The deviation columns in Table 1 quantify how each temperature-specific value differs from the temperature-averaged effective dimension of the corresponding representation,

$$\text{Dev}(T) = d_{\text{eff}}(T) - \frac{1}{|\mathcal{T}|} \sum_{T' \in \mathcal{T}} d_{\text{eff}}(T').$$

These values therefore characterize how strongly the effective dimensionality varies across operating conditions. The latent representation shows smaller deviations across temperature than the raw state. This trend is also reflected in the summary statistics at the bottom of the table: the standard deviation across temperature decreases from 0.117 for the raw state to 0.074 in the latent space, while the overall range decreases from 0.346 to 0.219.

Overall, these results suggest that the encoder produces a representation that is both more compact and less sensitive to temperature variations than the raw finite-memory state. This supports the use of the latent space as a stable basis for learning lightweight temperature-specific prediction heads.

4.6 Temperature-Adapted One-Step Prediction

After training the shared encoder on the multi-temperature dataset, we evaluate how well the learned latent representation supports one-step prediction across temperatures under lightweight per-temperature adaptation. Specifically, the encoder $\phi(\cdot)$ is frozen and used to map the observable state to a latent representation $\mathbf{h}_t$ as in (7), after which a temperature-specific linear head is fitted according to (25) and evaluated through $\mathrm{MSE}_{\mathrm{TF}}$. In this way, the nonlinear dynamical representation remains shared across temperatures, while only the final affine readout is adapted to each operating condition.

To contextualize the performance of the proposed model, we compare it with two reference predictors commonly used for nonlinear time-series modeling: regularized autoregression with exogenous input (*Ridge-ARX*) [40, Sec. 2.2] and *Boosted Trees*, implemented as a gradient-boosted regression-tree predictor [41].

Fig. 8 shows representative one-step prediction overlays at the held-out temperatures (-40°C, $+5$°C, and $+80$°C). The proposed model tracks the measured optical waveform well across all operating conditions, reproducing the dominant amplitude variations and temporal structure. The Ridge-ARX baseline remains slightly more accurate around some sharp transitions, while the boosted-tree predictor shows somewhat stronger smoothing in rapidly varying regions. Overall, the overlays indicate that the learned latent representation preserves the dominant waveform dynamics across a wide thermal range.

Fig. 9 (a) reports the one-step prediction RMSE across temperature for all methods, while Fig. 9 (b) shows the empirical distribution of the absolute one-step prediction errors aggregated over all temperatures. The proposed latent dynamical model yields higher one-step prediction error than the Ridge-ARX and boosted-tree baselines. This is consistent with the design of the method: the proposed model is trained to support stable recursive waveform generation rather than to optimize one-step accuracy alone. By contrast, the baseline methods are optimized directly for single-step regression in the observation space, which naturally favors lower one-step prediction error but does not enforce dynamically consistent waveform evolution over extended temporal windows.

4.7 Free-Running Rollout Prediction

We next evaluate the models in the deployment regime of interest, namely free-running multi-step prediction under temperature-specific readout adaptation. In contrast to the teacher-forced one-step setting, the predictor is initialized only once at the train–test boundary and is then recursively driven by its own past predictions. This evaluation therefore tests whether the learned shared representation supports a stable dynamical law under recursive rollout, rather than merely a local next-sample regression rule.

As in the previous subsection, each temperature sweep is split contiguously in time: the first 70% of the samples are used for training and the remaining 30% for testing. For the rollout-trained latent model, the frozen shared encoder is combined with a temperature-specific ridge head learned from the training prefix of the corresponding sweep. Rollout is then performed on the test suffix according to the closed-loop recursion in (11). To isolate the effect of the training objective, we also consider a teacher-forced latent model with the same architecture, but whose shared encoder is trained only using one-step teacher-forced loss. This ablation allows us to distinguish the effect of rollout-based training from that of the latent architecture itself. To ensure a fair comparison, all remaining baselines are trained on the same training prefix and evaluated in the same free-running manner.

Besides the rollout-trained latent model and its teacher-forced counterpart, we compare against three baselines. The first is the *Ridge-ARX* model, which provides a regularized linear autoregressive predictor with exogenous input in the original state space [40, Sec. 2.2]. The second is *Boosted Trees*, a nonlinear ensemble baseline based on gradient-boosted regression trees [41]. The third is a second-order *Volterra* model, representing a nonlinear memory-polynomial approximation of the system dynamics. Volterra series models are widely used for nonlinear system identification and have been applied to describe the modulation dynamics of semiconductor lasers, where nonlinear memory effects arise from carrier–photon interactions and thermal dynamics [42].

Fig. 11 shows representative free-running waveform overlays at three temperatures. In contrast to the one-step setting, the differences between methods become much more pronounced in rollout mode. The rollout-trained latent model (proposed) remains closely aligned with the measured waveform across

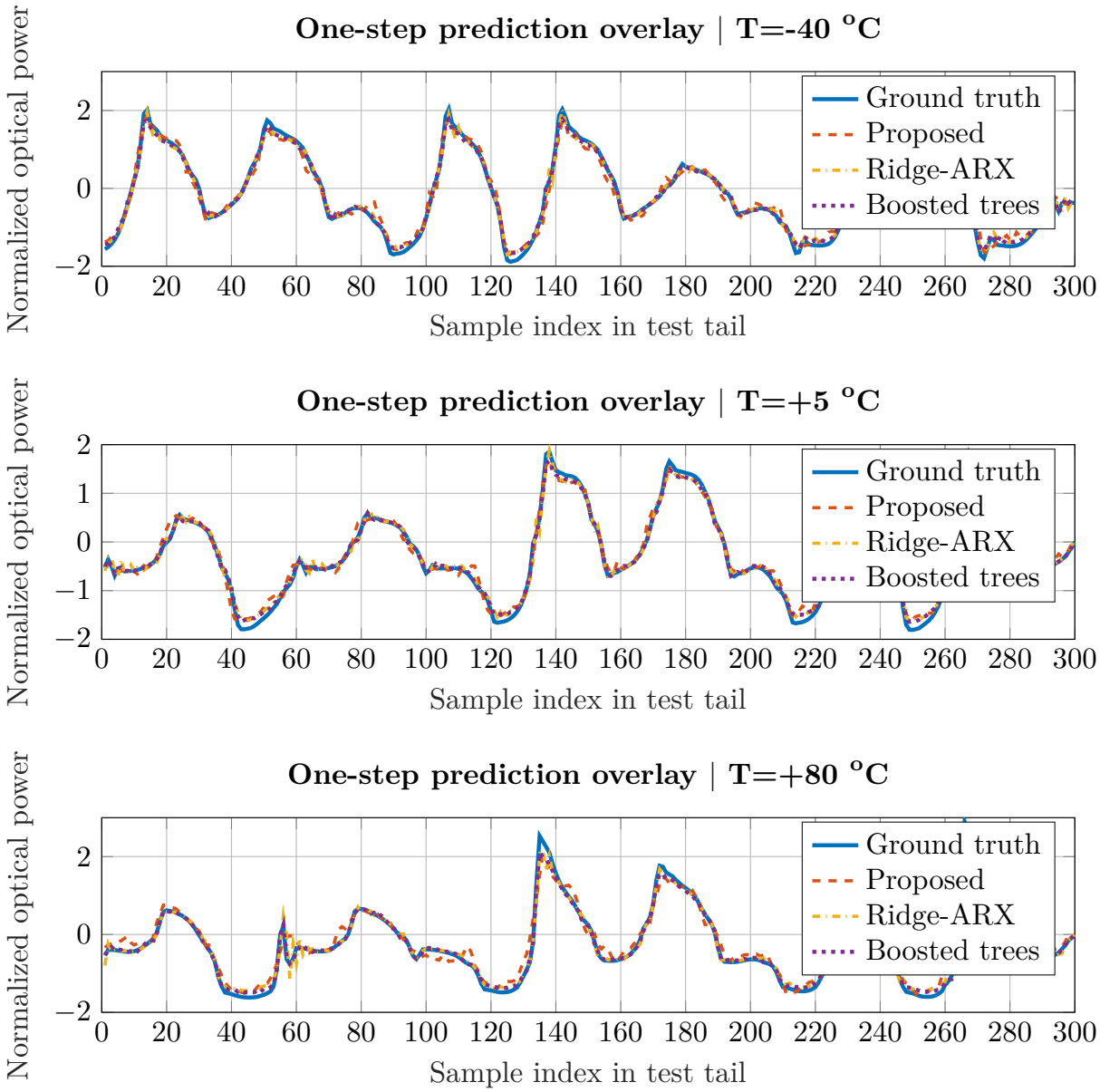

Figure 8: One-step prediction overlays at representative temperatures. The plots compare the ground-truth optical power trajectory with the predictions produced by the proposed latent dynamical model, the classical Ridge-ARX baseline, and the boosted tree baseline. The three panels correspond to low (-40°C), nominal ($+5$°C), and high ($+80$°C) operating temperatures. Across the full temperature range, the proposed model closely follows the measured dynamics while maintaining stable prediction behavior.

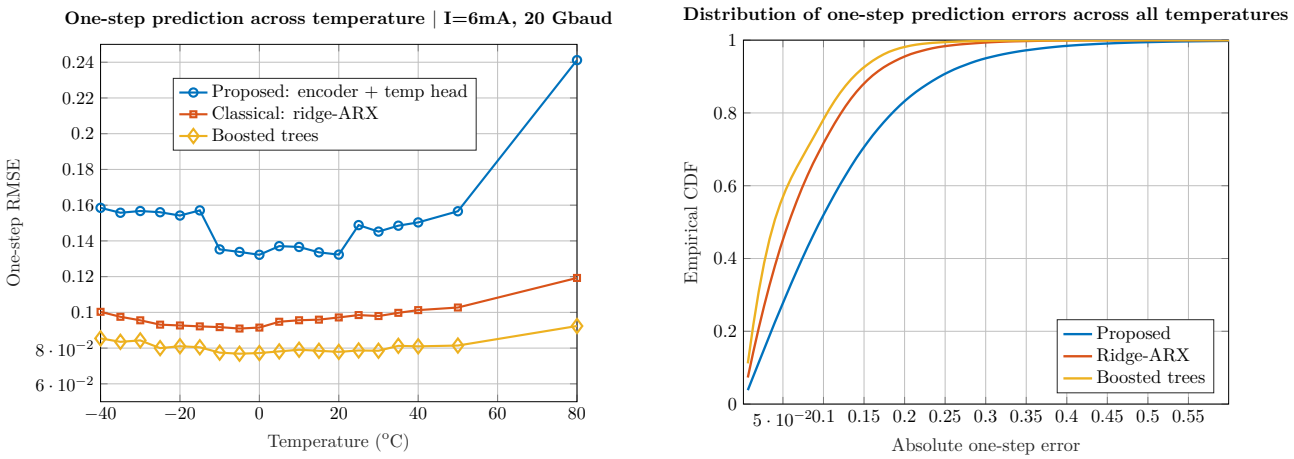

Figure 9: One-step prediction performance comparison. (a) **Left**: Root-mean-square error (RMSE) of one-step predictions across operating temperature for the proposed latent model and two baselines. (b) **Right**: Empirical cumulative distribution function (CDF) of the absolute one-step prediction errors aggregated over all temperatures. While the proposed method achieves higher one-step prediction error than the baselines, this behavior is expected because the model is trained to learn the underlying dynamical evolution of the system rather than to optimize purely one-step accuracy.

Table 2: PAM4 eye-opening (EO) comparison between the high-fidelity reference waveform and the proposed latent surrogate under free-running rollout. The minimum and mean EO values are computed over the three PAM4 eyes at the sampling instant and are reported in mW.

T	Ref. min EO (mW)	Sur. min EO (mW)	Ref. mean EO (mW)	Sur. mean EO (mW)
$-35°C$	0.40	0.30	0.50	0.50
$50°C$	0.60	0.60	0.60	0.70

the full prediction window and preserves both the oscillatory structure and the large-signal transitions. The Ridge-ARX predictor accumulates visible phase and amplitude drift as the horizon increases, while the boosted-tree model becomes increasingly unstable and departs from the underlying waveform. The Volterra baseline performs poorest over extended rollout windows, showing strong distortion and failing to reproduce the correct waveform evolution. These observations indicate that the latent model can capture a dynamical representation that remains stable under recursive prediction.

This trend is confirmed quantitatively in Fig. 12. Fig. 12 (a) reports the rollout RMSE across temperature for the rollout-trained latent model (proposed), the teacher-forced latent model, and the baseline methods. The rollout-trained latent model consistently achieves the lowest error over the full temperature range, whereas the teacher-forced latent model (with an identical architecture to the proposed) and the observation-space baselines degrade substantially in free-running mode. Fig. 12 (b) further supports this result through the empirical cumulative distribution of the absolute rollout errors aggregated over all temperatures. The rollout-trained latent model reaches high cumulative probability at much smaller error values, showing that its advantage persists throughout the rollout trajectories rather than being confined to a small subset of samples.

To complement the RMSE-based rollout evaluation with a waveform-level signal-integrity view, Fig. 10 shows eye diagrams obtained from the high-fidelity reference waveforms and from the proposed latent surrogate under free-running rollout. Two representative training temperatures are shown. The surrogate-generated eyes preserve the main PAM4 eye structure and relative level separation of the high-fidelity waveforms, although small differences in transition spread and level dispersion remain visible. Table 2 further reports the minimum and mean eye openings computed over the three PAM4 eyes at the sampling instant. The surrogate yields comparable eye-opening values to the high-fidelity reference, with particularly close agreement in the mean eye opening. These results indicate that the low rollout error is also reflected in the eye structure of the generated waveform.

Taken together, these results clarify the role of the proposed framework. The main gain in free-

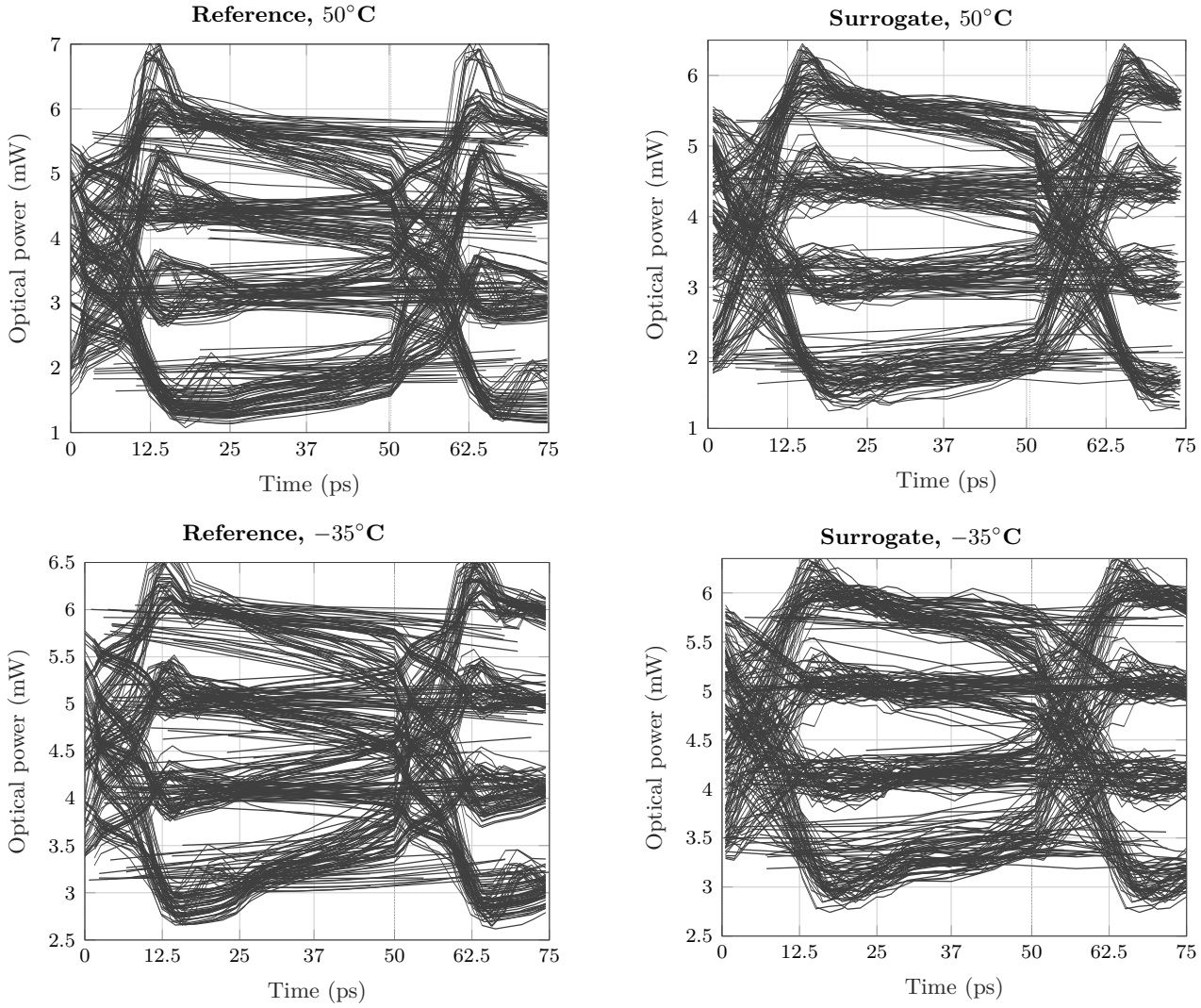

Figure 10: Eye diagrams obtained from the high-fidelity reference waveforms and the proposed latent surrogate under free-running rollout. Top row: 50°C. Bottom row: −35°C. Left column: high-fidelity electro-thermal reference simulation. Right column: latent surrogate with temperature-specific affine readout. The surrogate preserves the main PAM4 eye structure and level separation across the representative operating temperatures.

running prediction does not arise solely from the latent architecture, since the teacher-forced latent model uses the same structure but performs markedly worse under recursive rollout. Rather, the results indicate that exposing the model to its own predictions during training is important for learning a compact dynamical state whose recursive evolution remains stable over extended waveform durations. This also explains why the rollout-trained latent model is less competitive in the teacher-forced one-step test, yet clearly superior in the free-running regime relevant to surrogate waveform generation.

4.8 Transfer Across Operating Bias and Baud Rate

Finally, we evaluate whether the learned latent representation transfers across operating points. While the previous evaluation results considered temperature variation at a fixed bias current, practical deployment also requires the model to remain reliable under changes in operating bias. To examine this property, we train the shared encoder on waveform sweeps collected at 6 mA and evaluate it on a previously unseen bias of 10 mA.

The encoder learned on the source bias is kept fixed. For each temperature sweep in the target domain, a lightweight linear head is re-estimated using the training prefix of that sweep, while the encoder parameters remain unchanged. This setting tests whether the learned latent representation

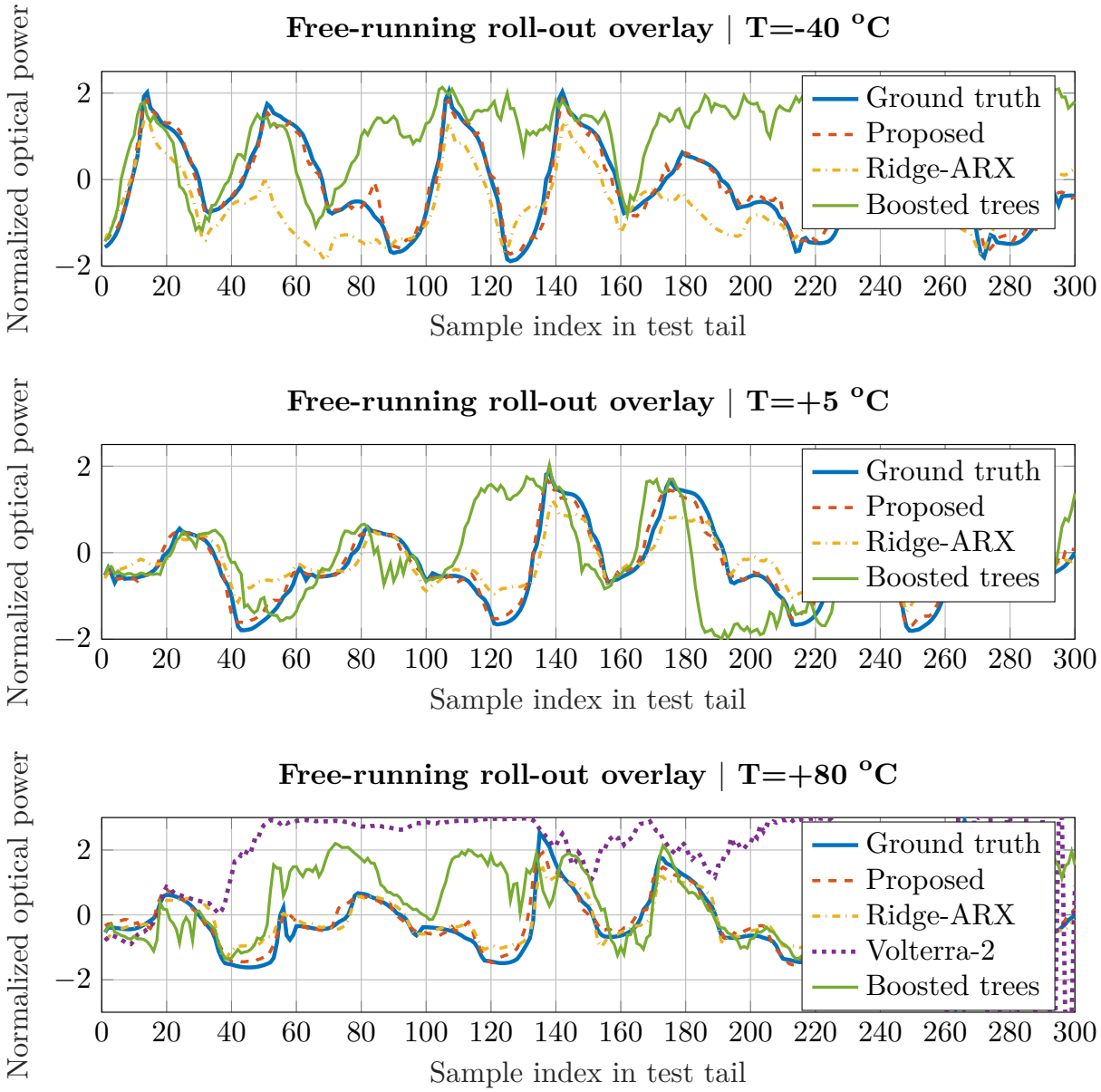

Figure 11: Free-running rollout overlays at representative temperatures. The plots compare the ground-truth optical power trajectory with the recursive predictions produced by the proposed latent dynamical model, the Ridge-ARX baseline, the Volterra-2 baseline, and the boosted-tree baseline. The three panels correspond to low (−40°C), nominal (+5°C), and high (+80°C) operating temperatures. Unlike the one-step setting, clear differences emerge under recursive prediction: the proposed model remains closely aligned with the measured waveform, whereas the baselines accumulate phase, amplitude, or structural distortion as the prediction horizon increases.

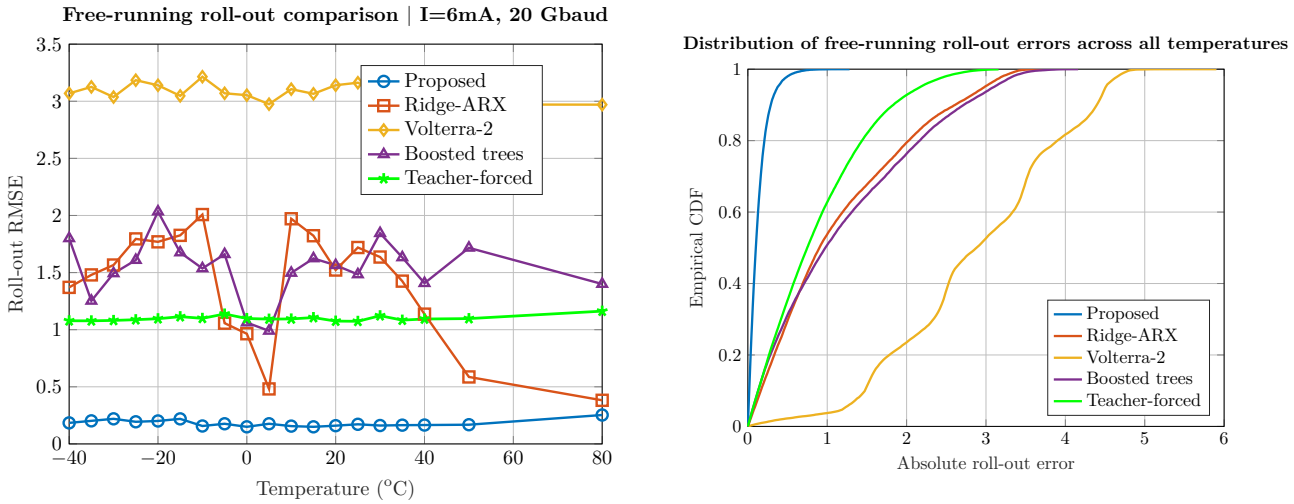

Figure 12: Free-running rollout performance comparison. (a) **Left**: Root-mean-square error (RMSE) of recursive rollout predictions across operating temperature for the proposed latent model and the baseline methods. (b) **Right**: Empirical cumulative distribution function (CDF) of the absolute rollout errors aggregated over all temperatures. The proposed model consistently attains the lowest rollout error and the most concentrated error distribution, indicating that the learned latent dynamics remain stable under recursive prediction, whereas the observation-space baselines degrade substantially when operated in free-running mode.

captures dynamical structure that transfers across operating bias. For reference, we also report a *zero-shot* configuration in which the source model is applied directly to the target bias without any adaptation. The same protocol is applied to all baselines, which are re-estimated using the identical training prefix.

Fig. 13 (a) reports the rollout RMSE across temperature under the unseen bias condition. Both the zero-shot and adapted versions of the proposed model maintain low prediction error across the full temperature range, indicating that the latent representation transfers well between bias regimes. In contrast, the baseline methods degrade substantially. The Ridge–ARX model exhibits pronounced temperature-dependent errors, while the boosted-tree predictor remains much less stable in recursive prediction.

The error distributions are summarized in Fig. 13 (b), which shows the empirical cumulative distribution of the absolute rollout errors aggregated over all temperatures. The proposed method reaches high cumulative probability at very small error values, indicating that accurate prediction is maintained throughout the rollout trajectories. By comparison, the baselines exhibit much heavier error tails, reflecting the accumulation of dynamical drift under recursive prediction.

Representative waveform overlays for three temperatures are shown in Fig. 14. The proposed model remains closely aligned with the measured optical waveform even under the unseen bias condition, preserving both the oscillatory structure and the large-signal transitions. The Ridge–ARX predictor departs from the correct phase trajectory, while the boosted-tree model produces increasing waveform distortion as the horizon grows.

To further assess the scalability of the surrogate beyond the operating point used for training, we also consider a cross-baud transfer experiment. In this experiment, the shared encoder trained at 20 Gbaud and 6 mA is kept fixed and evaluated on aligned 40 Gbaud waveforms at the same bias current. As in the unseen-bias experiment, only the lightweight affine readout is re-estimated from the training prefix of each target-temperature sweep. This experiment tests whether the latent representation learned at one symbol rate can be reused when the waveform dynamics are accelerated by a higher baud rate.

Fig. 15(a) reports the rollout RMSE across temperature. The proposed frozen-encoder model with an adapted affine readout achieves the lowest RMSE over almost the entire temperature range. Its error remains below that of the adapted Ridge–ARX and boosted-tree baselines, which exhibit much larger temperature-dependent variations and less stable recursive behavior. Fig. 15(b) shows the corresponding empirical CDF of the absolute rollout error. The proposed adapted model reaches high

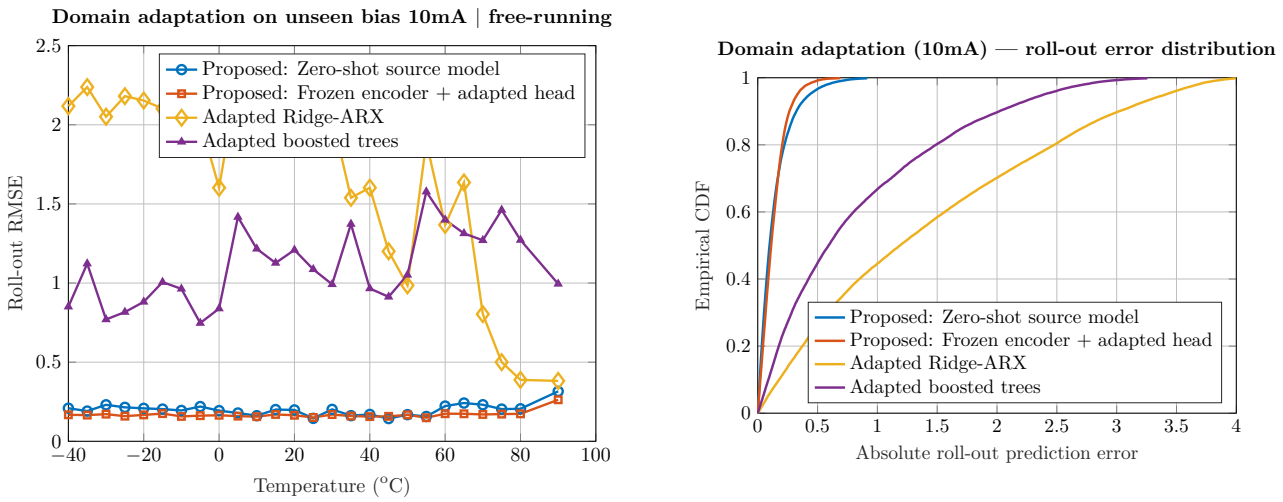

Figure 13: Transfer to an unseen bias current (10 mA). (a) Rollout RMSE across temperature for the zero-shot source model, the proposed method with a frozen encoder and adapted head, and the adapted baseline predictors. The proposed model maintains low prediction error across the full temperature range, while the baseline methods degrade substantially under recursive prediction. (b) Empirical cumulative distribution of the absolute rollout prediction errors aggregated over all temperatures. The proposed model concentrates most errors at significantly smaller values, indicating stable prediction over extended rollout windows under the unseen operating condition.

cumulative probability at smaller error values than the baselines, confirming that its rollout errors are more concentrated and that the learned latent representation provides useful predictive structure at the higher baud rate.

Compared with the unseen-bias transfer case in Fig. 13, however, the cross-baud experiment leads to a clear increase in prediction error. This indicates that transfer across baud rate is more challenging than transfer across bias. Changing the bias mainly shifts the operating regime of the VCSEL while preserving the symbol-rate time scale, whereas changing the baud rate modifies the effective temporal memory, transition density, and high-frequency waveform content. Therefore, although the proposed method remains the strongest among the tested models, adapting only the final affine readout is not sufficient to fully recover the waveform fidelity obtained in the fixed-baud experiments.

The waveform overlays in Fig. 16 provide a time-domain view of this behavior. At -40°C, $+5$°C, and $+80$°C, the proposed adapted model follows the main waveform trajectory and remains closer to the ground truth than the Ridge–ARX and boosted-tree baselines. However, the agreement is weaker than in the unseen-bias overlays in Fig. 14, especially around sharp peaks, fast transitions, and deep falls. This suggests that the frozen encoder captures part of the transferable dynamics, but the change from 20 to 40 Gbaud introduces temporal features that cannot be fully compensated by readout adaptation alone.

Overall, these results indicate that the encoder learns a latent representation that is highly transferable across operating bias and remains useful under a moderate baud-rate shift. For bias transfer, moving the surrogate to a new operating current requires at most a lightweight adjustment of the affine readout, while the underlying dynamical representation remains valid. For baud-rate transfer, the same adaptation strategy still outperforms the baseline predictors and captures the dominant waveform evolution, but the increased rollout error and reduced peak fidelity show that changes in symbol rate affect the temporal dynamics more strongly than changes in bias. Thus, the proposed surrogate provides efficient adaptation across operating conditions that preserve the main temporal structure, while robust transfer across baud rates motivates baud-aware or multi-baud encoder training.

4.9 Computational Footprint of the Surrogate

Beyond waveform accuracy, it is also useful to characterize the computational footprint of the proposed surrogate itself. For the architecture used in this work, the shared encoder contains 35,680 trainable

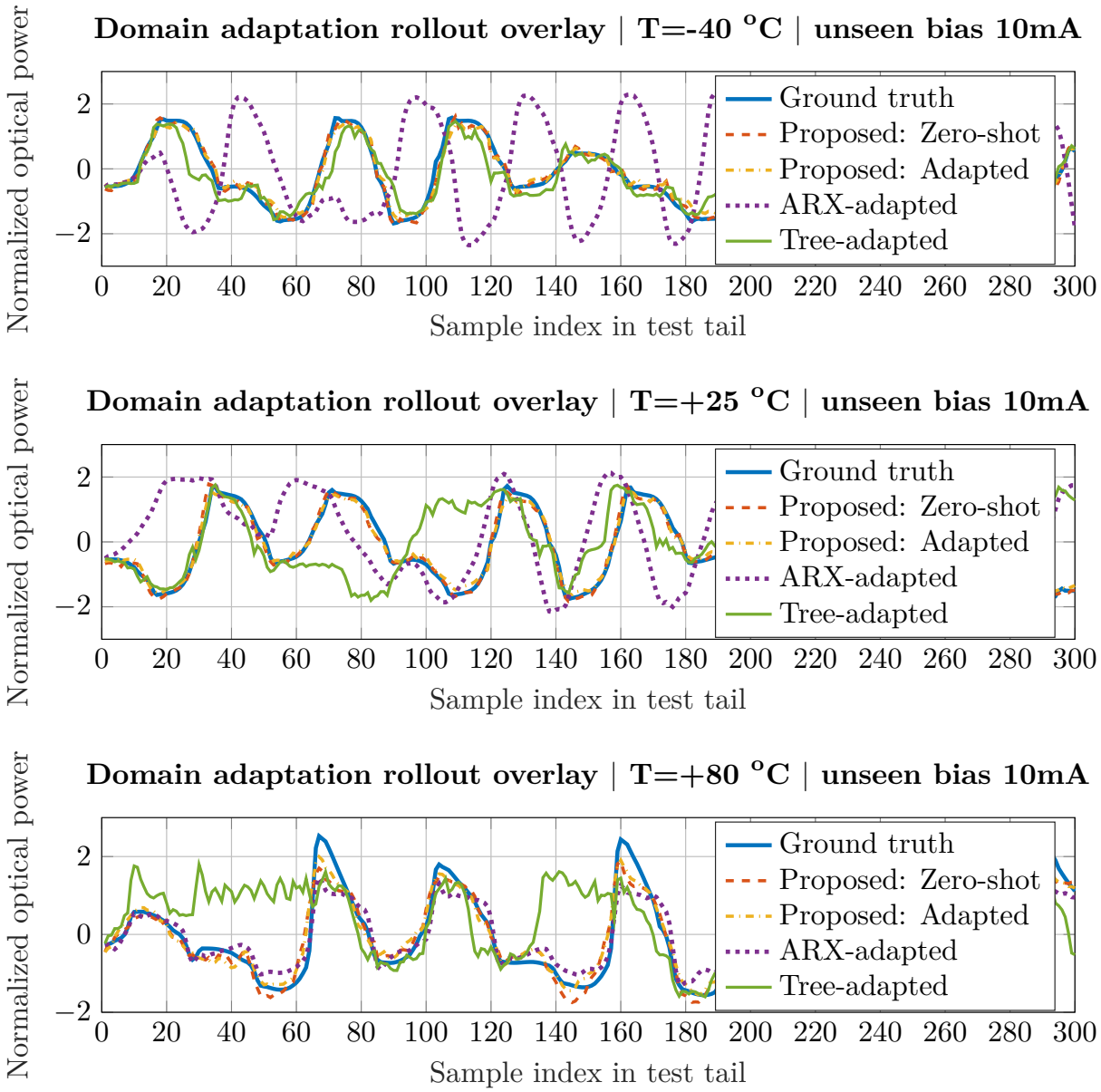

Figure 14: Free-running rollout waveform overlays under the unseen bias condition (10 mA). The plots compare the measured optical power trajectory with predictions from the zero-shot source model, the proposed adapted model, and the adapted baselines. Three representative temperatures are shown: low (−40°C), nominal (+25°C), and high (+80°C) . Both the zero-shot and adapted versions of the proposed model remain closely aligned with the ground-truth waveform across the full prediction window, while the baseline predictors accumulate phase and amplitude distortions during recursive prediction.

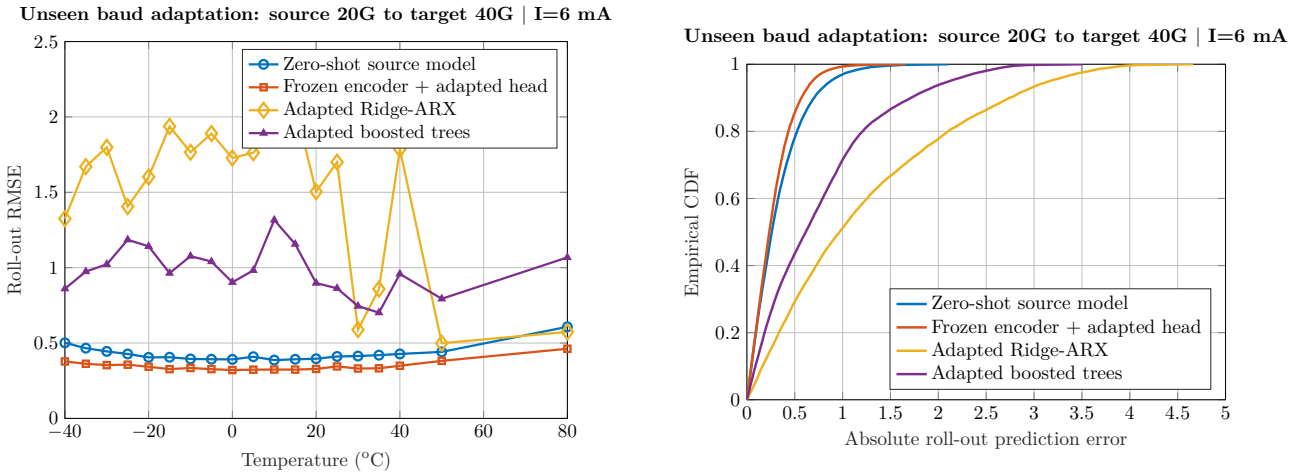

Figure 15: Transfer to an unseen baud rate. The shared encoder is trained at 20 Gbaud and 6 mA and evaluated on aligned 40 Gbaud waveforms at the same bias current. Only the affine readout is adapted using the target training prefix, while the encoder remains fixed. (a) Rollout RMSE across temperature for the zero-shot source model, the proposed frozen-encoder model with adapted readout, and the adapted baseline predictors. The proposed adapted model achieves the lowest error over almost the entire temperature range, although the cross-baud errors are higher than in the unseen-bias case. (b) Empirical cumulative distribution of the absolute rollout prediction errors aggregated over all temperatures. The proposed adapted model concentrates the errors at smaller values than the adapted Ridge–ARX and boosted-tree baselines, indicating more stable recursive prediction under the higher-baud target condition.

parameters, corresponding to approximately 0.136 MB in single precision, while each temperature-specific affine head adds only 33 parameters. Including the auxiliary shared scalar output layer used during Stage 1 training, the full training network contains 35,713 trainable parameters. Thus, once the shared encoder has been learned, adaptation to a new operating condition reduces to solving a small ridge-regression problem in the latent space.

On a Windows 64-bit machine with a 13th Gen Intel(R) Core(TM) i7-1370P CPU, 31.7 GB RAM, and MATLAB R2023b Update 7, Stage 1 training required approximately 10.85 minutes, while Stage 2 head fitting required on average 45.3 ms per temperature sweep. In addition, free-running generation of a rollout of length 1,200 required 854.9 ± 115.6 ms. For reference, batched latent-feature extraction on precomputed states required 13.1 ± 0.9 ms, while batched evaluation of the affine head on precomputed latent vectors required 2.14 ± 0.37 ms. [2] These measurements do not constitute a full benchmark against the reference circuit-simulation workflow, which is outside the scope of the present study, but they confirm that the surrogate itself is computationally modest and suitable for repeated waveform-level evaluation.

5 Conclusion and Practical Implications

We presented a compact data-driven surrogate model for thermally coupled VCSEL dynamics aimed at waveform-level evaluation during fabrication-stage studies. The proposed approach combines an empirically selected finite-memory state, a shared nonlinear latent representation, and lightweight operating-condition-specific affine readouts. By training the shared representation under closed-loop rollout, the surrogate is optimized for recursive waveform generation rather than only one-step prediction.

The results show that the proposed surrogate can reproduce optical waveforms accurately over extended durations, while maintaining a compact parameterization and low-cost per-condition adaptation. The shared latent representation transfers effectively across temperature and operating bias through lightweight affine readout calibration. Additional cross-baud evaluation at 40 Gbaud indicates that the

[2] Note that the individual timings are not directly additive to the free-running rollout time, since recursive generation involves 1200 sequential model evaluations together with state updates and feedback.

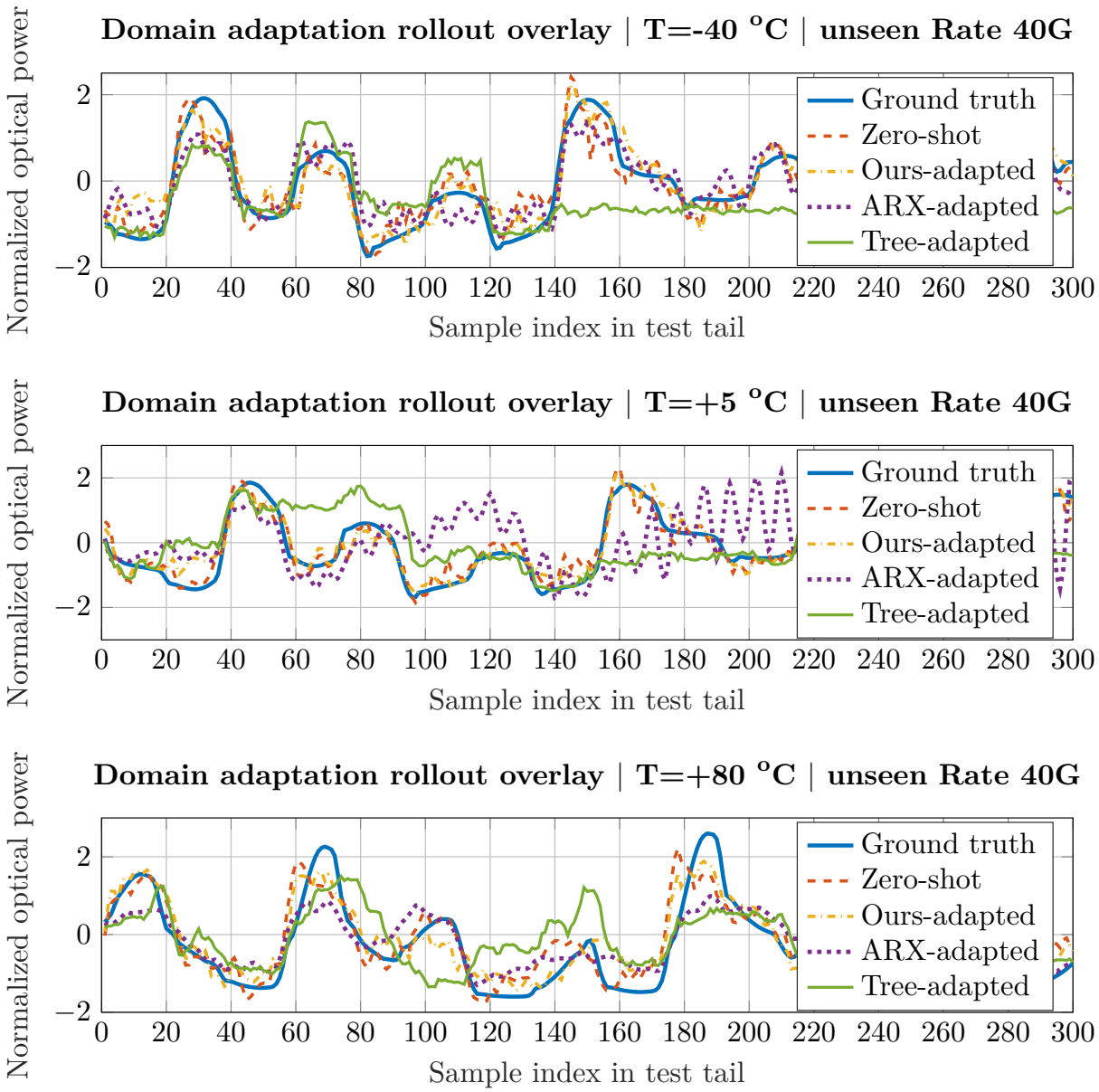

Figure 16: Free-running rollout waveform overlays under the unseen-baud condition. The encoder is trained at 20 Gbaud and evaluated at 40 Gbaud for the same 6 mA bias current, with only the affine readout adapted from the target training prefix. Three representative temperatures are shown: low (−40°C), nominal (+5°C), and high (+80°C). The proposed adapted model follows the dominant waveform trend and remains closer to the ground truth than the adapted Ridge–ARX and boosted-tree baselines. Compared with the unseen-bias transfer case, however, the agreement around sharp peaks, rapid transitions, and deep falls is reduced, showing that baud-rate changes modify the effective temporal dynamics more strongly than bias changes.

learned representation remains useful beyond the training symbol rate, although baud-rate changes impose a stronger dynamical shift than temperature or bias variations. This suggests that lightweight readout adaptation is well suited to operating-condition changes that preserve the main temporal structure, while robust multi-rate operation will benefit from baud-aware or multi-baud training.

From a practical perspective, the proposed surrogate is relevant in settings where waveform simulations must be repeated many times during model exploration or operating-condition sweeps. The framework can be implemented using either high-fidelity simulator outputs or measured waveform data from fabricated VCSEL devices. In a laboratory setting, such data could be obtained from a temperature-controlled VCSEL setup with controlled biasing, high-speed electrical modulation, optical detection, and real-time waveform acquisition. Repeating the acquisition across bias currents, modulation conditions, and ambient temperatures would provide the paired electrical-input–optical-output trajectories required for surrogate training and evaluation. Practical implementation would require careful calibration of the drive waveform and detected optical power, synchronization between electrical excitation and optical measurement, and attention to measurement noise, bandwidth limitations, packaging parasitics, and device-to-device variability.

The present study focuses on waveform fidelity under teacher-forced and free-running prediction. A full benchmark inside a complete circuit-simulation loop, as well as evaluation of downstream signal-integrity metrics such as eye opening, extinction ratio, TDECQ, BER, and related link-level measures, remains an important next step.

More broadly, the results suggest that the dominant waveform-level dynamics of the considered VCSEL system can be captured through a compact observable representation, and that much of this structure is shared across operating conditions. This supports the use of latent dynamical surrogates as an intermediate modeling layer between high-fidelity device physics and repeated waveform-level analysis. Future work should focus on baud-aware surrogate design and more robust latent representations for multi-rate operation. The 40 Gbaud stress test indicates that baud-rate changes modify the effective temporal dynamics, so a frozen single-baud encoder with readout adaptation alone cannot fully recover the waveform fidelity obtained at the training baud rate. However, the model still follows the dominant waveform evolution, suggesting that multi-baud encoder training, symbol-rate conditioning, or temporal rescaling may enable broader operation across 50 Gbaud and higher regimes relevant to modern short-reach links. Further extensions should include tighter integration into circuit-simulation environments, link-level metrics, and validation using measured device data in addition to simulator-generated waveforms.

Author Contributions

Conceptualization, A.P., S.M. and M.S.; methodology, A.P.; software, A.P., S.M; validation, A.P.; formal analysis, A.P.; investigation, A.P.; data curation, S.M. and A.P.; writing—original draft preparation, A.P., S.M. and M.S.; writing—review and editing, A.P., L.S., P.A.A. and H.W.; visualization, A.P.; supervision, H.W., L.S., and P.A.A.; project administration, L.S., P.A.A. and H.W.; funding acquisition, H.W., L.S., and P.A.A. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by HOT OPTICS under Grant CHI19-0004.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The data and code supporting the results of this study are planned to be made publicly available at the following GitHub repository:
https://github.com/alirezap94/vcsel-surrogate-modeling.

Conflicts of Interest

The authors declare no conflicts of interest.

A Predictive Error Analysis

This appendix provides the full predictive-error analysis used to determine the empirical memory order in Sec. 3.1. The goal is to quantify how much observable input–output history is required to predict the next optical-power sample, while keeping the estimation procedure statistically well posed and reproducible.

A.1 Finite-memory regression problem

Stacking states $\mathbf{s}_t^{(L)}$ in (1) row-wise yields the design matrix

$$\mathbf{S}^{(L)} \triangleq \begin{bmatrix} (\mathbf{s}_L^{(L)})^\mathsf{T} \\ \vdots \\ (\mathbf{s}_{N-1}^{(L)})^\mathsf{T} \end{bmatrix} \in \mathbb{R}^{M\times(2L+1)}, \qquad M = N - L, \tag{31}$$

and the corresponding target vector

$$\mathbf{y}^+ \triangleq \begin{bmatrix} y_{L+1}^{(L)} \\ \vdots \\ y_N^{(L)} \end{bmatrix} \in \mathbb{R}^M. \tag{32}$$

The resulting regression problem asks how well $y_{t+1}^{(L)}$ can be predicted from the observable state $\mathbf{s}_t^{(L)}$.

Remark 5. The term x_{t+1} is included in (1) because the physical device is a controlled dynamical system: the next optical-power sample depends not only on past state information but also on the current drive action applied over the next interval. Including x_{t+1} therefore makes the one-step predictor causal with respect to the known input sequence.

Let $\rho \in (0, 1)$ denote the training fraction. With $M = N - L$ state-target pairs, define

$$M_{\mathrm{tr}} = \lfloor \rho M \rfloor, \qquad M_{\mathrm{te}} = M - M_{\mathrm{tr}}. \tag{33}$$

The training and test blocks are then

$$\mathbf{S}_{\mathrm{tr}}^{(L)} = \mathbf{S}^{(L)}(1 : M_{\mathrm{tr}}, :), \qquad \mathbf{S}_{\mathrm{te}}^{(L)} = \mathbf{S}^{(L)}(M_{\mathrm{tr}} + 1 : M, :), \tag{34}$$

and similarly

$$\mathbf{y}_{\mathrm{tr}}^+ = \mathbf{y}^+(1 : M_{\mathrm{tr}}), \qquad \mathbf{y}_{\mathrm{te}}^+ = \mathbf{y}^+(M_{\mathrm{tr}} + 1 : M). \tag{35}$$

The coordinates of $\mathbf{S}^{(L)}$ may have different empirical scales, particularly when output lags, input lags, and the current input sample are combined in the same regression vector. To make the ridge penalty comparable across coordinates, we standardize features column-wise using training statistics only. Let

$$\boldsymbol{\mu}_S^{(L)} = \frac{1}{M_{\mathrm{tr}}} \sum_{i=1}^{M_{\mathrm{tr}}} \mathbf{S}_{\mathrm{tr}}^{(L)}(i, :)^\mathsf{T}, \tag{36}$$

and let $\boldsymbol{\sigma}_S^{(L)} \in \mathbb{R}^{2L+1}$ denote the vector of column-wise sample standard deviations computed from $\mathbf{S}_{\mathrm{tr}}^{(L)}$. The standardized matrices are

$$\tilde{\mathbf{S}}_{\mathrm{tr}}^{(L)} = (\mathbf{S}_{\mathrm{tr}}^{(L)} - \mathbf{1}(\boldsymbol{\mu}_S^{(L)})^{\mathsf{T}}) \oslash (\mathbf{1}(\boldsymbol{\sigma}_S^{(L)})^{\mathsf{T}}), \tag{37}$$

$$\tilde{\mathbf{S}}_{\mathrm{te}}^{(L)} = (\mathbf{S}_{\mathrm{te}}^{(L)} - \mathbf{1}(\boldsymbol{\mu}_S^{(L)})^{\mathsf{T}}) \oslash (\mathbf{1}(\boldsymbol{\sigma}_S^{(L)})^{\mathsf{T}}), \tag{38}$$

where $\oslash$ denotes elementwise division and $\mathbf{1}$ is an all-ones column vector.

A.2 Ridge-regularized one-step predictor

For each memory length L, we fit a linear one-step predictor using ridge regularization. This choice is appropriate because lagged waveform coordinates are strongly collinear, and the conditioning of the regression problem typically worsens as L increases [34, Sec. 3.4.1].

To include an intercept, augment the standardized design matrices as

$$\tilde{\mathbf{S}}_{\mathrm{tr},b}^{(L)} = \begin{bmatrix} \tilde{\mathbf{S}}_{\mathrm{tr}}^{(L)} & \mathbf{1} \end{bmatrix} \in \mathbb{R}^{M_{\mathrm{tr}} \times (2L+2)}, \tag{39}$$

$$\tilde{\mathbf{S}}_{\mathrm{te},b}^{(L)} = \begin{bmatrix} \tilde{\mathbf{S}}_{\mathrm{te}}^{(L)} & \mathbf{1} \end{bmatrix} \in \mathbb{R}^{M_{\mathrm{te}} \times (2L+2)}. \tag{40}$$

The ridge estimator is

$$\hat{\mathbf{w}}^{(L)} = \arg \min_{\mathbf{w} \in \mathbb{R}^{2L+2}} \left\| \mathbf{y}_{\mathrm{tr}}^{+} - \tilde{\mathbf{S}}_{\mathrm{tr},b}^{(L)} \mathbf{w} \right\|_2^2 + \lambda \left\| \mathbf{D}^{(L)} \mathbf{w} \right\|_2^2, \tag{41}$$

where

$$\mathbf{D}^{(L)} = \mathrm{diag}(1, \ldots, 1, 0) \in \mathbb{R}^{(2L+2) \times (2L+2)} \tag{42}$$

penalizes all regression coefficients except the intercept. The ridge parameter was fixed to $\lambda = 0.01$ throughout the predictive-error analysis.

The corresponding closed-form solution is

$$\hat{\mathbf{w}}^{(L)} = \left((\tilde{\mathbf{S}}_{\mathrm{tr},b}^{(L)})^{\mathsf{T}} \tilde{\mathbf{S}}_{\mathrm{tr},b}^{(L)} + \lambda (\mathbf{D}^{(L)})^{\mathsf{T}} \mathbf{D}^{(L)} \right)^{-1} (\tilde{\mathbf{S}}_{\mathrm{tr},b}^{(L)})^{\mathsf{T}} \mathbf{y}_{\mathrm{tr}}^{+}. \tag{43}$$

The one-step test predictions are

$$\hat{\mathbf{y}}_{\mathrm{te}}^{+,(L)} = \tilde{\mathbf{S}}_{\mathrm{te},b}^{(L)} \hat{\mathbf{w}}^{(L)}, \tag{44}$$

and the corresponding test mean-squared error is

$$\mathrm{MSE}(L) = \frac{1}{M_{\mathrm{te}}} \left\| \mathbf{y}_{\mathrm{te}}^{+} - \hat{\mathbf{y}}_{\mathrm{te}}^{+,(L)} \right\|_2^2. \tag{45}$$

Remark 6. The purpose of this stage is not to construct the final nonlinear surrogate. It is to determine how much predictive information is retained by an observable history of length L. A ridge-regularized linear predictor is therefore appropriate: it is stable, inexpensive, and sufficiently informative for detecting the point at which additional history ceases to improve one-step prediction.

References

[1] Q. Cheng, M. Bahadori, M. Glick, S. Rumley, and K. Bergman, "Recent Advanced in Optical Technologies for Data Centers: a Review," *Optica*, vol. 5, no. 11, pp. 1354–1370, 2018.

[2] A. Maharry, L. A. Valenzuela, H. Andrade, I. Kalifa, I. Cestier, M. Galanty, B. Atias, A. Sandomirsky, E. Mentovich, L. Coldren *et al.*, "A 50 Gbps 9.5 pJ/bit VCSEL-based Optical Link," in *IEEE Int. Conf. Photonics (IPC)*. IEEE, 2021, pp. 1–2.

[3] D. Mahgerefteh, C. Thompson, C. Cole, G. Denoyer, T. Nguyen, I. Lyubomirsky, C. Kocot, and J. Tatum, "Techno-economic Comparison of Silicon Photonics and Multimode VCSELs," *IEEE J. Lightw. Technol.*, vol. 34, no. 2, pp. 244–242, 2016.

[4] H.-T. Cheng, Y.-C. Yang, T.-H. Liu, and C.-H. Wu, "Recent Advances in 850 nm VCSELs for High-Speed Interconnects," *Photonics*, vol. 9, no. 2, 2022.

[5] S. Rumley, M. Bahadori, R. Polster, S. D. Hammond, D. M. Calhoun, K. Wen, A. Rodrigues, and K. Bergman, "Optical Interconnects for Extreme Scale Computing Systems," *Parallel Comput.*, vol. 64, pp. 65–80, 2017.

[6] T. Aoki, R. Kubota, H. Hiiro, S. Yoshimoto, and M. Yanagisawa, "50 gb/s pam-4 vcsels operating up to 125°c," in *Proc. Optoelectron. Commun. Conf. (OECC)*. Optica Publishing Group, 2021, p. T3D.4.

[7] R. King, "VCSEL Design for Automotive Datacom Experimental Results for 980 nm versus 850 nm," Tech. Rep., 2021, [Accessed Nov. 12, 2022]. [Online]. Available: https://www.ieee802.org/3/cz/public/may_2021/king_3cz_01a_0521.pdf

[8] F. Z. Jasim, K. Omar, Z. Hassan *et al.*, "Temperature effect on vcsel output performance," *J. Optoelectron. Adv. Mater.*, vol. 3, pp. 1136–1138, 2009.

[9] T. E. Sale, J. S. Roberts, J. P. David, R. Grey, J. Woodhead, and P. N. Robson, "Temperature effects in vcsels," in *Vertical-Cavity Surface-Emitting Lasers*, vol. 3003. SPIE, 1997, pp. 100–110.

[10] H. Li, P. Wolf, X. Jia, J. A. Lott, and D. Bimberg, "Thermal analysis of high-bandwidth and energy-efficient 980nm vcsels with optimized quantum well gain peak-to-cavity resonance wavelength offset," *Applied Physics Letters*, vol. 111, no. 24, p. 243508, 12 2017.

[11] S. Entezam, A. Zarifkar, and M. H. Sheikhi, "Thermal equivalent circuit model for coupled-cavity surface-emitting lasers," *IEEE Journal of Quantum Electronics*, vol. 51, no. 4, pp. 1–8, 2015.

[12] P. V. Mena, J. Morikuni, S.-M. Kang, A. Harton, and K. Wyatt, "A simple rate-equation-based thermal vcsel model," *Journal of lightwave Technology*, vol. 17, no. 5, p. 865, 1999.

[13] P.-C. Pan, D. Nag, Z. Khan, C.-J. Chen, J.-W. Shi, A. Laha, and R.-H. Horng, "Effect of thermal management on the performance of vcsels," *IEEE Transactions on Electron Devices*, vol. 67, no. 9, pp. 3736–3739, 2020.

[14] A. Grabowski, J. Gustavsson, Z. S. He, and A. Larsson, "Large-signal equivalent circuit for datacom VCSELs," *Journal of Lightwave Technology*, vol. 39, no. 10, pp. 3225–3236, 2021.

[15] M. B. Aziz, S. Giannakopoulos, A. Grabowski, and P. Andrekson, "Word length dependent sensitivity penalty in high speed vcsel-based optical interconnects," *IEEE Photonics Technology Letters*, vol. 36, no. 19, pp. 1181–1184, 2024.

[16] M. Srinivasan, J. Song, A. Grabowski, K. Szczerba, H. K. Iversen, M. N. Schmidt, D. Zibar, J. Schröder, A. Larsson, C. Häger *et al.*, "End-to-end learning for vcsel-based optical interconnects: State-of-the-art, challenges, and opportunities," *Journal of Lightwave Technology*, vol. 41, no. 11, pp. 3261–3277, 2023.

[17] M. Bou Sanayeh, W. Hamad, and W. Hofmann, "Equivalent circuit model of high-performance vcsels," *Photonics*, vol. 7, no. 1, 2020. [Online]. Available: https://www.mdpi.com/2304-6732/7/1/13

[18] S. Wu and S. Kang, "Modeling and time domain simulation of vcsel using vhdl-ams," in *Southwest Symposium on Mixed-Signal Design, 2003.*, 2003, pp. 170–174.

[19] K. Zhong, X. Zhou, J. Huo, C. Yu, C. Lu, and A. P. T. Lau, "Digital signal processing for short-reach optical communications: A review of current technologies and future trends," *IEEE J. Lightw. Technol.*, vol. 36, no. 2, pp. 377–400, 2018.

[20] T. S. R. Shen and A. P. T. Lau, "Fiber Nonlinearity Compensation using Extreme Learning Machine for DSP-based Coherent Communication Systems," in *Proc. Optoelectron. Commun. Conf. (OECC)*, 2011, pp. 816–817.

[21] S. Gaiarin, X. Pang, O. Ozolins, R. T. Jones, E. P. Da Silva, R. Schatz, U. Westergren, S. Popov, G. Jacobsen, and D. Zibar, "High Speed PAM-8 Optical Interconnects with Digital Equalization based on Neural Network," in *2016 Asia Commun. Photon. Conf. (ACP)*, 2016, pp. 1–3.

[22] L. Ge, W. Zhang, C. Liang, and Z. He, "Compressed Neural Network Equalization based on Iterative Pruning Algorithm for 112-Gbps VCSEL-enabled Optical Interconnects," *IEEE J. Lightw. Technol.*, vol. 38, no. 6, pp. 1323–1329, 2020.

[23] S. Deligiannidis, A. Bogris, C. Mesaritakis, and Y. Kopsinis, "Compensation of fiber nonlinearities in digital coherent systems leveraging long short-term memory neural networks," *IEEE J. Lightw. Technol.*, vol. 38, no. 21, pp. 5991–5999, 2020.

[24] A. Argyris, J. Bueno, and I. Fischer, "Photonic Machine Learning Implementation for Signal Recovery in Optical Communications," *Sci. Rep.*, vol. 8, no. 1, pp. 1–13, 2018.

[25] P. W. Berenguer, M. Nölle, L. Molle, T. Raman, A. Napoli, C. Schubert, and J. K. Fischer, "Nonlinear Digital Pre-distortion of Transmitter Components," *IEEE J. Lightw. Technol.*, vol. 34, no. 8, pp. 1739–1745, 2015.

[26] D. Lavery, R. Maher, G. Liga, D. Semrau, L. Galdino, and P. Bayvel, "On the Bandwidth Dependent Performance of Split Transmitter-receiver Optical Fiber Nonlinearity Compensation," *Opt. Express*, vol. 25, no. 4, pp. 4554–4563, 2017.

[27] G. Paryanti, H. Faig, L. Rokach, and D. Sadot, "A Direct Learning Approach for Neural Network based Pre-distortion for Coherent Nonlinear Optical Transmitter," *IEEE J. Lightw. Technol.*, vol. 38, no. 15, pp. 3883–3896, 2020.

[28] Y. Wu, U. Gustavsson, A. G. I. Amat, and H. Wymeersch, "Low Complexity Joint Impairment Mitigation of I/Q Modulator and PA Using Neural Networks," *IEEE J. Sel. Areas Commun*, vol. 40, no. 1, pp. 54–64, 2022.

[29] L. Minelli, F. Forghieri, T. Shao, A. Shahpari, and R. Gaudino, "Tdecq-based optimization of nonlinear digital pre-distorters for vcsel-mmf optical links using end-to-end learning," *Journal of Lightwave Technology*, 2023.

[30] Y. Khorrami, D. Fathi, and R. C. Rumpf, "Fast optimal design of optical components using the cultural algorithm," *Opt. Express*, vol. 28, no. 11, pp. 15 954–15 968, May 2020. [Online]. Available: https://opg.optica.org/oe/abstract.cfm?URI=oe-28-11-15954

[31] M. Schaedler, M. Kuschnerov, S. Calabrò, F. Pittalà, C. Bluemm, and S. Pachnicke, "AI-based Digital Predistortion for IQ Mach-Zehnder Modulators," in *2019 Asia Commun. Photon. Conf. (ACP)*, 2019, pp. 1–3.

[32] M. Srinivasan, J. Song, C. Häger, K. Szczerba, H. Wymeersch, and J. Schröder, "Learning optimal pam levels for vcsel-based optical interconnects," in *2022 European Conference on Optical Communication (ECOC)*. IEEE, 2022, pp. 1–4.

[33] A. Grabowski, J. S. Gustavsson, and A. Larsson, "Large-signal equivalent circuit for datacom VCSELs – Including Intensity Noise," *Journal of Lightwave Technology*, vol. 40, no. 22, pp. 7342–7350, 2022.

[34] T. Hastie, R. Tibshirani, J. Friedman *et al.*, "The elements of statistical learning," 2009.

[35] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural networks*, vol. 4, no. 2, pp. 251–257, 1991.

[36] K. Safarihamid, A. Pourafzal, and A. Fereidunian, "A joint-entropy approach to time-series classification," in *2021 7th International Conference on Signal Processing and Intelligent Systems (ICSPIS)*. IEEE, 2021, pp. 1–7.

[37] A. Pourafzal, H. D. Kaimre, C. Häger, P. A. Andrekson, and H. Wymeersch, "Pilot-free vcsel temperature monitoring via statistical complexity," *IEEE Photonics Journal*, 2026.

[38] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 627–635.

[39] P. Gao, E. Trautmann, B. Yu, G. Santhanam, S. Ryu, K. Shenoy, and S. Ganguli, "A theory of multineuronal dimensionality, dynamics and measurement," *BioRxiv*, p. 214262, 2017.

[40] S. A. Billings, *Nonlinear system identification: NARMAX methods in the time, frequency, and spatio-temporal domains*. John Wiley & Sons, 2013.

[41] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[42] L. Hassine, Z. Toffano, F. Lamnabhi-Lagarrigue, A. Destrez, and C. Birocheau, "Volterra functional series expansions for semiconductor lasers under modulation," *IEEE Journal of Quantum Electronics*, vol. 30, no. 4, pp. 918–928, 2002.